

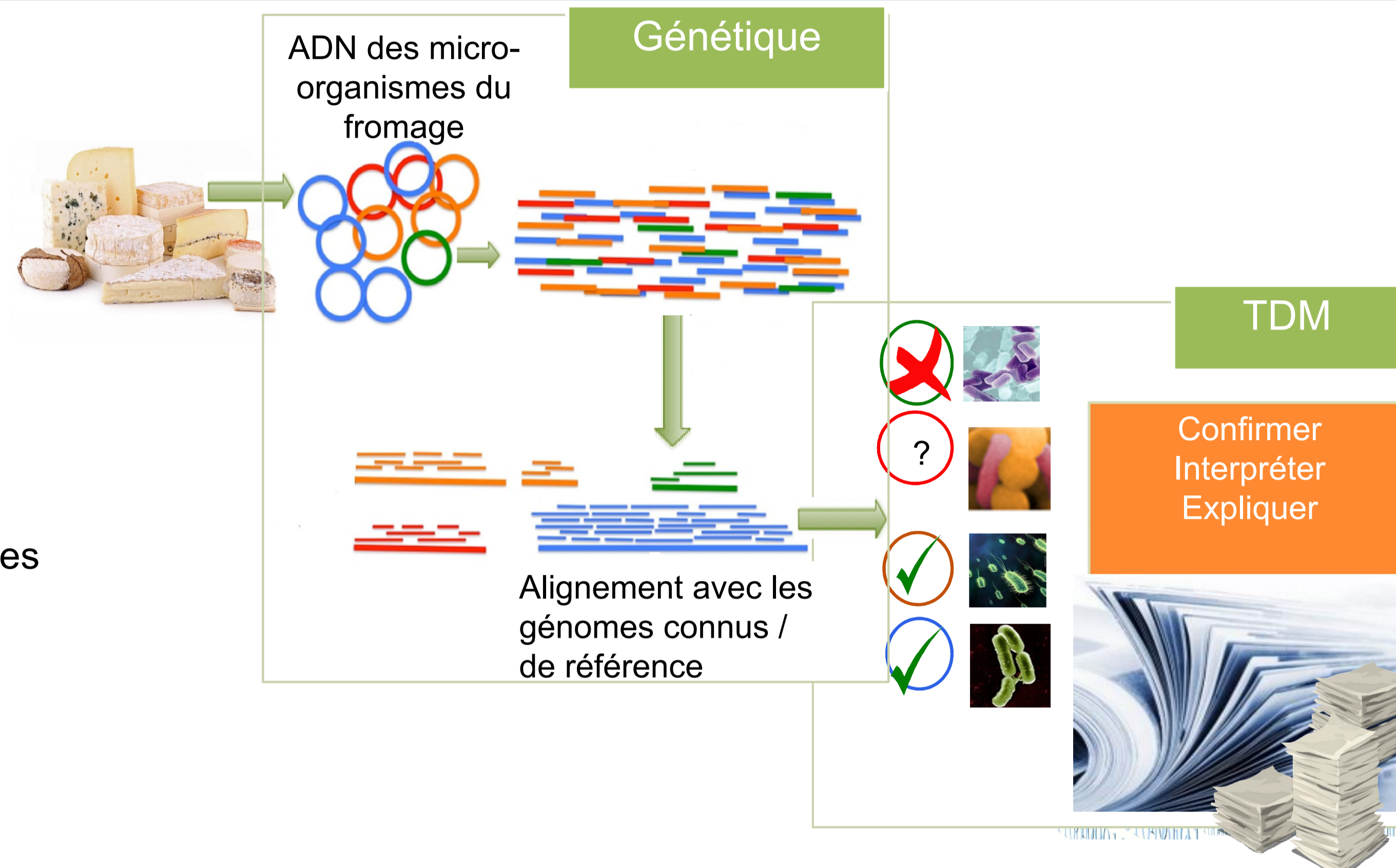
La suite Alvis, plateforme d'analyse de contenu textuel, et son intégration dans OpenMinTeD

Mouhamadou Ba, Robert Bossy, Claire Nedellec
 Équipe Bibliome, Département MaLAGE, INRA – Jouy-en-Josas, France
 mouhamadou.ba@inra.fr

Introduction

La littérature scientifique constitue un gisement de connaissances scientifiques de grande valeur, mais largement inexploité parce qu'uniquement sous forme textuelle. La croissance très rapide du nombre de publications au niveau mondial rend impossible une veille scientifique manuelle. C'est pour permettre de fouiller ces données que la suite Alvis et le projet OpenMinTeD (H2020 E-INFRA 2016-2018) existent. La suite Alvis offre des méthodes et des technologies pour l'extraction d'information et de connaissances en Sciences de la Vie. Le projet OpenMinTeD vise à faciliter l'utilisation des solutions existantes de fouille de textes et de données (TDM) telles qu'Alvis.

