:probabl.

Own Your Data Science

Inrap

Institut national de recherches archéologiques préventives

Synaptix, un framework open source pour la génération pilotée de graphes de connaissances, appliqué au projet Archipel de l'INRAP

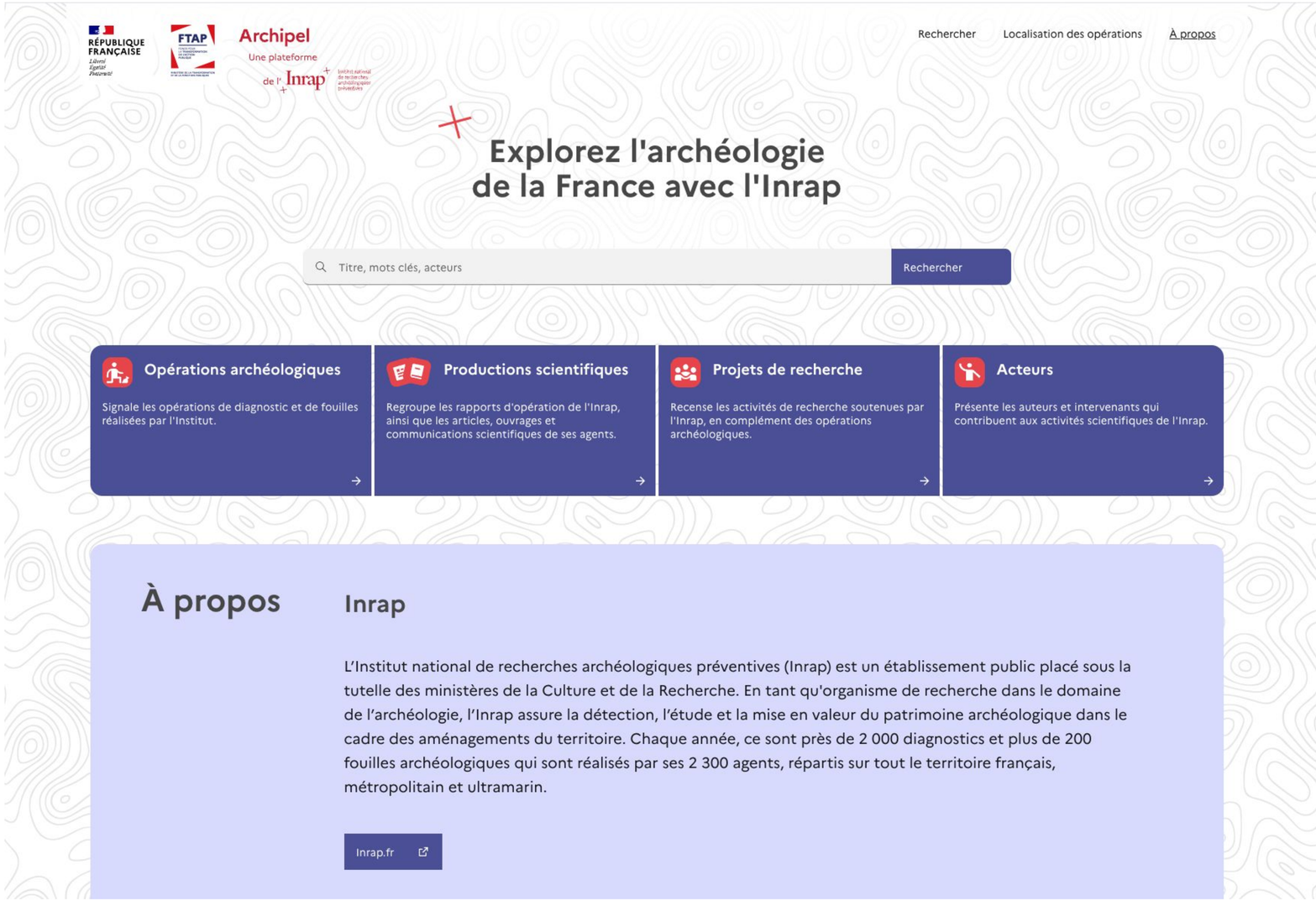
Freddy Limpens, freddy@probabl.ai

SemWeb.Pro

27 Novembre 2025

archipel.inrap.fr

- vitrine documentaire de l'inrap
- accès centralisé et ouvert aux données issues de 6 sources
- développé en partenariat avec reciproque (ui/ux), wavestone (amoa) et devops.works (infra)



Objectifs du projet Archipel Inrap

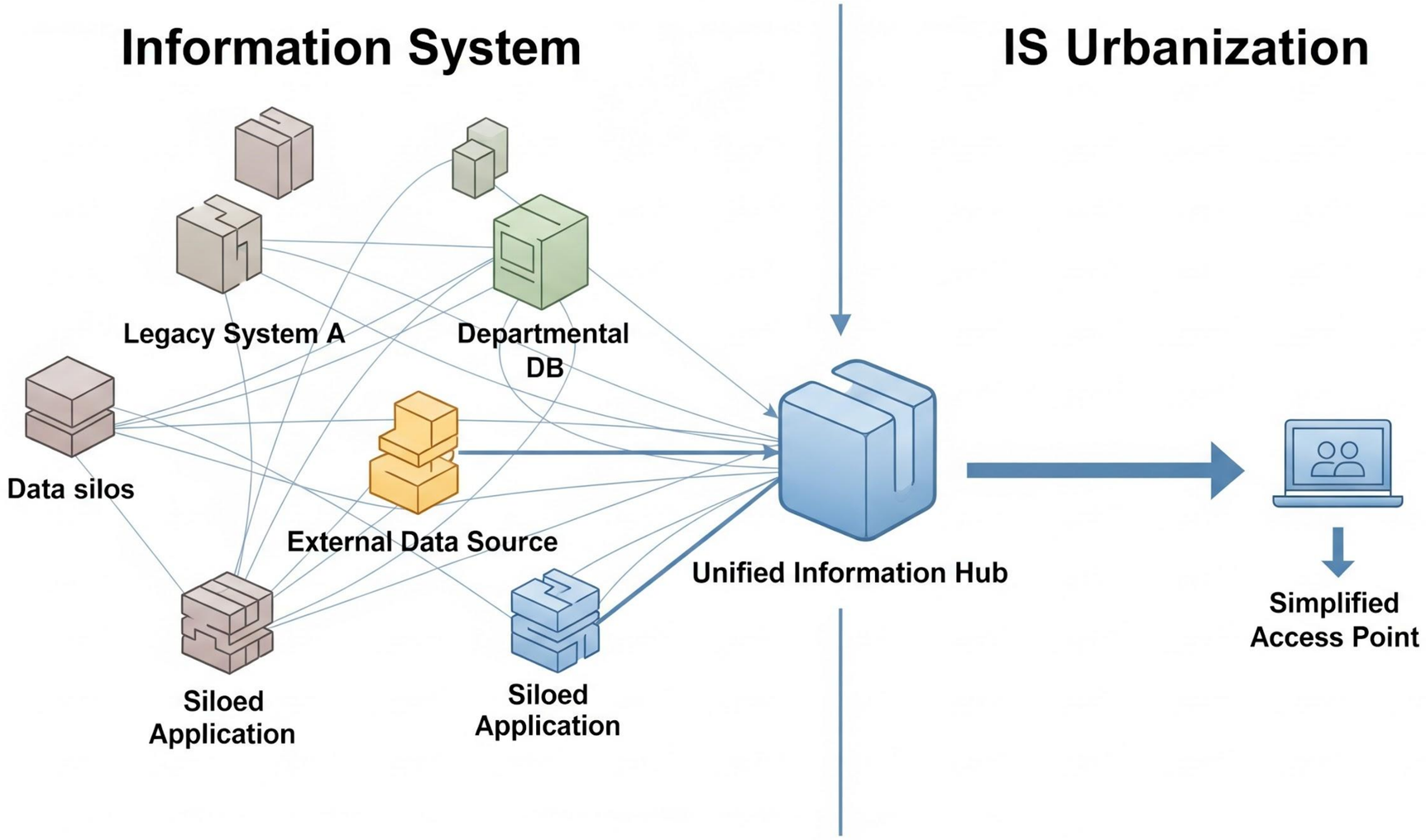
La Graph Data Science pour une meilleure visibilité de la production scientifique Inrap

- **Recréer du lien** entre les données
- **Valoriser** pour le public
- **Enrichir** les données grâce à l'analyse et le Knowledge Graph
- Garantir **interopérabilité**

