Extraction d'information pour la reconstruction des réseaux de régulation biologiques impliquées dans le développement de la graine chez *A. Thaliana*

Dialekti VALSAMOU

sous la direction de Claire Nédellec (INRA-MIG)
et Pierre Zweigenbaum(LIMSI-CNRS)
en collaboration avec Bertrand Dubreucq et Loïc Lepiniec (INRA/IJPB)

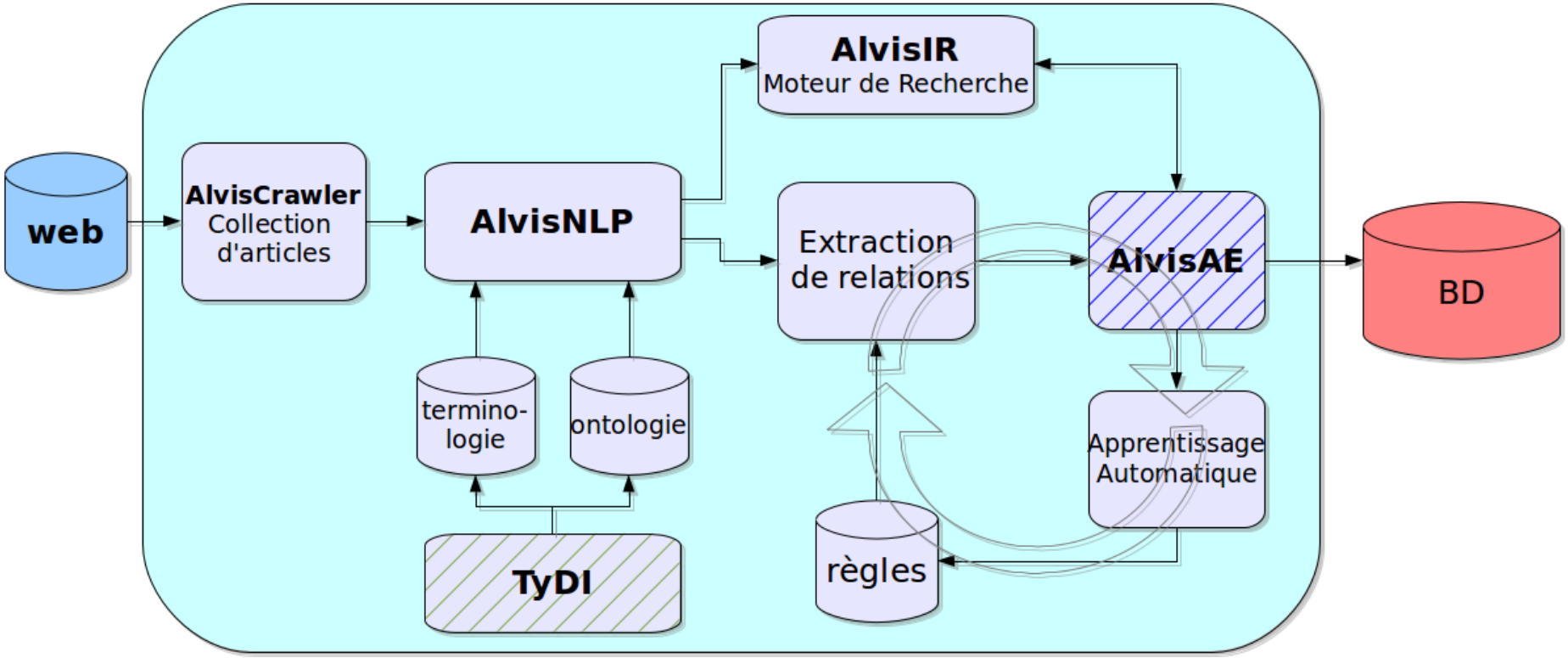
Intuition

- Idées clés
  - « Lecture » automatique des publications scientifiques
  - Apprentissage automatique
  - « Extraction » du réseau, aide à la lecture
  - Complémentaire aux données expérimentales
- Enjeu de l'étude du développement de la graine chez *A.Thaliana*
  - compréhension du fonctionnement biologique avec multiples applications
  - espèce modèle : généralisation + disponibilité de données