Identifier des microbes par leur ADN



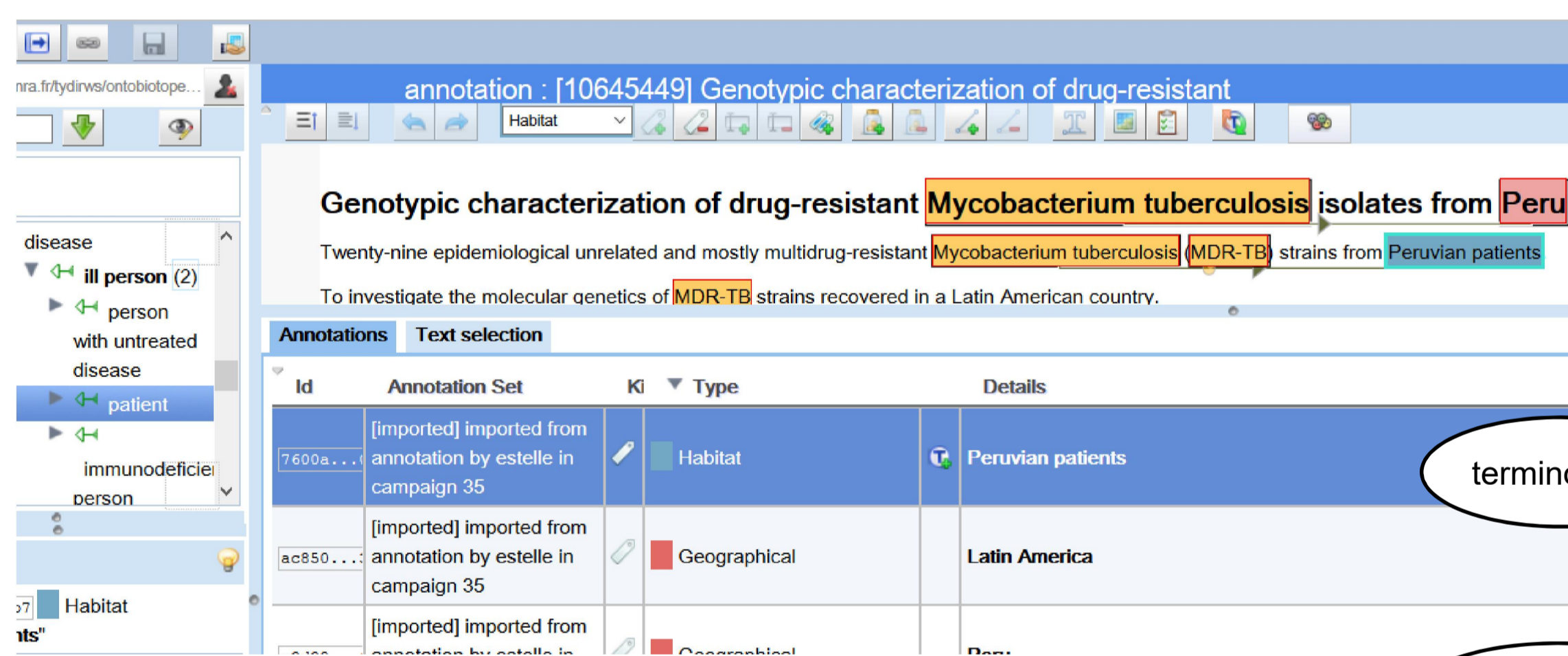
Exemple d'application de méthodes de fouille de textes avec la suite Alvis

Exemple d'application, à rendre plus accessible et facilement utilisable par la communauté grâce à OpenMinTeD