Synaptix, méthode pour l'urbanisation des SI

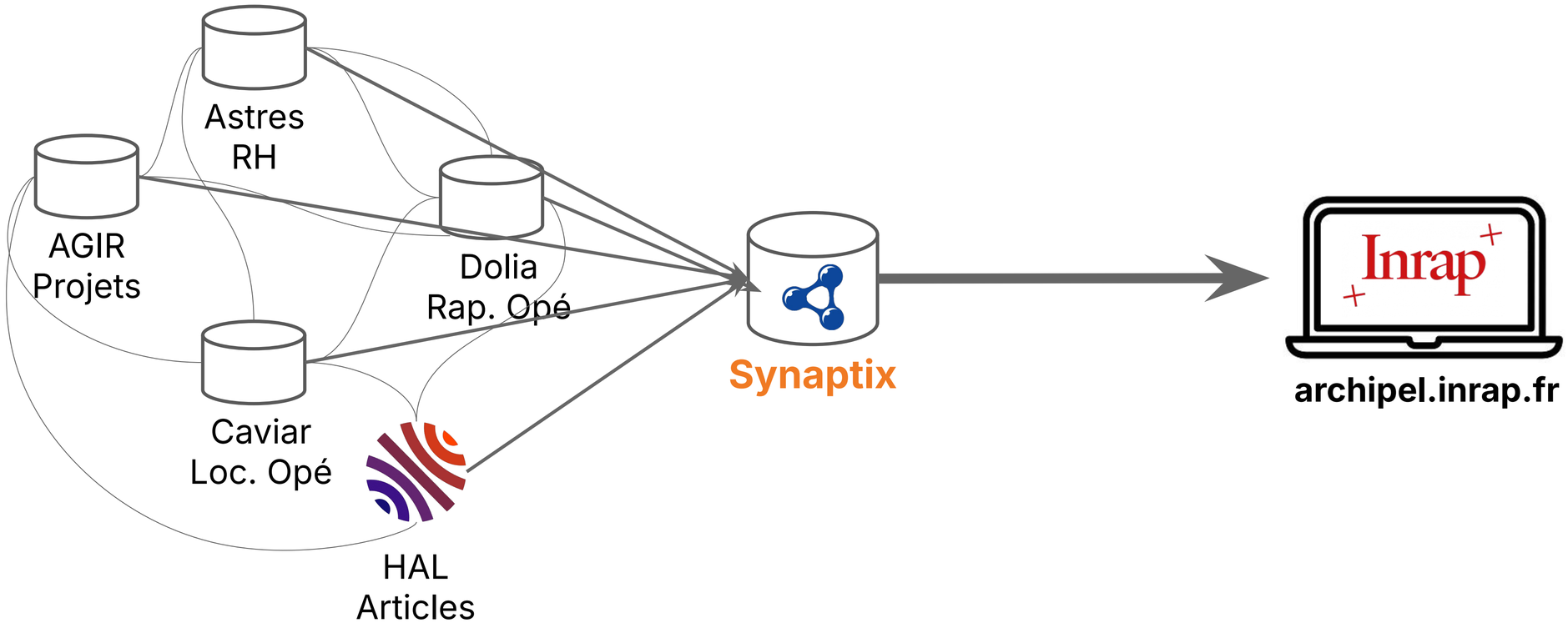
Urbanisation de SI

Urbaniser = (re)créer du lien entre les données + simplifier l'accès

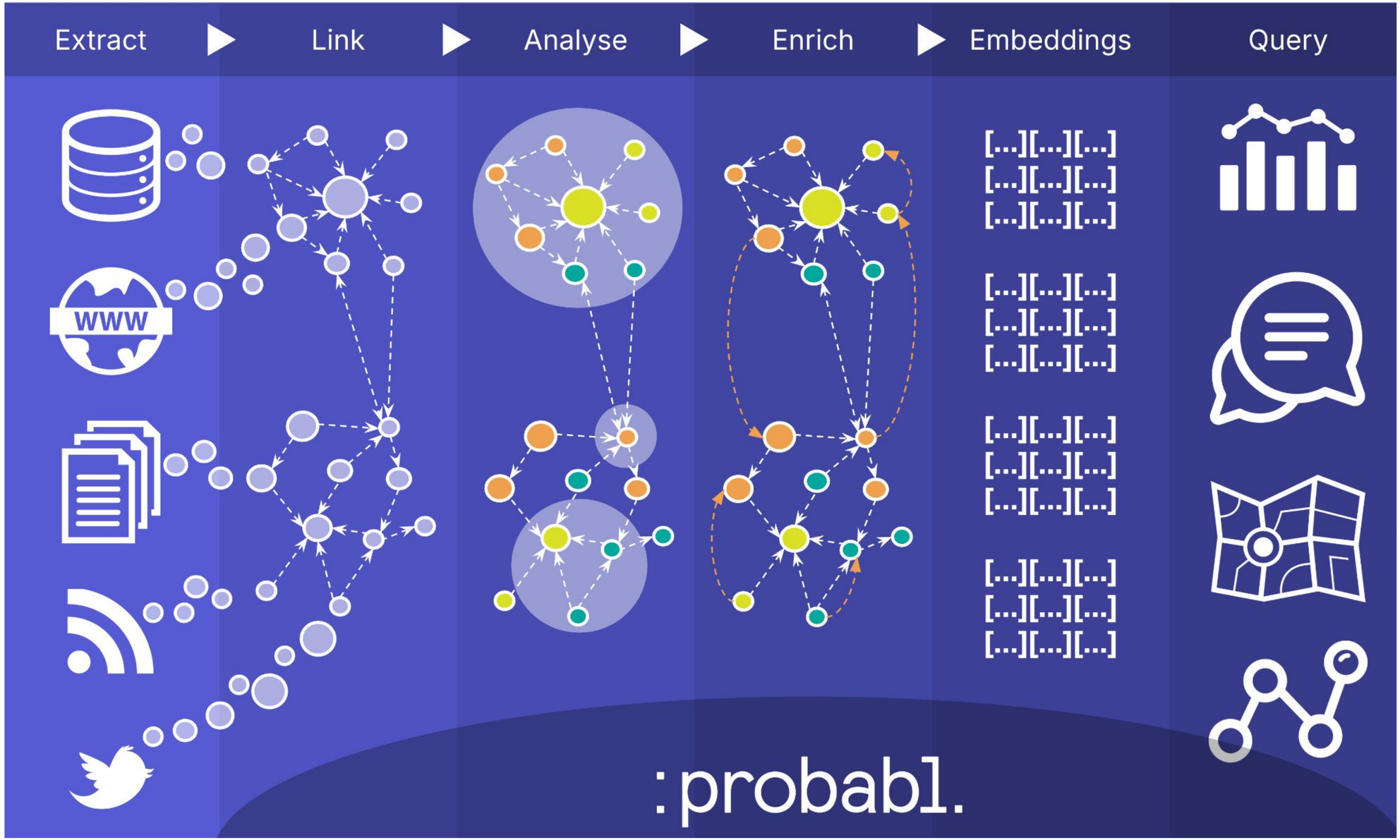


Urbanisation de SI

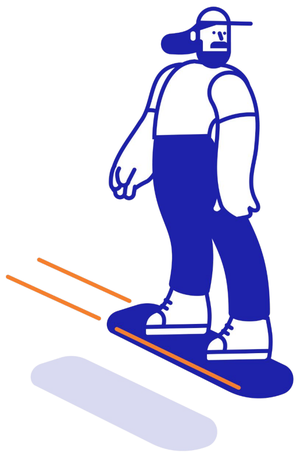
Urbaniser = (re)créer du lien entre les données + simplifier l'accès



Approche Synaptix de l'urbanisation



Synaptix, appliqué au projet Archipel



Data modeling

Construction d’un modèle pivot

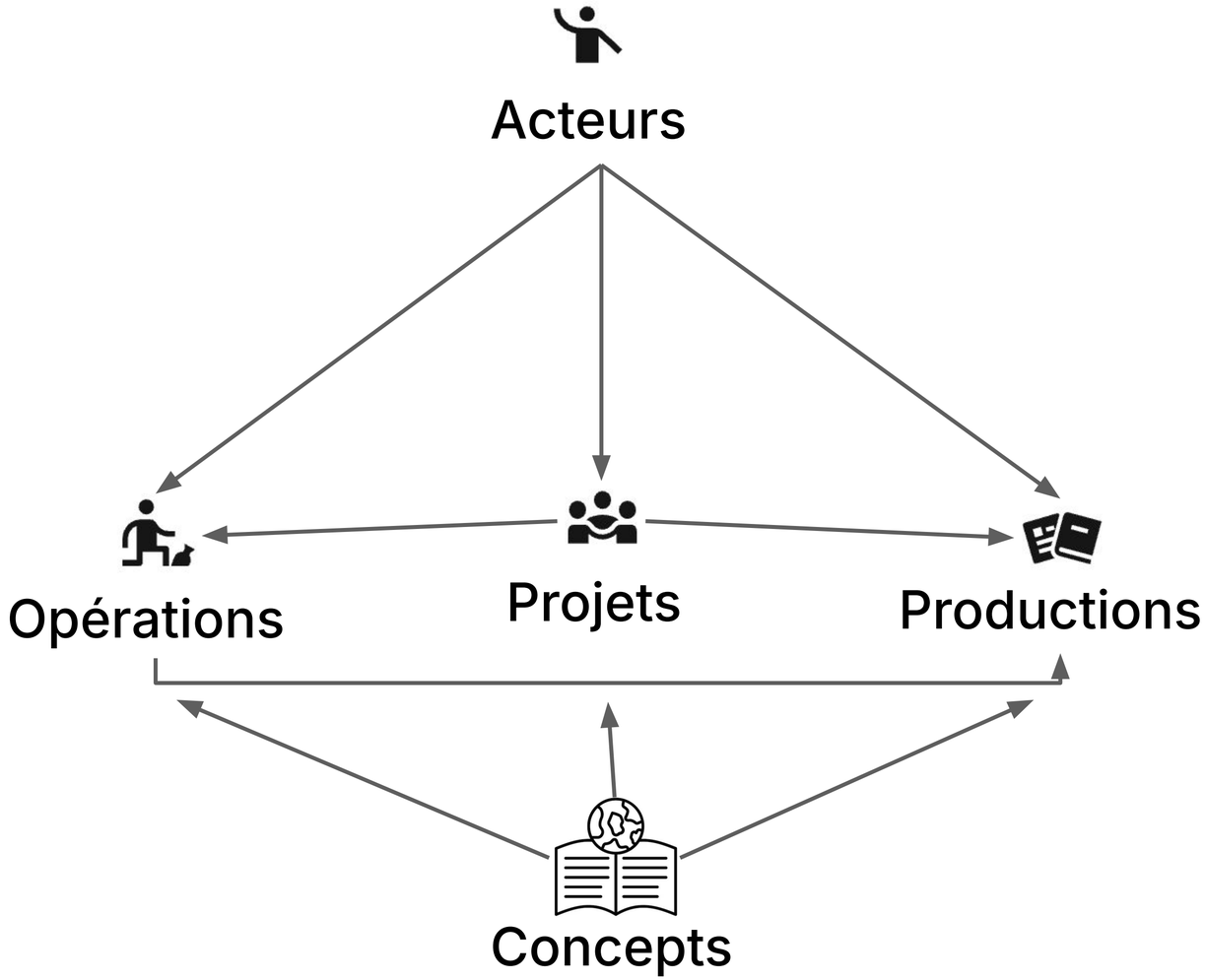
Objectifs :

- Articuler les données sources autour des concepts clés
- interopérabilité Ariadne, syst de partage des données archéo européen