Composantes

- Collection de documents
- Modèle formel
  - Schéma d'annotation
  - Document de *Guidelines*
- Extraction d'information
  - Traitement Linguistique
  - Apprentissage automatique
- Reconstruction du réseau / Visualisation

Architecture du système d' EI

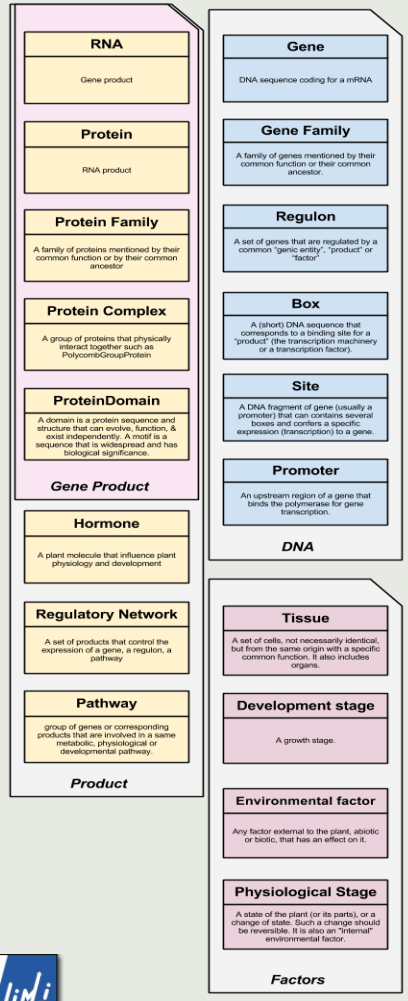


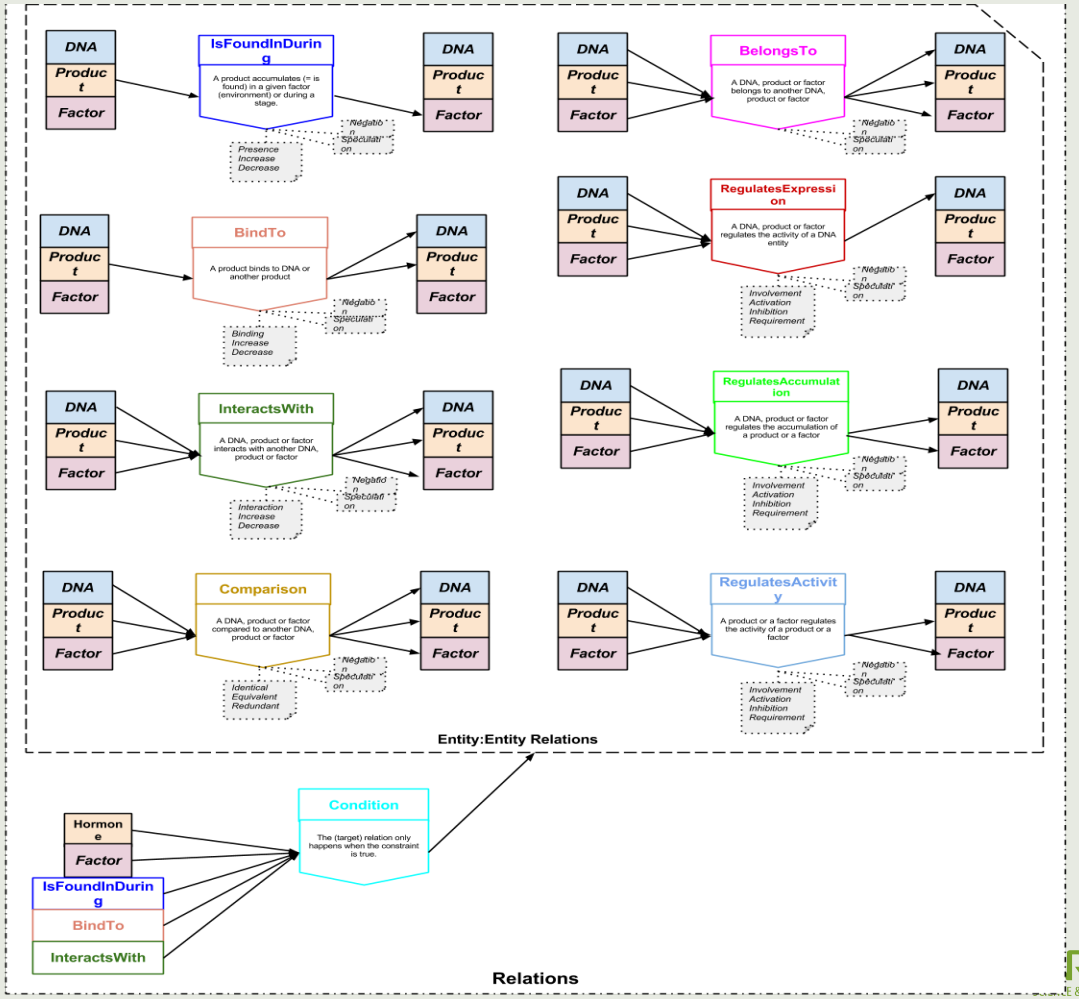
Modélisation

- Représentation formelle de la connaissance biologique
- Travail multi-experts
- Cas d' *A.Thaliana* : modèle riche