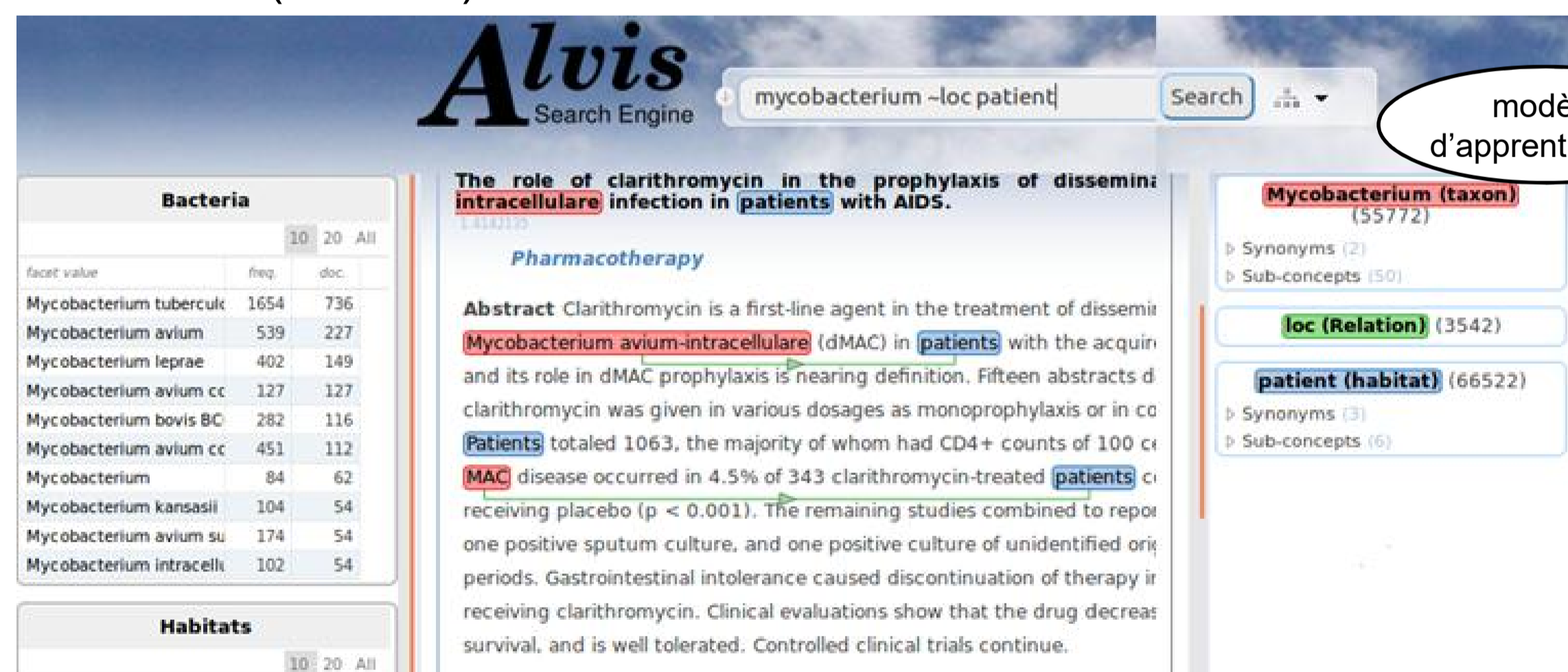
Suite Alvis

Un panel d'outils et de solutions pour l'extraction de connaissances par le TAL (Traitement Automatique de la Langue) et la fouille de textes