Data modeling

Construction d'un modèle pivot

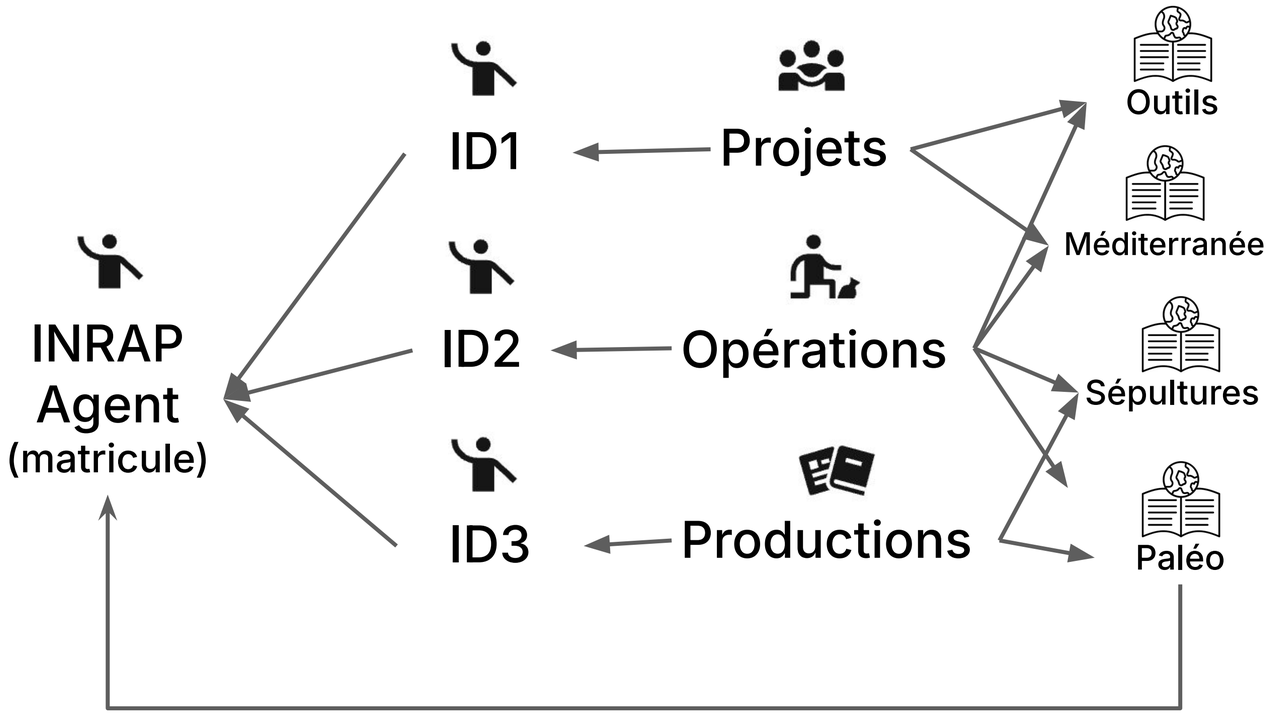
- 3 grands types de ressources
- 2 pivots principaux :
  - ↳ Acteurs (Person)
  - ↳ Concept (Pactols Thésau)



Data modeling

Graphes et enrichissement des profils des agents INRAP

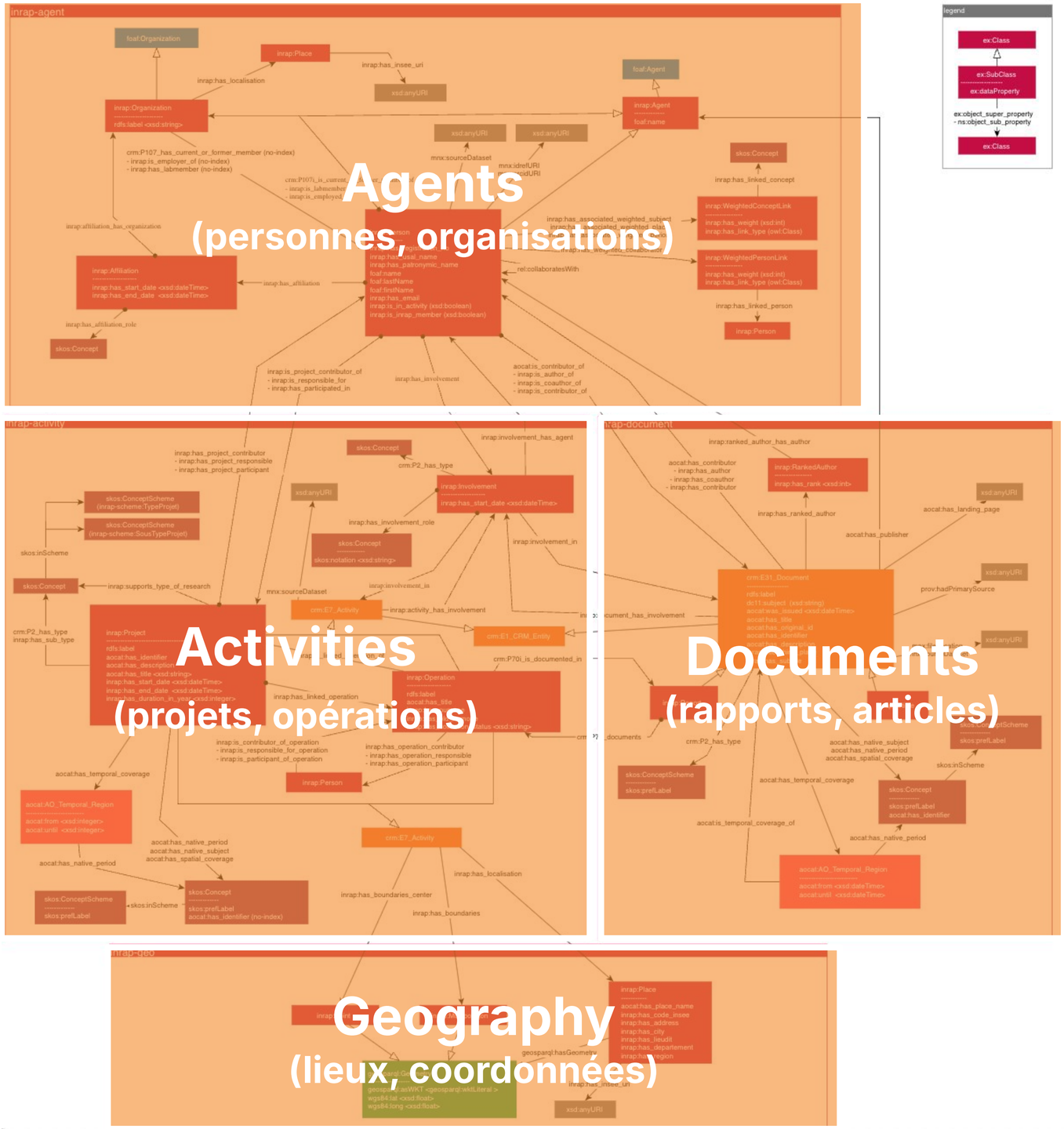
- Les ID des personnels INRAP sont alignés et articulés autour de leur ID RH
- Concepts associés aux acteurs via leurs contributions pondérées

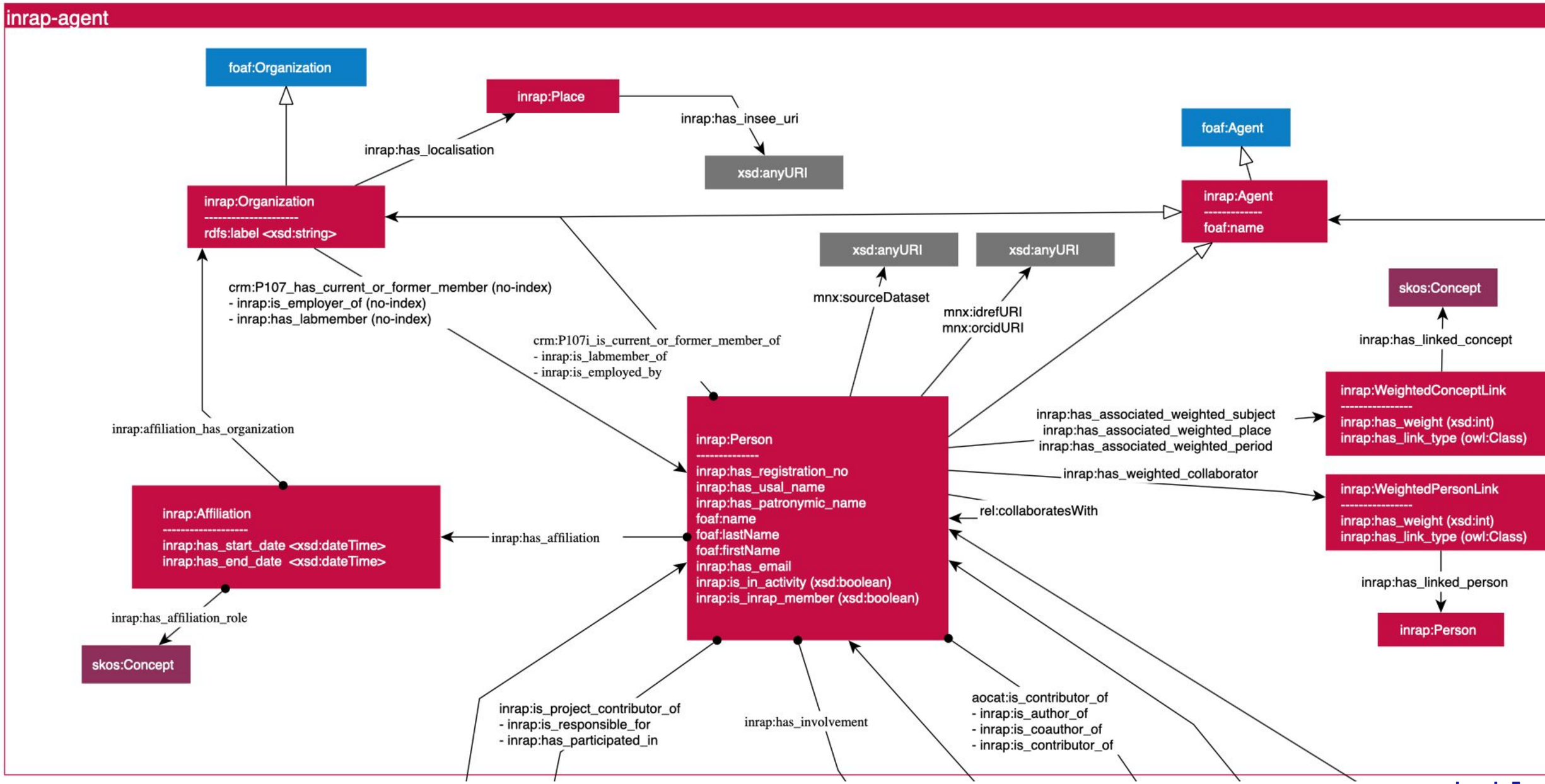


Modèle pivot

Ontologie RDFS/OWL

- Basée sur CidocCRM + ontologies standards (foaf, skos, etc.)
  - Variante AOCAT - projet Ariadne
  - Principales classes héritent de CidocCRM
  - réemploi des properties Aocat
- Découpages en modules
- Documentation complète (lien)