- *Idée* : Données expérimentales + résultat de l'Extraction d'Information: « image complète »

Modélisation: Schéma





Détail du Schéma

Gene

DNA sequence coding for a mRNA

Gene Family

A family of genes mentioned by their common function or their common ancestor.

Regulon

A set of genes that are regulated by a common "genic entity", "product" or "factor"

Box

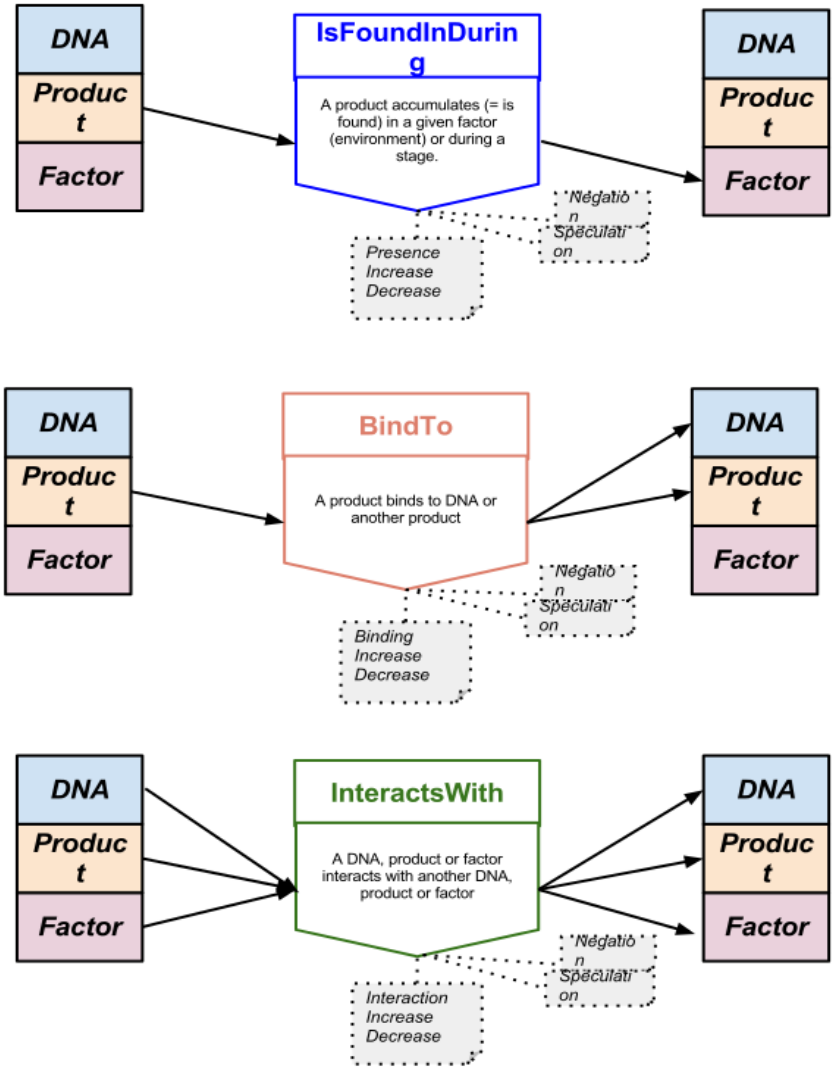
A (short) DNA sequence that corresponds to a binding site for a "product" (the transcription machinery or a transcription factor).