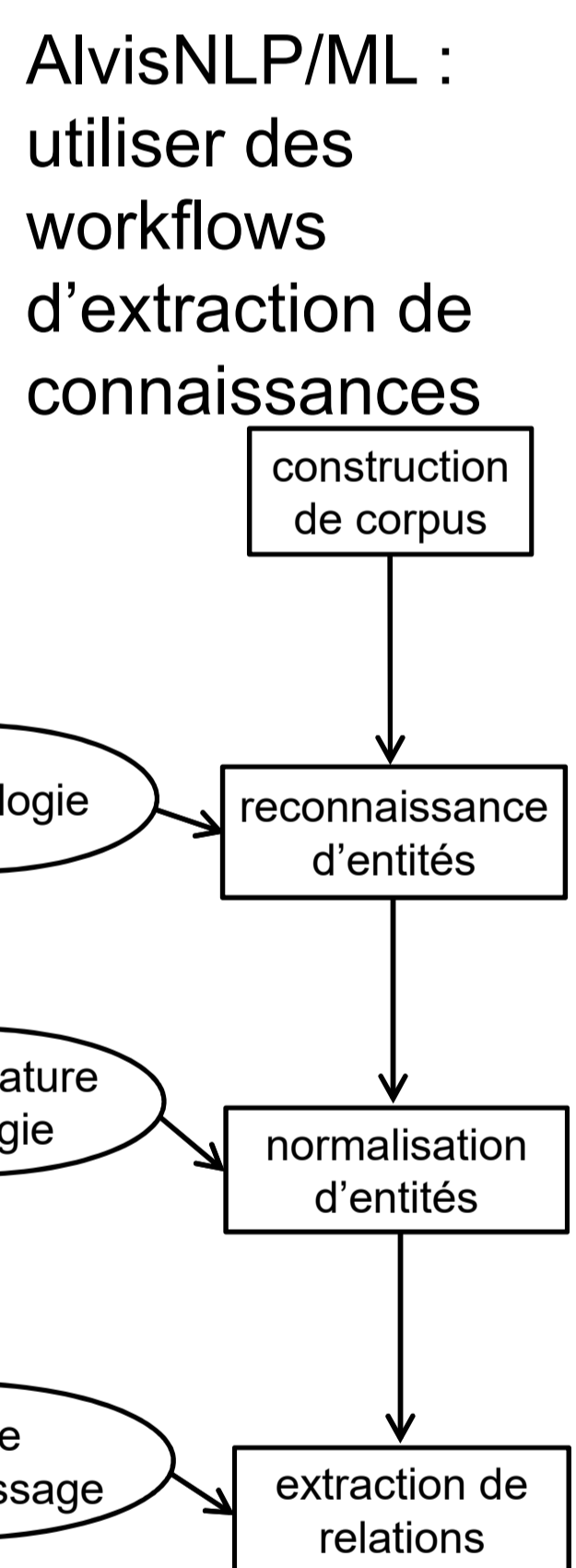
- documents scientifiques textuels
- domaines de l'agriculture et de la biodiversité



Annoter de manière expressive des corpus textuels par des utilisateurs finaux avec utilisation d'ontologies adaptées aux domaines (AlvisAE)



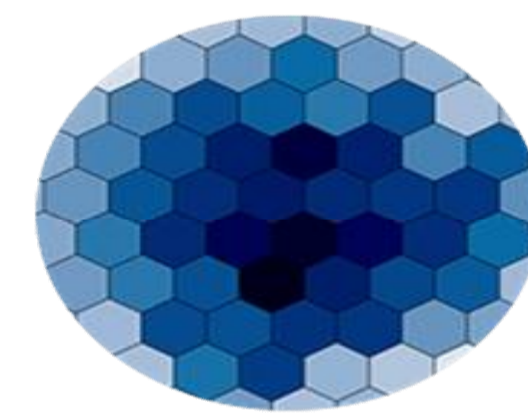
Rechercher à travers des corpus annotés en utilisant AlvisIR, un moteur générique d'indexation et de recherche sémantique



OpenMinTeD

- Pour une infrastructure collaborative de fouille de textes et de données
- Intégration des solutions existantes de fouille de textes et de données (Alvis, UIMA, Gate, Argo, DkPro, Core, OpenAire, etc.).
 - Public cible : ingénieurs, chercheurs dans divers domaines

COMMUNITY DRIVEN APPLICATIONS



RESEARCH ANALYTICS (ARC)



LIFE SCIENCES (CNIO)



AGRICULTURE & BIODIVERSITY (INRA)



SOCIAL SCIENCES (UKP-TUDA)

Définition d'un *framework* pour l'interopérabilité

- les besoins (ressources, licences, description, représentation, implémentation, etc.)
- les guidelines et la documentation

Développement des outils de base et gestion des ressources matérielles

- registre de services
- moteur de workflows
- service d'annotations
- bases de connaissances
- service de gestion des utilisateurs

Intégration : OpenMinTeD propose les règles, les partenaires s'approprient...

OpenMinTeD	Schéma unique et uniforme de description des ressources	Protocoles et formats d'échange des données	Spécification d'interfaces pour les modules (Standalone, Services Web, Docker)	Définition uniforme des workflows (framework Galaxy)	Registre de ressources, Moteur de workflows, Service d'annotations, Base de connaissances	Espaces personnels (Identification, authentification, autorisation, partage, accès)
	méta-données	données	modules	workflows	applications	utilisateurs
Alvis	Description et déclaration des ressources d'Alvis (corpus, ontologies, modules, etc.)	Prise en charge des formats et protocoles (AERO, RDF, XMI). Préparation des données	Docker pour les modules d'AlvisNLP/ML. Gestion des dépendances par maven et conda	Adoption de la solution Workflow de Galaxy. Mapping de données avec UIMA	Déploiement des applications clientes (ex. microbes, blé, arabette)	Prise en main et test avec des partenaires de Bibliome qui sont les développeurs des applications clientes