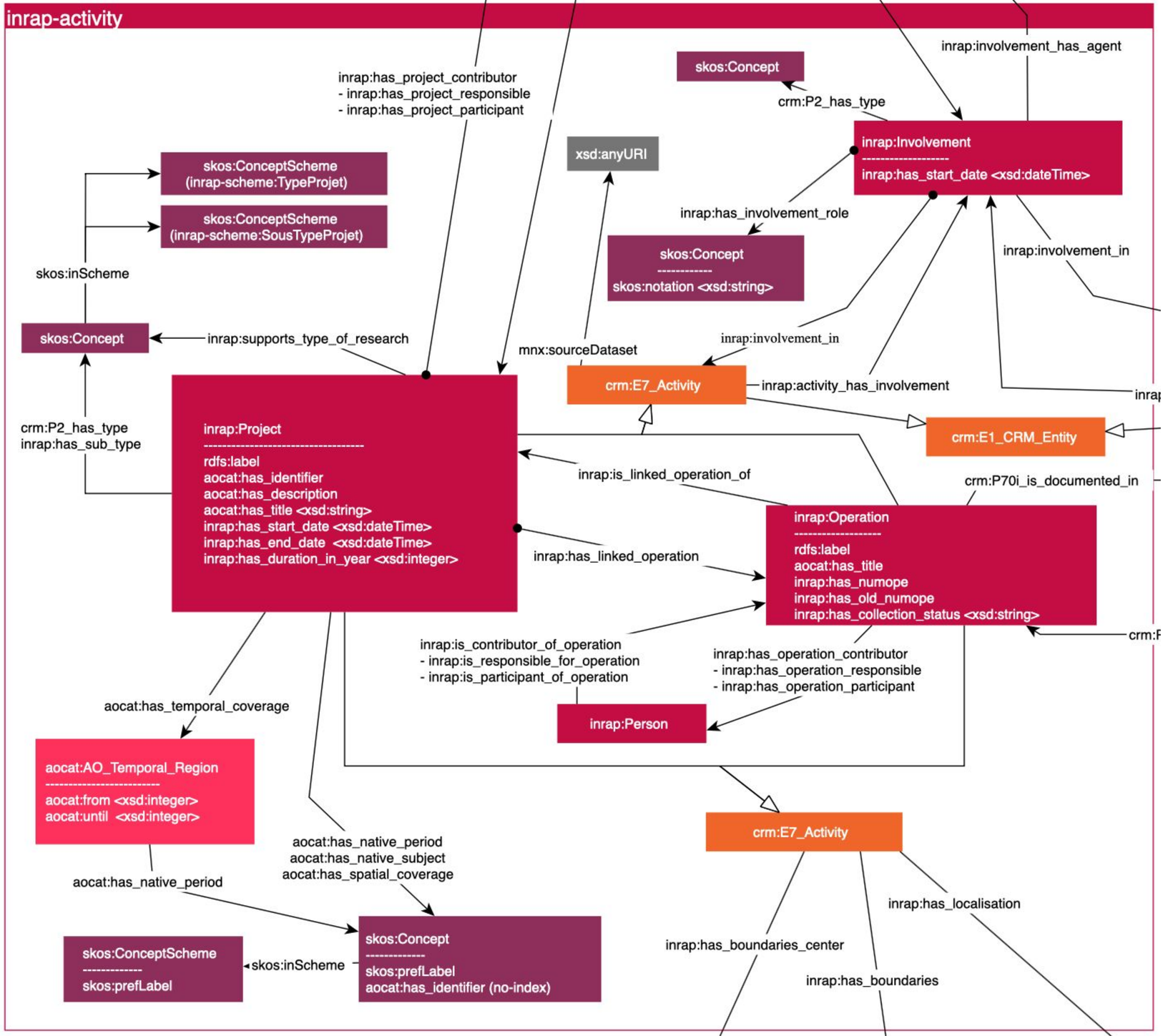


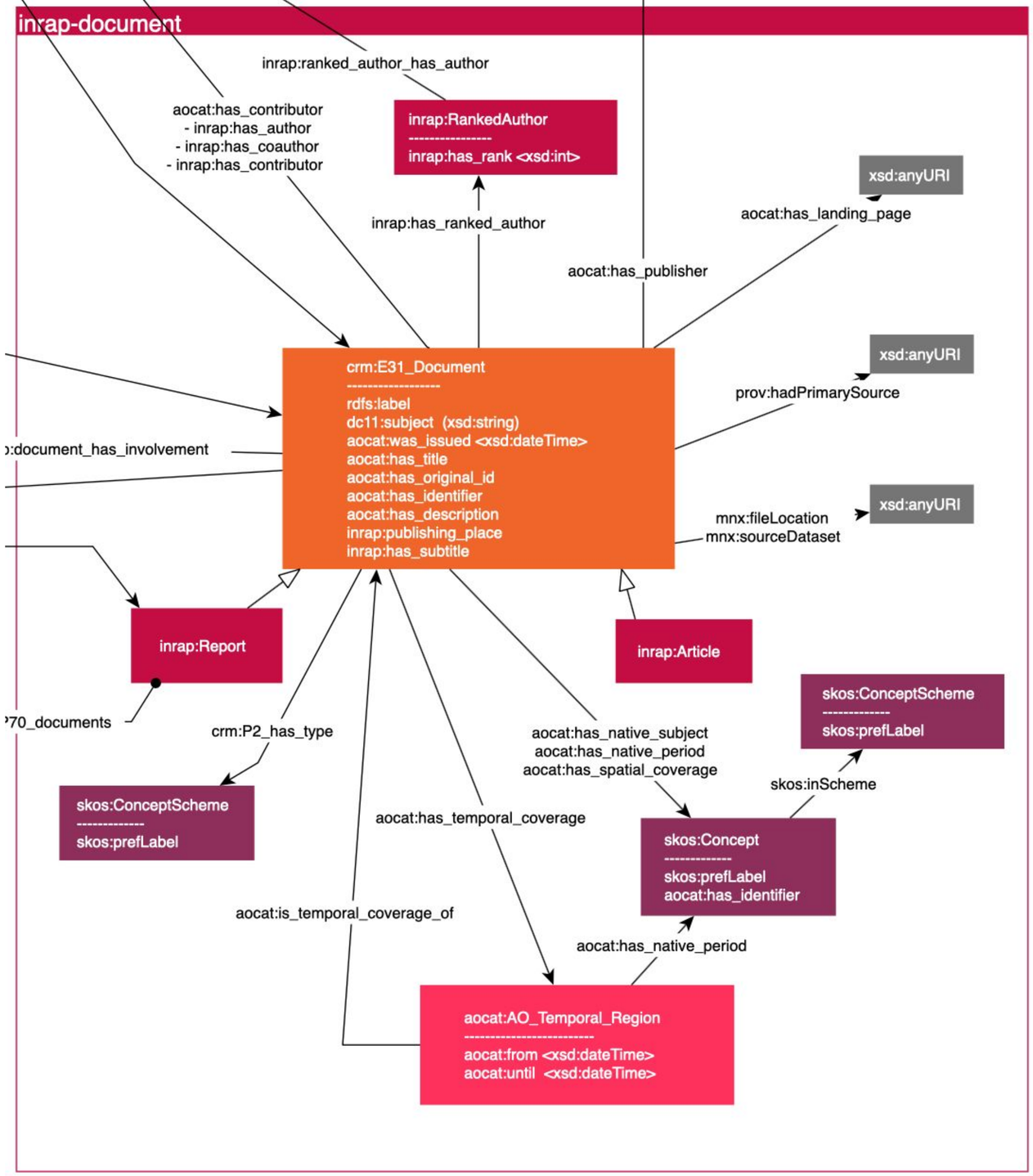


Modèle pivot - Agents

:probabl.
Own Your Data Science

Modèle pivot - Activity





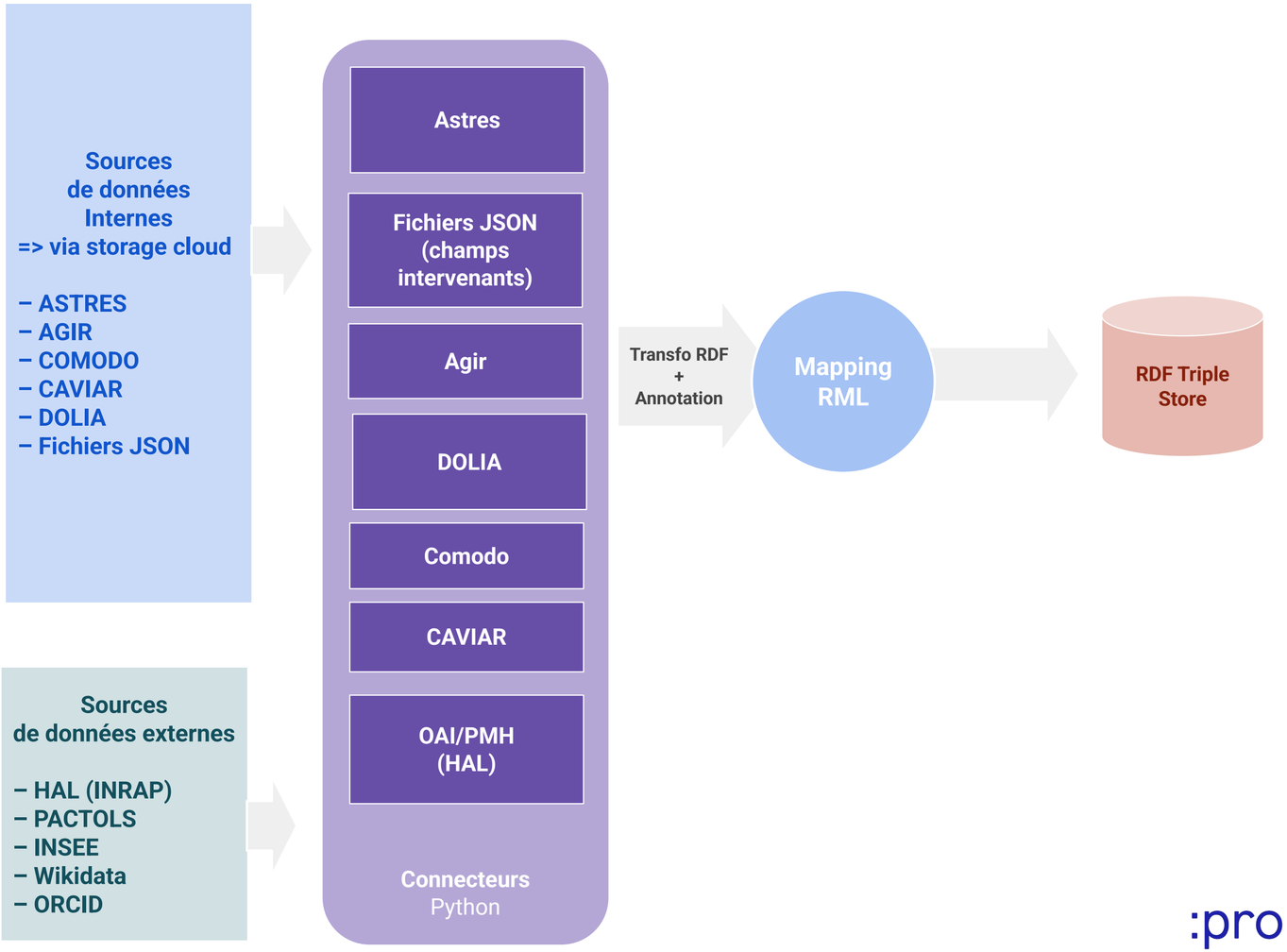
Modèle pivot - Document

:probabl.
Own Your Data Science

Data mapping graph generation

Transformation des données sources en graphe RDF

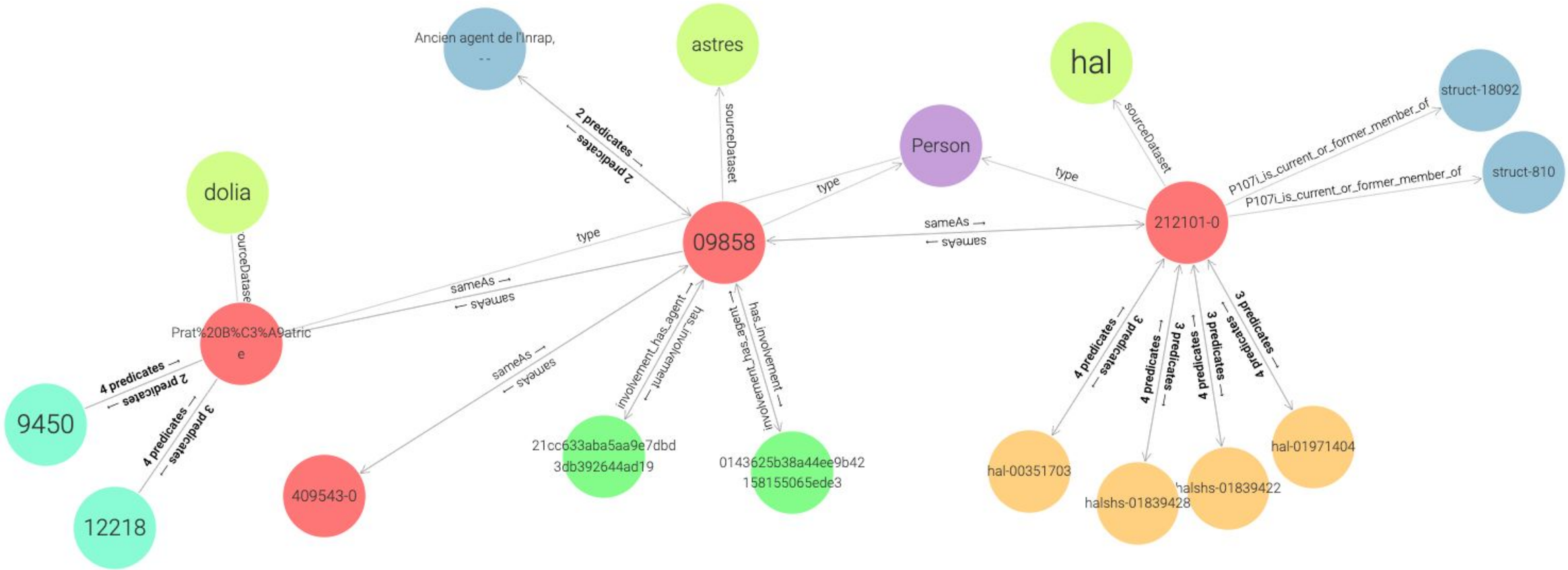
- Inrap exporte données CSV/JSON/XML
- Graph Connectors Synaptix transforme en RDF via schémas de mapping RML.io
- RDF chargés sur TripleStore
- Graphe enrichi via Analyse (cf plus bas)



:probabl.
Own Your Data Science

Présentation du KG Archipel

Exemples de décompositions des liens entre identité d'une même personne en graphe



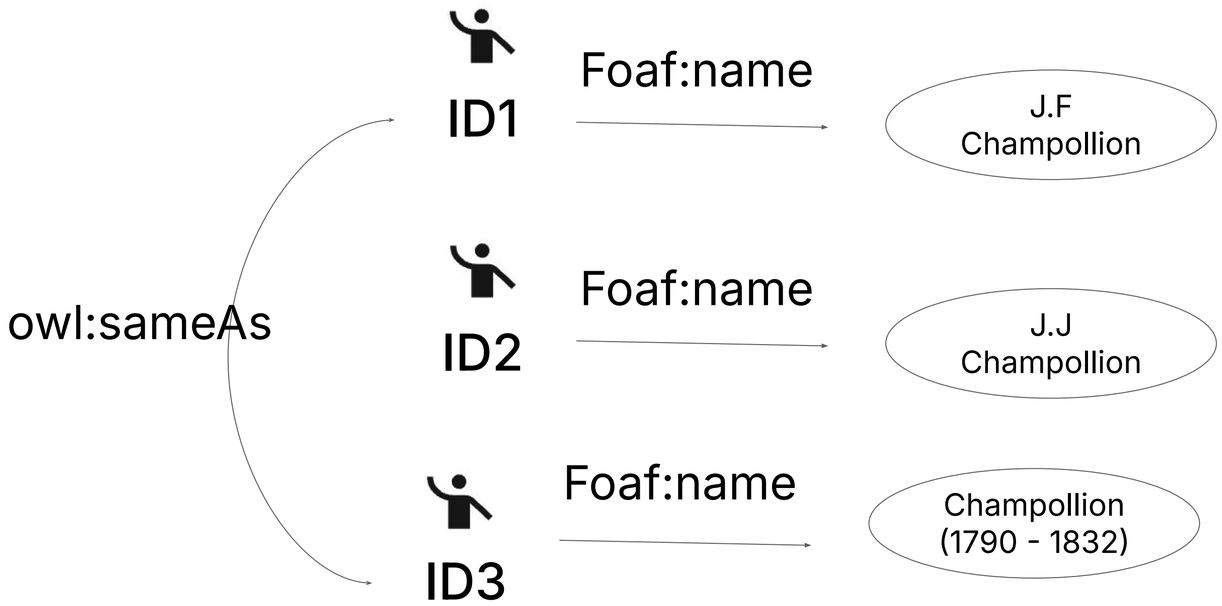
:probabl.
Own Your Data Science

Enrichissement du graphe

1. Alignement des personnes

- **Objectif :** rapprocher des entités représentant la même personne dans le graphe RDF

- **Approche :** utilisation de MinHashLSH (*min-wise independent permutation locality sensitive hashing scheme*) [1]



1 - Broder, Andrei Z. (1997), "On the resemblance and containment of documents", Compression and Complexity of Sequences: Proceedings, Positano, Amalfitan Coast, Salerno, Italy, June 11-13, 1997