Site

A DNA fragment of gene (usually a promoter) that can contains several boxes and confers a specific expression (transcription) to a gene.

Promoter

An upstream region of a gene that binds the polymerase for gene



Modélisation : *Guidelines*

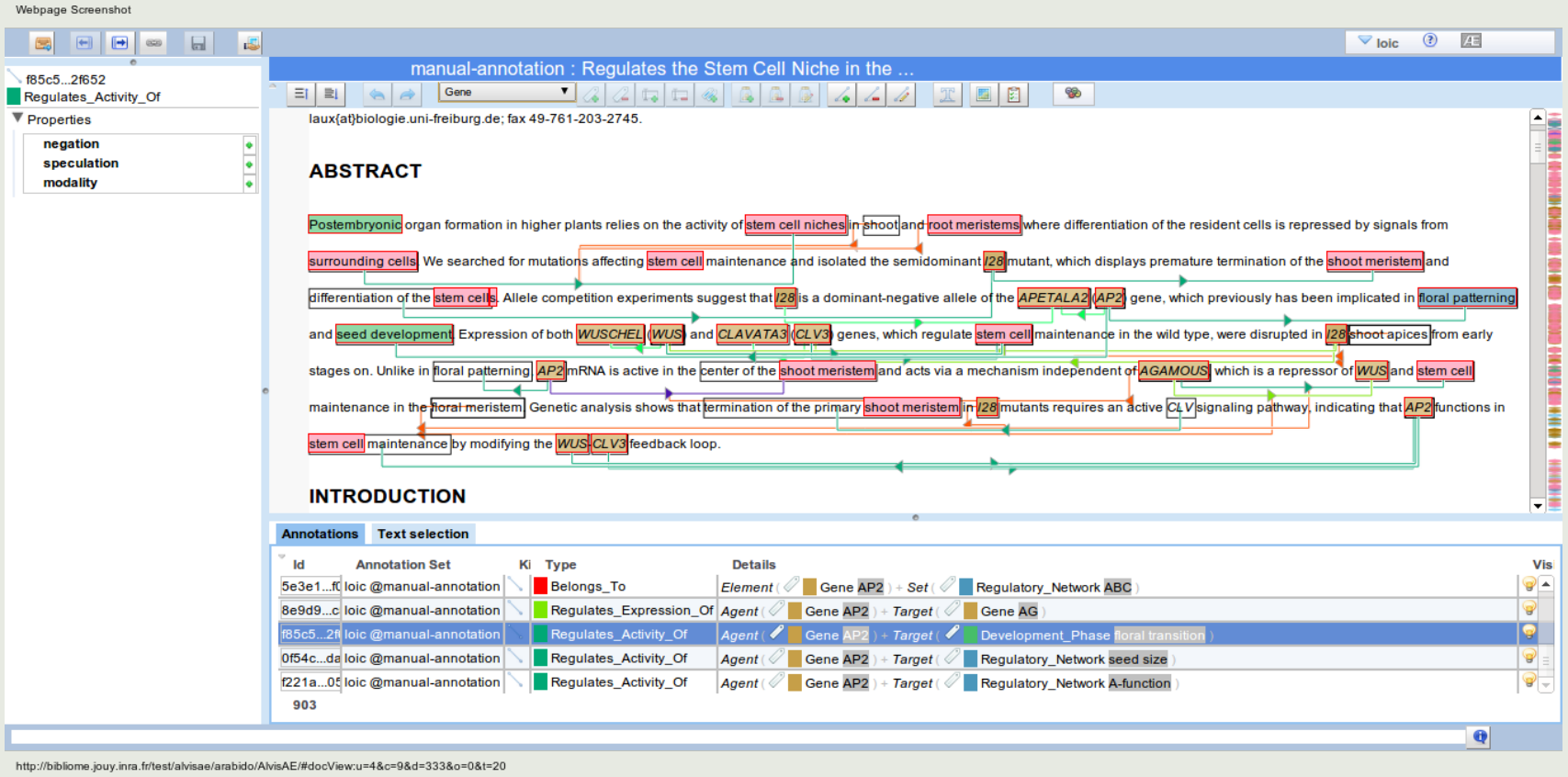
- Relie texte – modèle
- Produit d'un consensus
- Document unique de référence
  - Définitions complètes
  - Exemples
- ici :
  - Riche
  - Disponible en ligne