Enrichissement du graphe

1. Alignement des personnes non supervisé - Méthodologie

- Pipeline MinHashLSH
  - ↳ **TextCleaner**: Jean-François Champollion (1790 - 1832) → jean françois champollion
  - ↳ **NGramTransformer** (n=3) : jean françois champollion →[jea],[jean],[ean],[an]...[ion]
  - ↳ **Vectorizer** : [160][1,0,0,1,....0]
  - ↳ **MinHashLSH**
- **Validation**
  - ↳ **Compromis** sur le choix du **seuil de similarité** :
    - Seuil trop haut : risque de faux négatifs (perte de vrais doublons)
    - Seuil trop bas : risque de faux positifs (mauvais appariements)
  - ↳ Choix basé sur des tests :
    - Plusieurs seuils comparés (0.75, 0.8, 0.9, 0.95)

Enrichissement du graphe

2. Recommandation d'entités similaires

Objectif

- Identifier des entités proches dans le graphe RDF

Feature Engineering

- Sélection des **features les plus pertinentes** pour la similarité
- Utilisation de TF-IDF
  - ↳ Limité aux 25 termes les plus fréquents.
  - ↳ **Fréquence minimale = 1** (faible recouvrement entre concepts ce qui évite de perdre l'information rare)

Pondération et Boosting

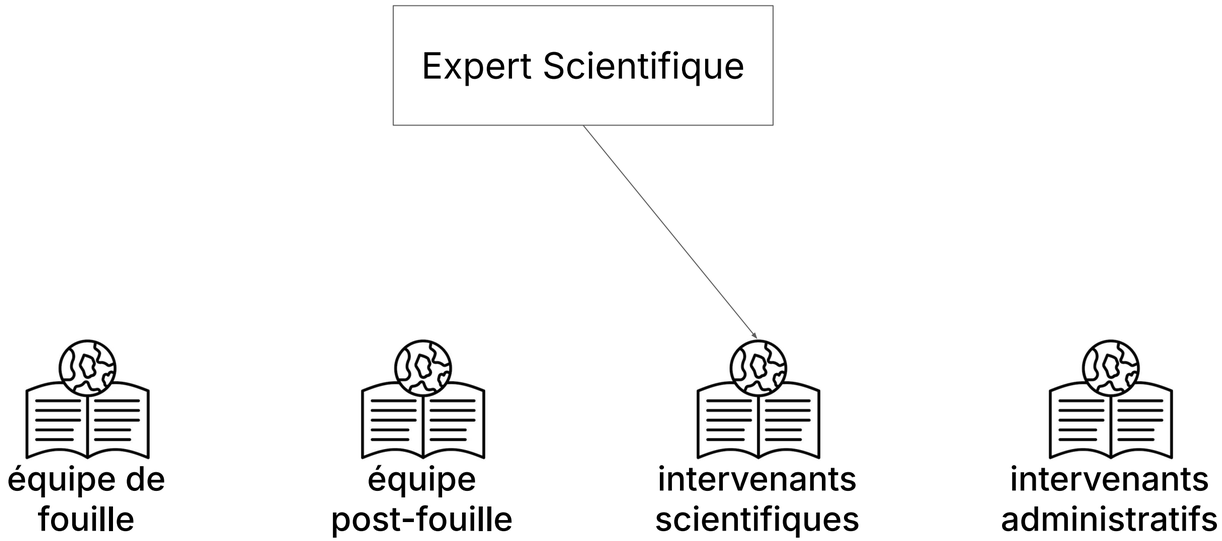
- Attribution de **poids différenciés** aux features.
- **Boost** sur certaines dimensions jugées discriminantes.

Enrichissement du graphe

3. Harmonisation des labels de concepts

Objectif

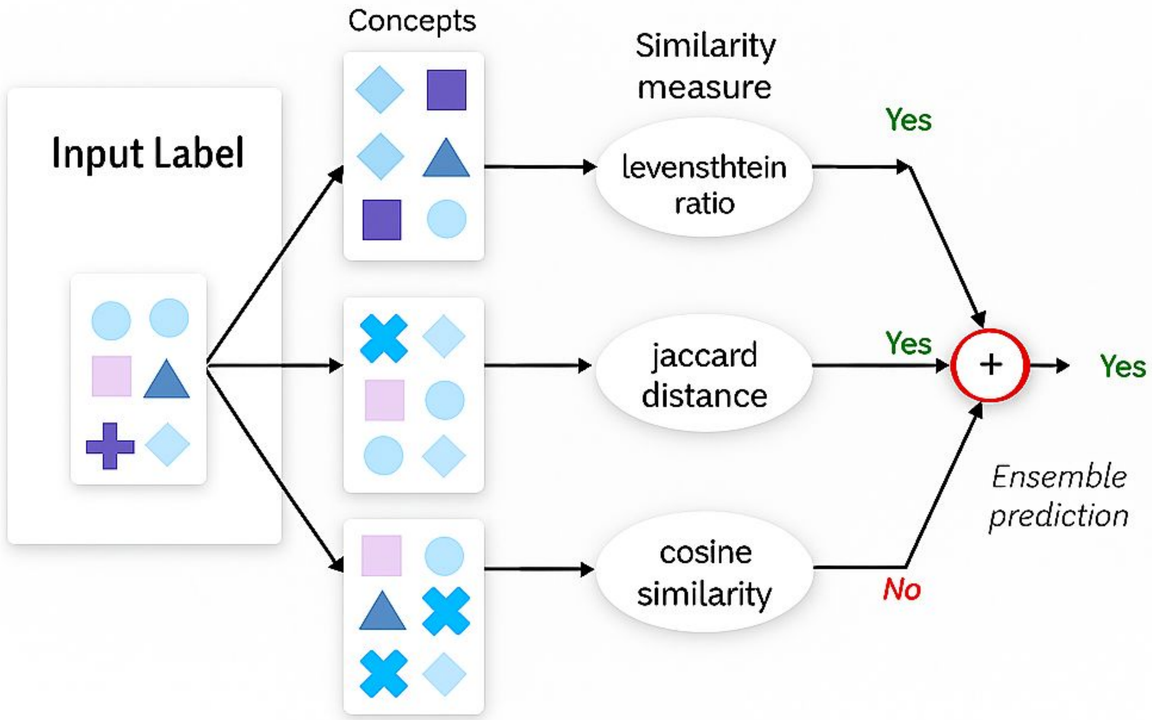
- Aligner des termes spécifiques issus du corpus avec les concepts normalisés d'une taxonomie



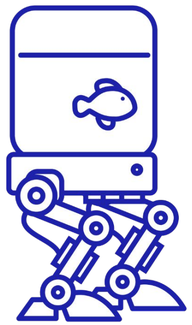
Enrichissement du graphe

3. Harmonisation des labels de concepts - méthodologies

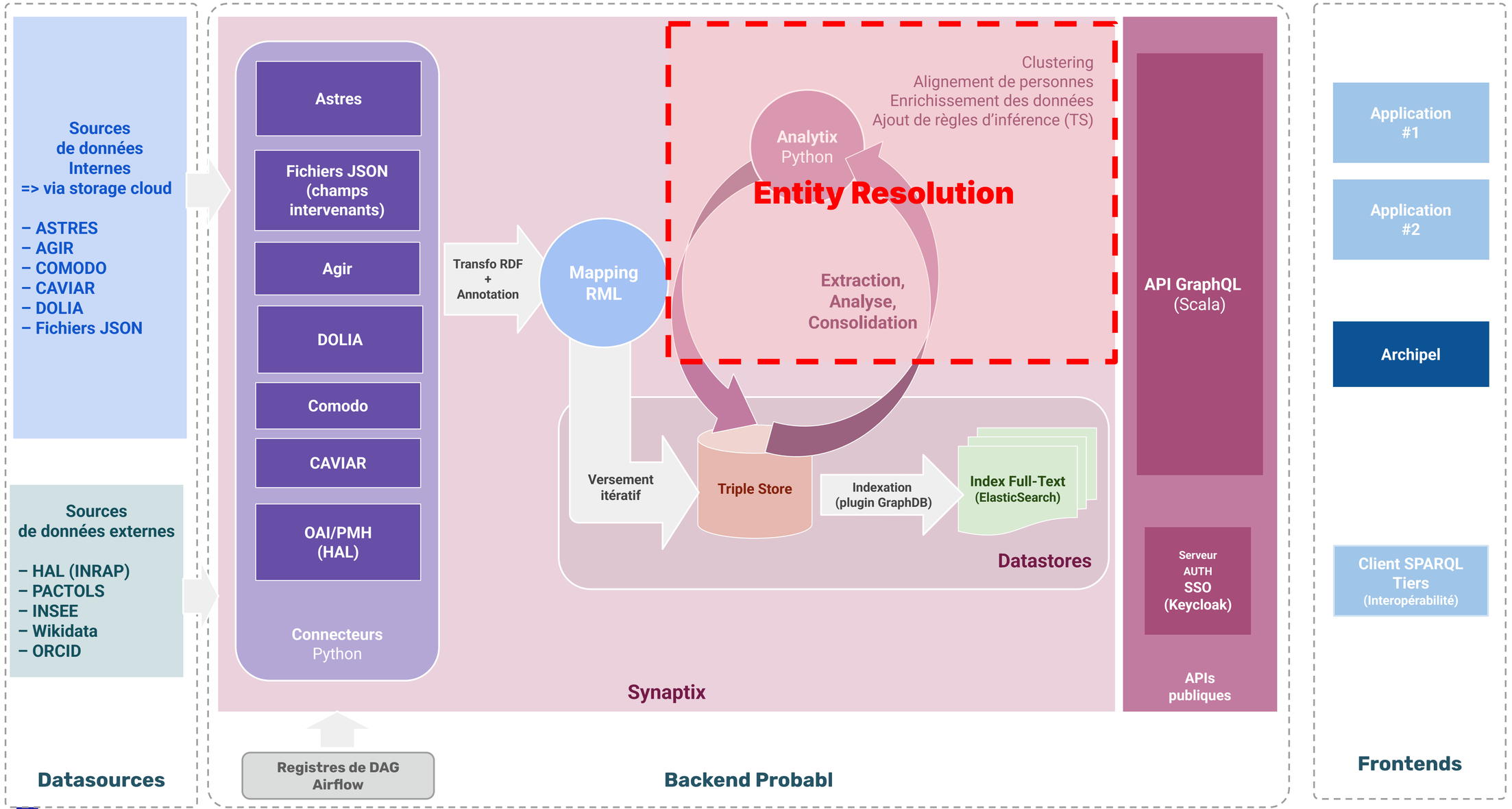
- Prétraitement des données
  - ↳ **TextCleaner**
- Rapprochement
  - ↳ Application d'une méthode d'ensemble ("Ensemble Method")
  - ↳ Calcul de 3 **mesure de similarité** pour chaque label :
    - **Levenshtein ratio**
    - **Similarité de Jaccard**
    - **Similarité du Cosinus** : similarité du cosinus entre embeddings générés avec all-MiniLM-L6-v2 (transformer Hugging Face)