Annotation

- Homogénéité, cohérence
- 15-30' par ½ page *
- Deux experts
  - Accord entre deux annotateurs
  - Cohérence temporelle d'un annotateur

* * © B. Dubreucq 2013

Annotation : éditeur



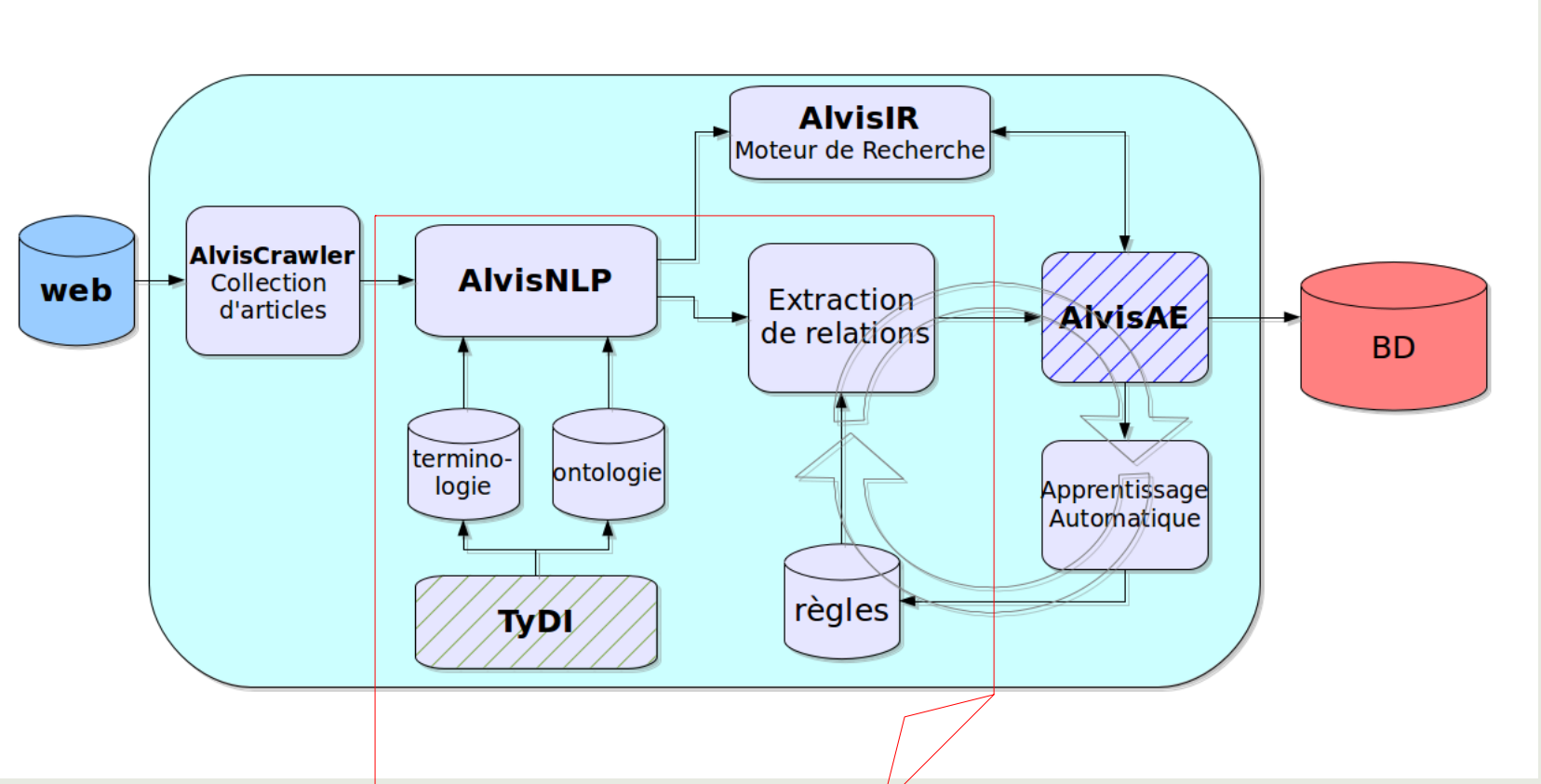
Défis linguistiques

- Traitement de la métonymie
  - Ex. produits de gènes vs gènes
- Expressions, paraphrases
- Anaphores
- Conditions, hypothèses, négation
- Richesse d'information → complexité linguistique

Extraction d'Information : Tâches

- Reconnaissance d'entités nommées