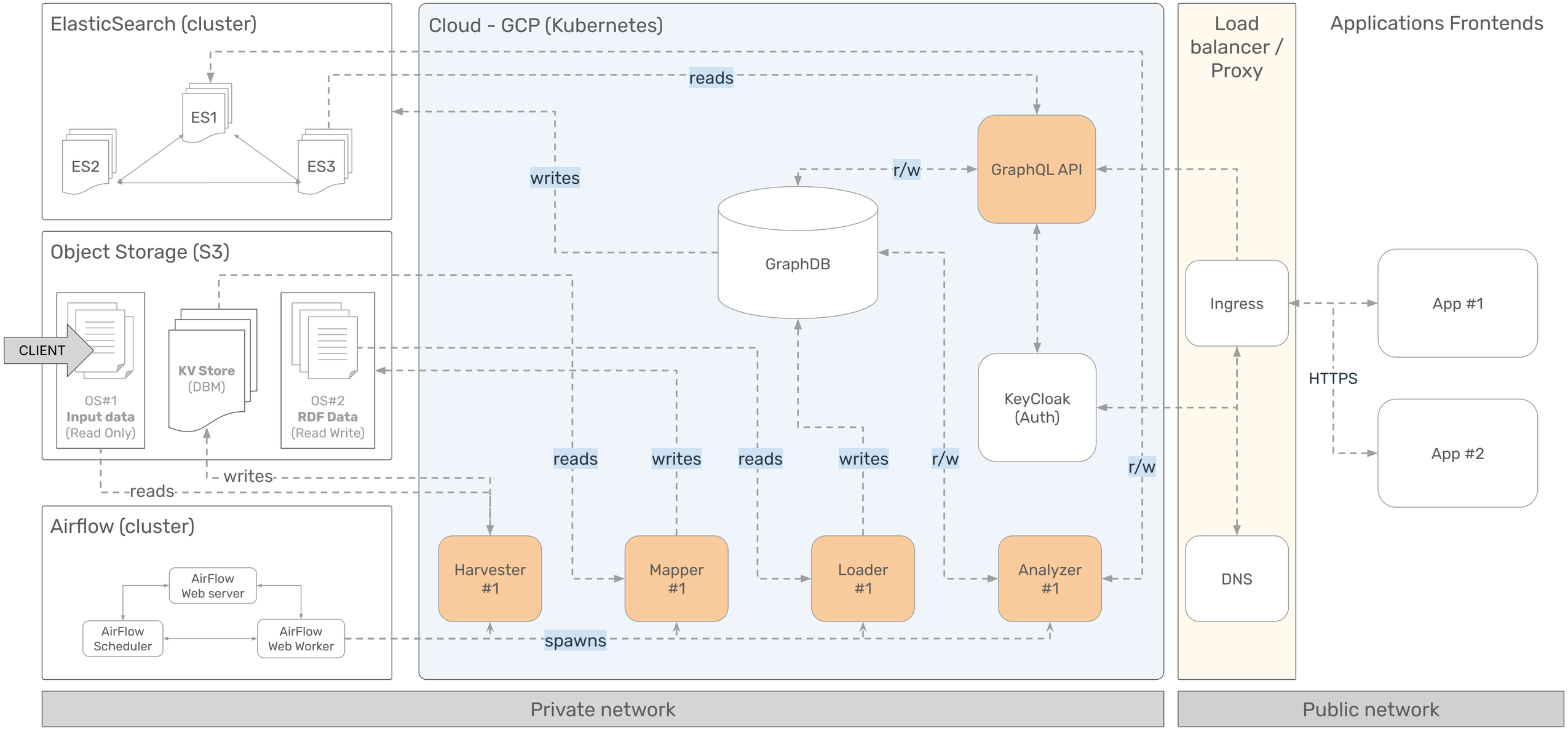
Architecture technique & fonctionnelle



Archipel - Vue fonctionnelle

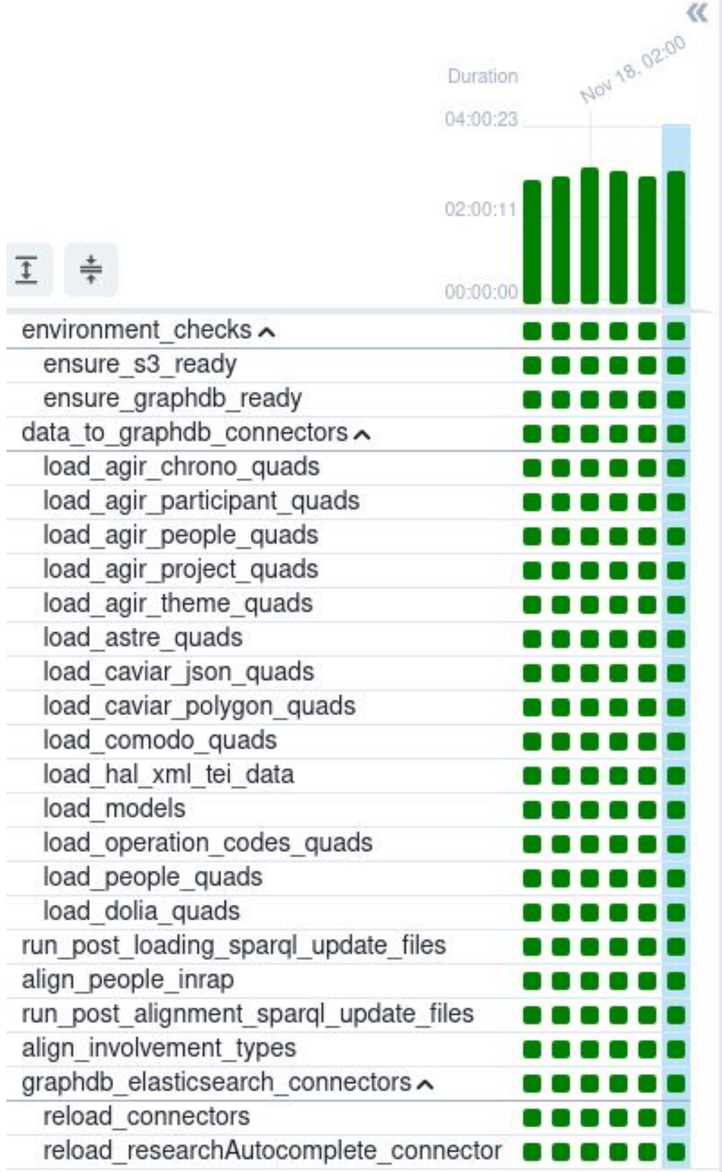


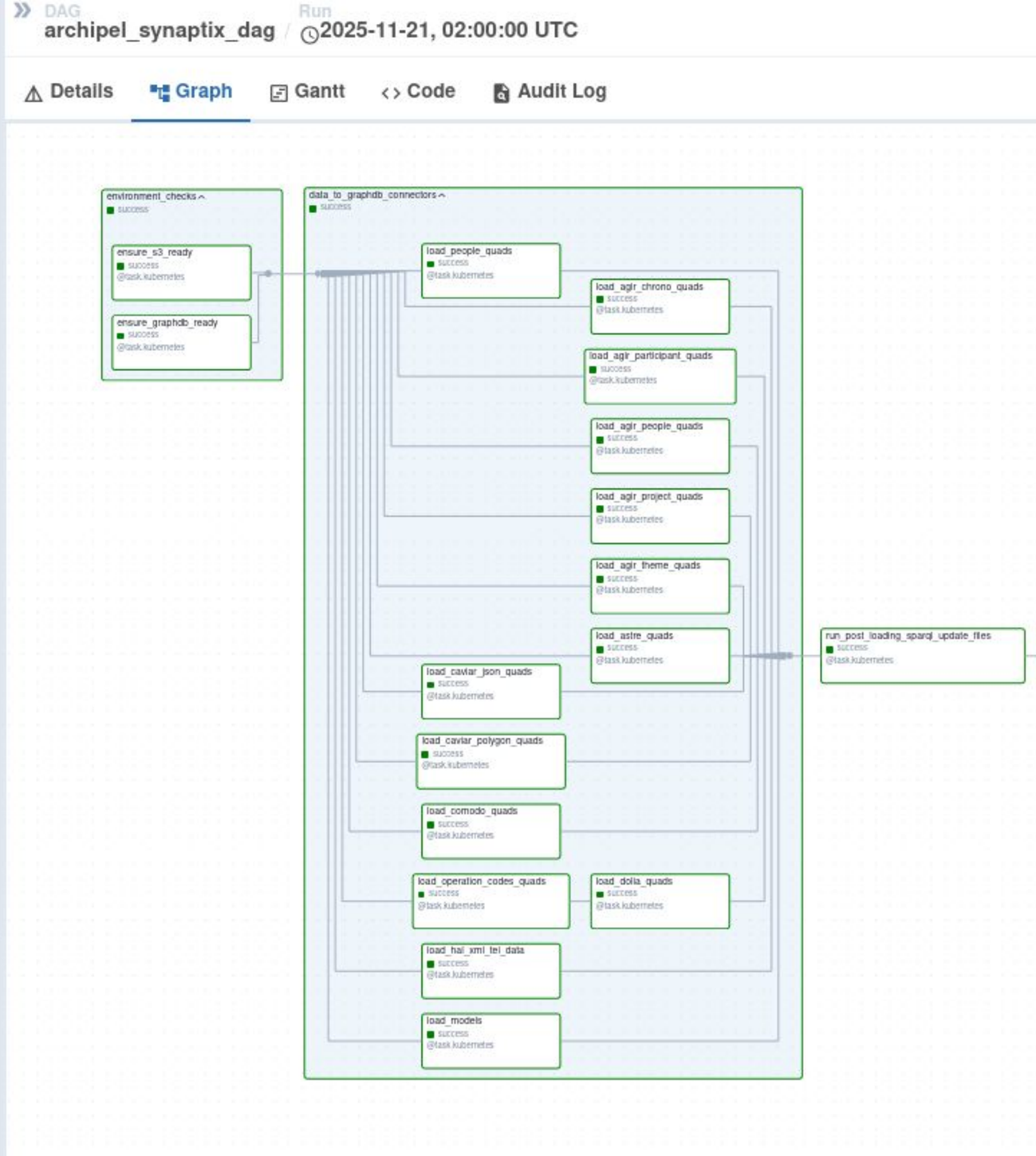
Architecture technique



Airflow pour le pilotage et monitoring

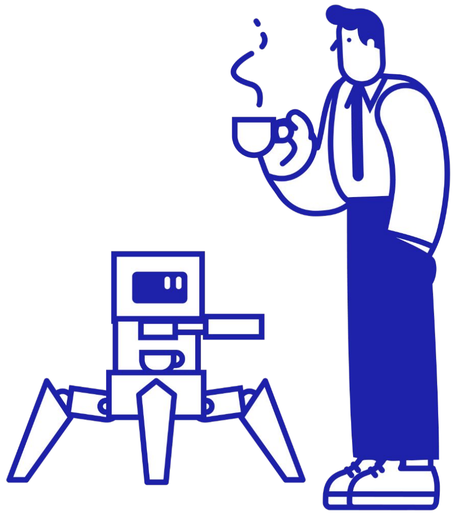
https://airflow.apache.org





Conclusion

- Archipel permet d'explorer la production archéologique de l'Inrap ⇒ https://archipel.inrap.fr/
- Archipel s'appuie sur Synaptix
  - ↳ une approche de l'urbanisation de SI
  - ↳ exploitant un graphe RDF comme pivot entre données et connaissances
  - ↳ un framework open source
    - ⇒ https://gitlab.com/probabl/services/synaptix



Questions ?

https://hello.probabl.ai/contact-us

Freddy Limpens

- freddy@probabl.ai
- +33 6 65 07 64 35

Nicolas Delaforge

- nicolas@probabl.ai
- +33 6 74 93 26 52

:probabl.

Own Your Data Science

Back-Up slides

:probabl?