  - terminologie
  - normalisation
- Résolution d'anaphores
- Extraction de relations
  - binaires ou n-naires
  - relation syntaxique ↔ relation sémantique

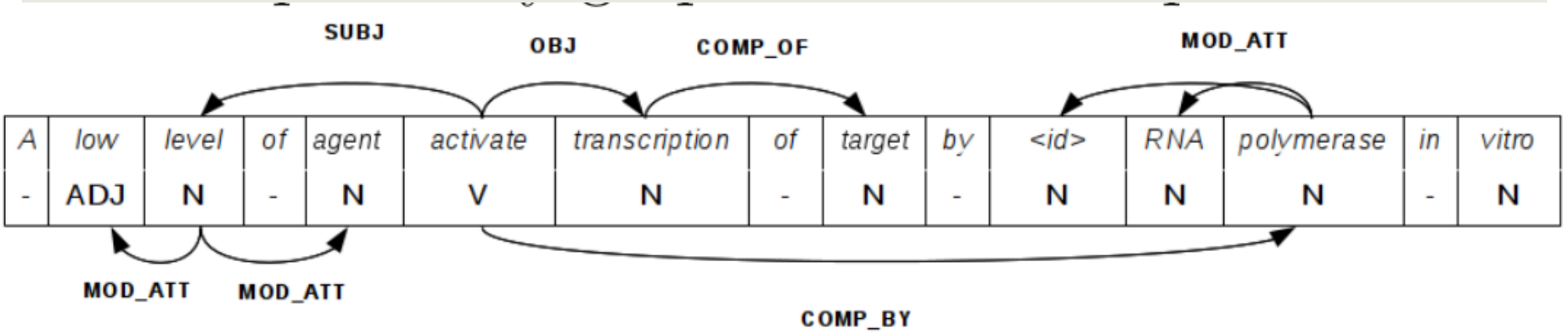
Architecture du système d' EI



Extraction d'information

- AlvisNLP : traitement linguistique
  - NER
  - analyse syntaxique, données linguistiques
  - données sémantiques
- Calcul de données pour l'apprentissage (candidats)
- Représentation formelle des candidats
- Apprentissage automatique