Own Your Data Science

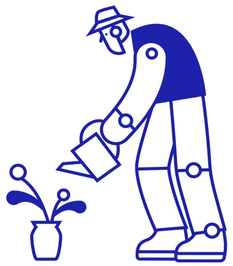
Open source

Open science

Souveraineté Numérique

Inspired by RedHat

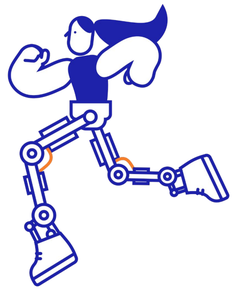
Une entreprise à mission



“Développer, maintenir à l'état de l'art et pérenniser un ensemble complet d'outils open source pour la science des données au bénéfice de la France, de l'Union Européenne et du Monde.” *(extrait des statuts Probabl.)*

- Créée dans le cadre de la Stratégie Nationale IA (France 2030, 24M€ d’investissement sur 5 ans)
- Spin off Inria
- Gouvernance tripartite : 33% public, 33% privé, 33% co-fondateurs et employés

Expertise reconnue internationalement

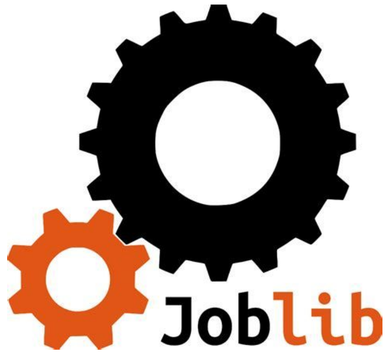


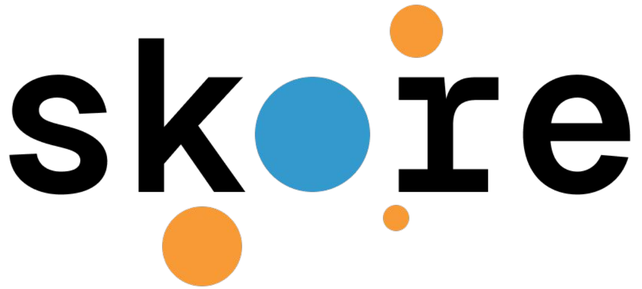
- Principaux contributeurs de Scikit Learn (Bibliothèque mondialement connue de Machine Learning, ~2 millions d’utilisateurs)
- Contribution aux initiatives de standardisation ouvertes du W3C
- Experts reconnus mondialement (chercheurs, enseignants, conférenciers)
- Collaborations avec les plus grands fabricants de hardware (Intel, Nvidia, AMD, Ampere)

W3C

scikit learn

Nos Marques OSS





Enrichissement du graphe

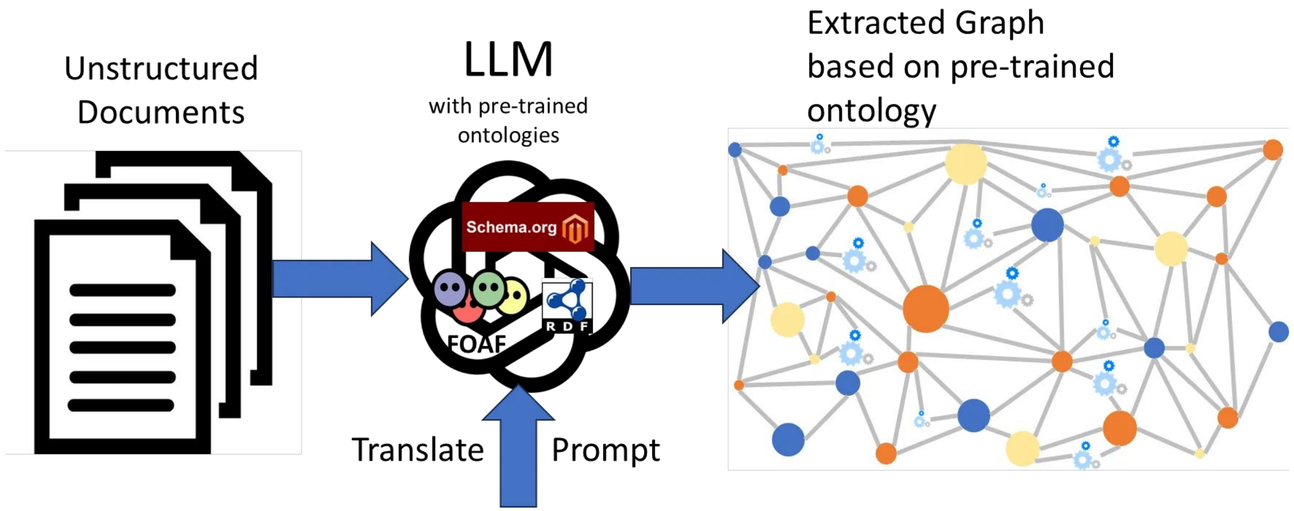
1. Alignement des personnes - Amélioration par le contexte du graphe

- Utilisation **d'heuristiques** en s'appuyant sur le graphe RDF
  - **Identifiant externe**: Si deux personnes ont un **même identifiant externe**, cela valide le rapprochement
  - **Entités voisines communes** : Si elles apparaissent dans des **projets communs**, partagent des **co-auteurs** cela améliore la confiance dans l'alignement
  - **Caractéristiques inférées** : Génération de nouvelles données en exploitant les liens dans le graphe. Par exemple : Transitivité ⇒ Exploiter les dates ou les localisations associées aux entités liées aux personnes pour les rattacher directement aux personnes et les utiliser pour valider un alignement.

“Structure the Unstructured”

Giving structure and ML capabilities to any kind of data

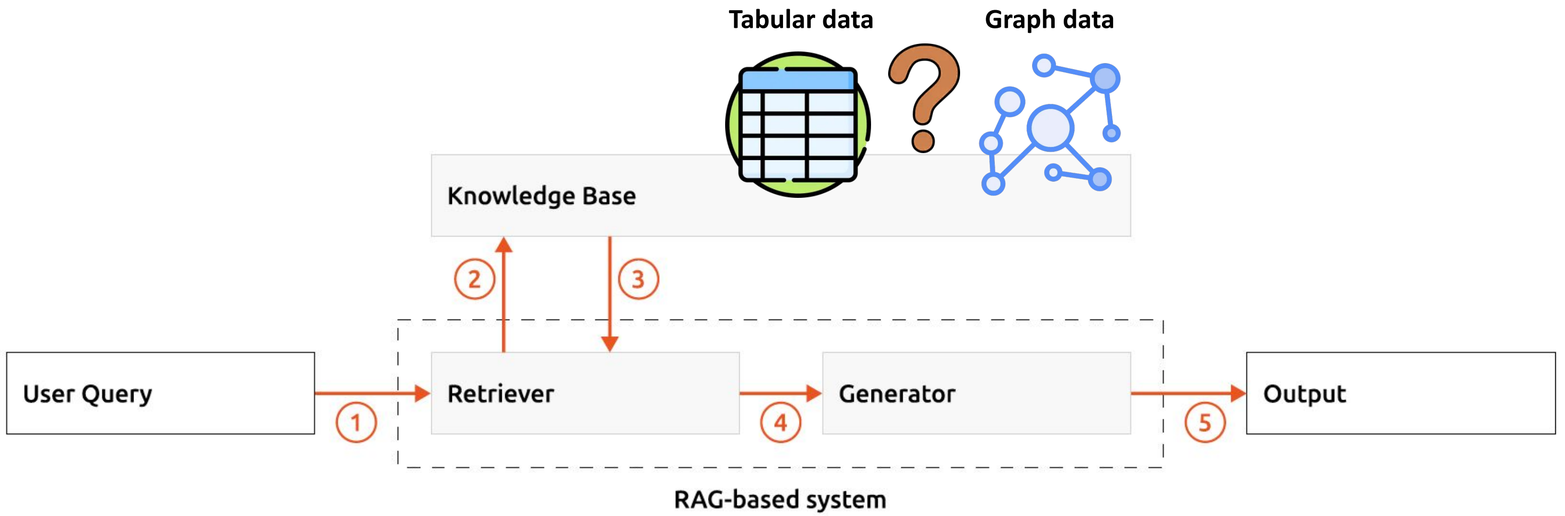
- Historically, Knowledge Graph generation used to be a complicated process and somehow fastidious to maintain.
- Today, with the power of LLMs it has never been so easy to extract information from unstructured data and put it into a knowledge graph.



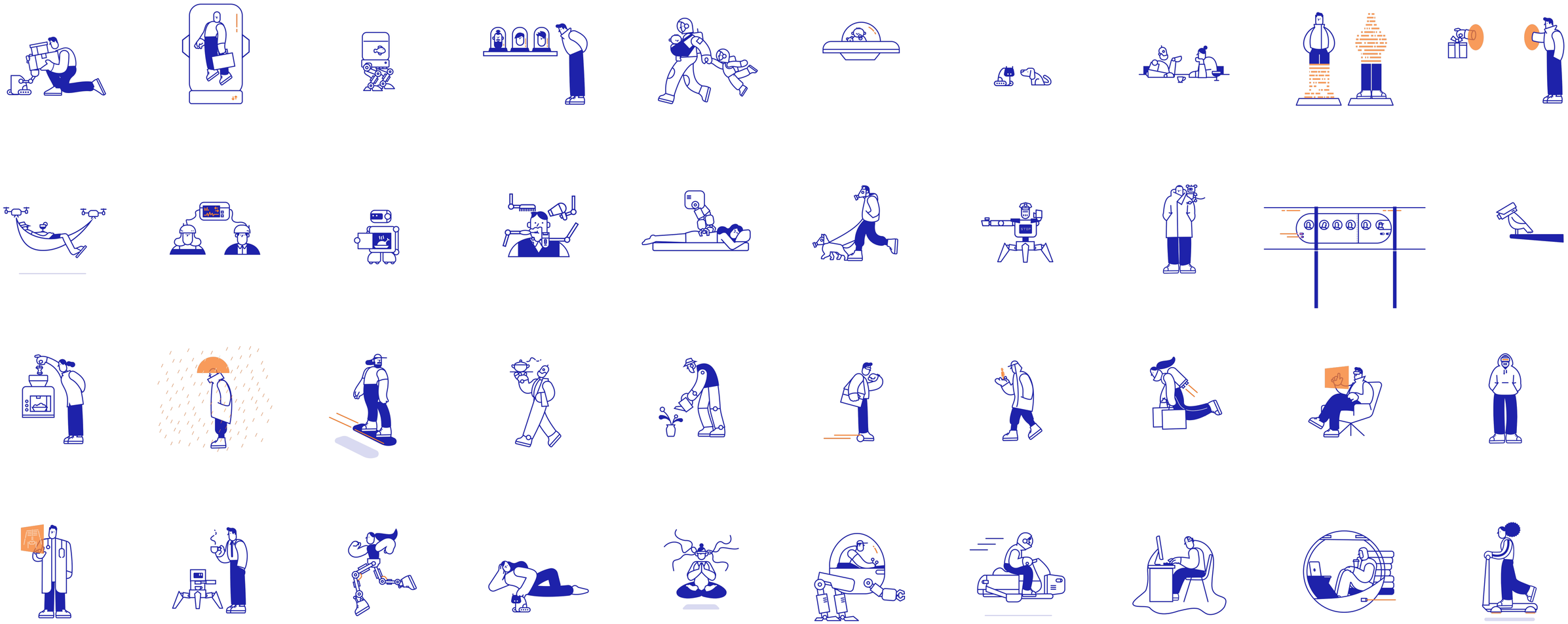
Neuro-semantic GraphRAG

Enrich your GenAI with a Knowledge Graph

- The RAG (Retrieval Augmented Generation) is a technique used to provide “expertise” to a LLM which is basically trained on generic data.
- This mainly consists of creating a hook between the prompt and the generator in order to enrich the context of the query.
- From there, any kind of data can be used to enrich the prompt, although some researchers have proved that knowledge graphs provide a richest context and give a better explainability.



Official Icons



Official Logos

:probabl.

:probabl.

:probabl.

:probabl.