Représentation en graphe



Analyse syntaxique

<agent>	MOD_ATT	level	SUBJ	activate	OBJ	transcription	COMP_OF	<target>
N	←	N	←	V	→	N	→	N

Chemin le plus court

Distance entre graphes

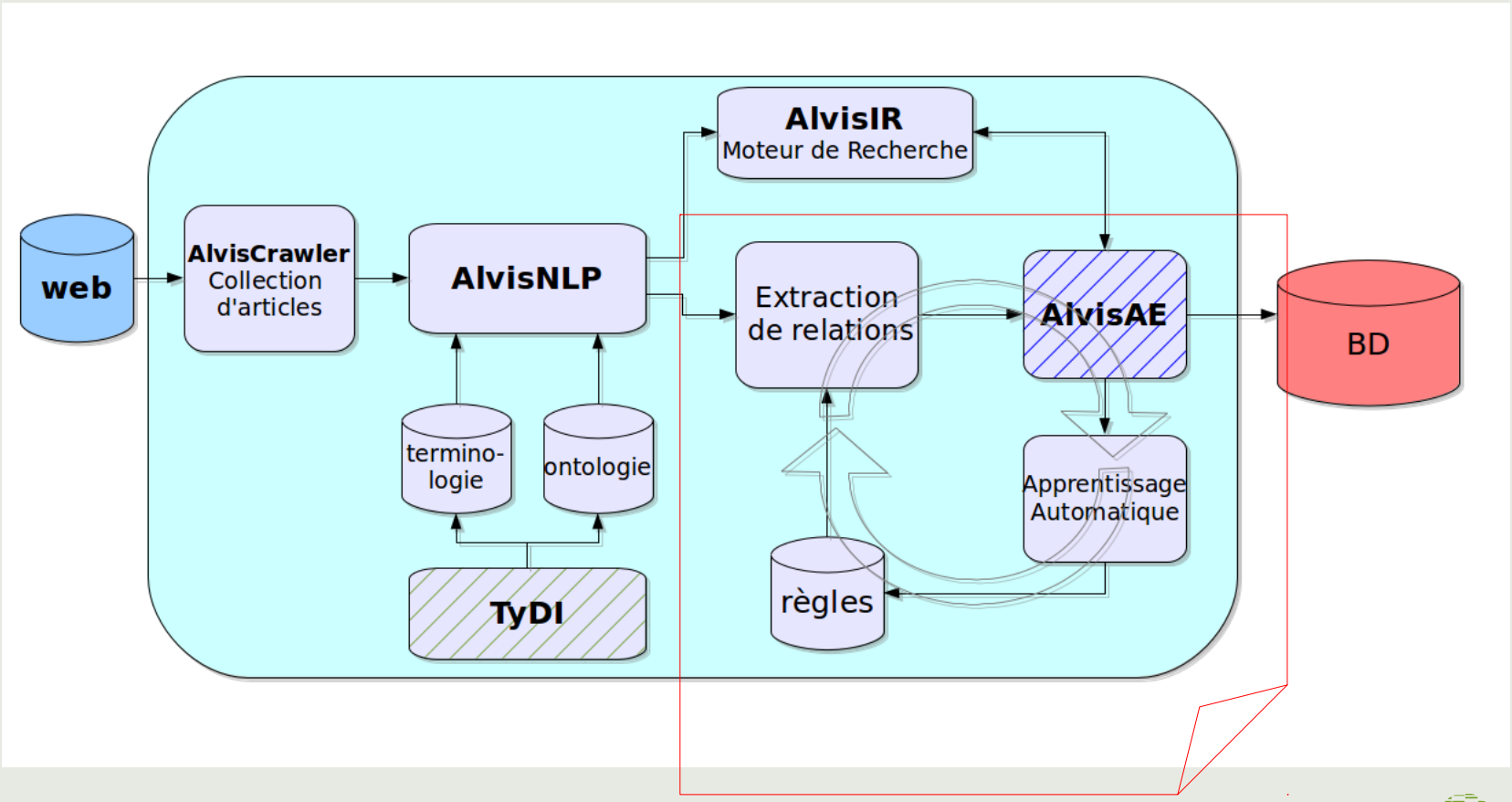
A low level of <AGENT> activated transcription of <TARGET> by sigmaK RNA polymerase in vitro

<AGENT>	MOD_ATT	level	SUBJ	activate	OBJ	transcription	COMP_OF	<TARGET>
|			|	|	|	|	|
<AGENT>	–	–	SUBJ	control	OBJ	expression	MOD_ATT	<TARGET>

<AGENT> appears to control <TARGET> expression

Alignement de deux phrases pour calculer leur distance

Processus itératif



Avancement

- Extraction d'Information
  - relations binaires
  - représentation en graphes, distance alignée
  - informations syntaxiques, sémantiques
  - résultats comparables à l'état de l'art sur *benchmark*
- Campagne d'Annotation
  - *Guidelines* à finaliser
  - Travail d'annotation en cours

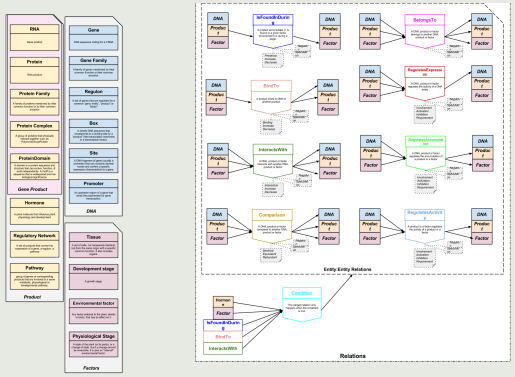
À venir

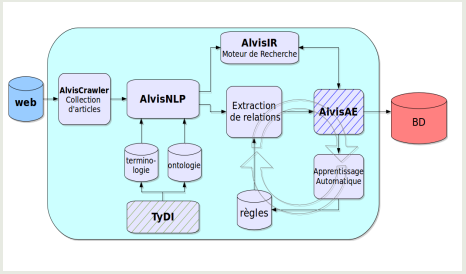
- Achèvement d'une première campagne d'annotation manuelle
- Utilisation des annotations manuelles
- Étendre et adapter le système d'extraction d'informations pour le modèle d' *A. Thaliana*
- Annotations prédites de relations dans l'éditeur

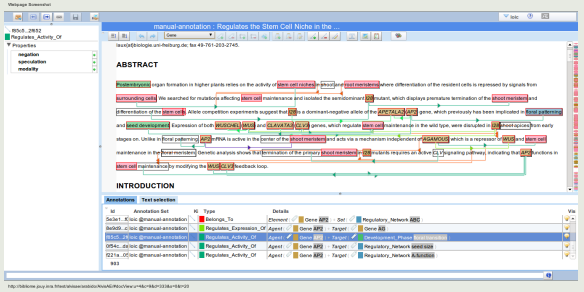
Usage des résultats de l'extraction

- dans le futur :
  - Inférence du graphe/réseau
  - Visualisation
- déjà possible :
  - représentation « base de données »
  - utilisation du système AlvisAE+AlvisIR

Merci pour votre attention







Apprentissage Automatique : Architecture

1. Calcul de candidats
2. Transformation de candidats en format de chemin sur arbre syntaxique :
   - Noeuds : mots et relations syntaxiques
3. Calcul de la matrice de similarité, pour chaque couple :
   - Calcul de distance de type Needleman-Wunsch
4. Donnée d'apprentissage : vecteur de distances d'un candidat avec tout autre candidat
5. Classification par SVM linéaire