

# 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)

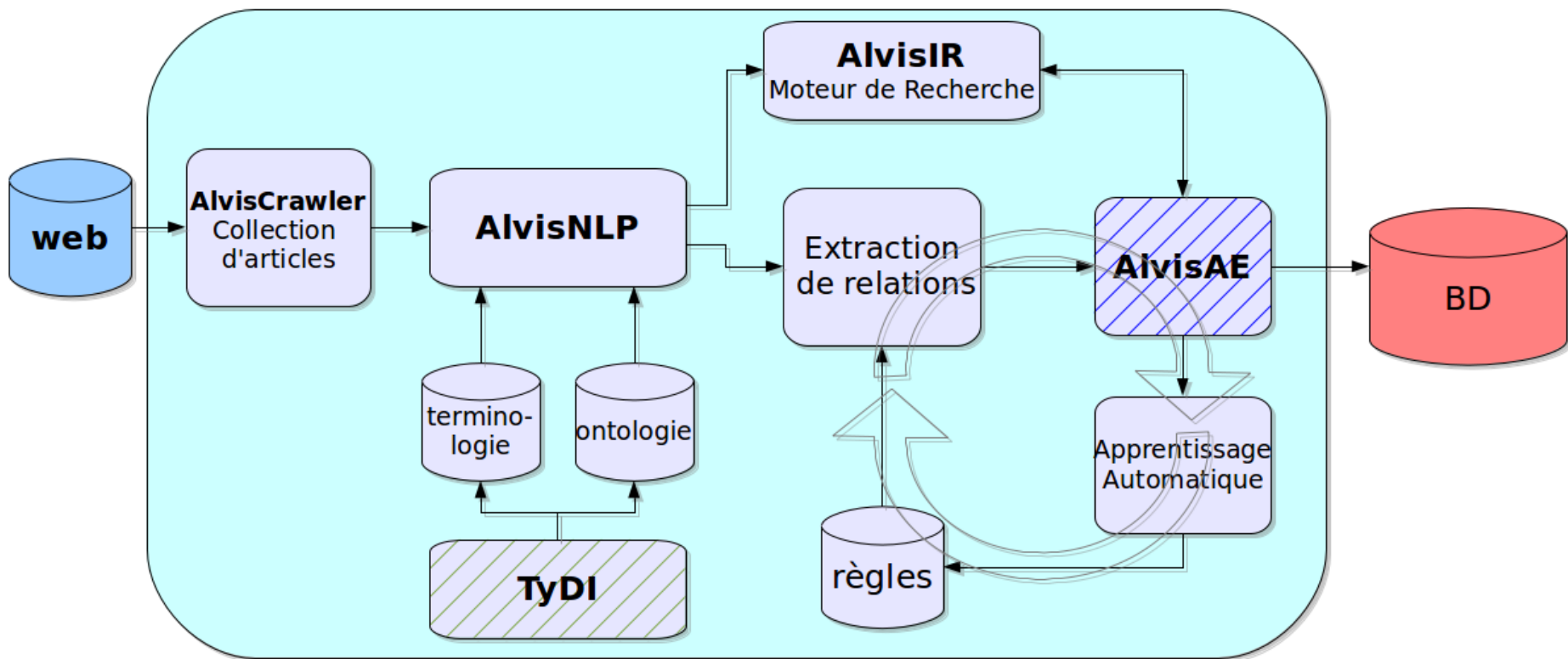
# Intuition

- 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

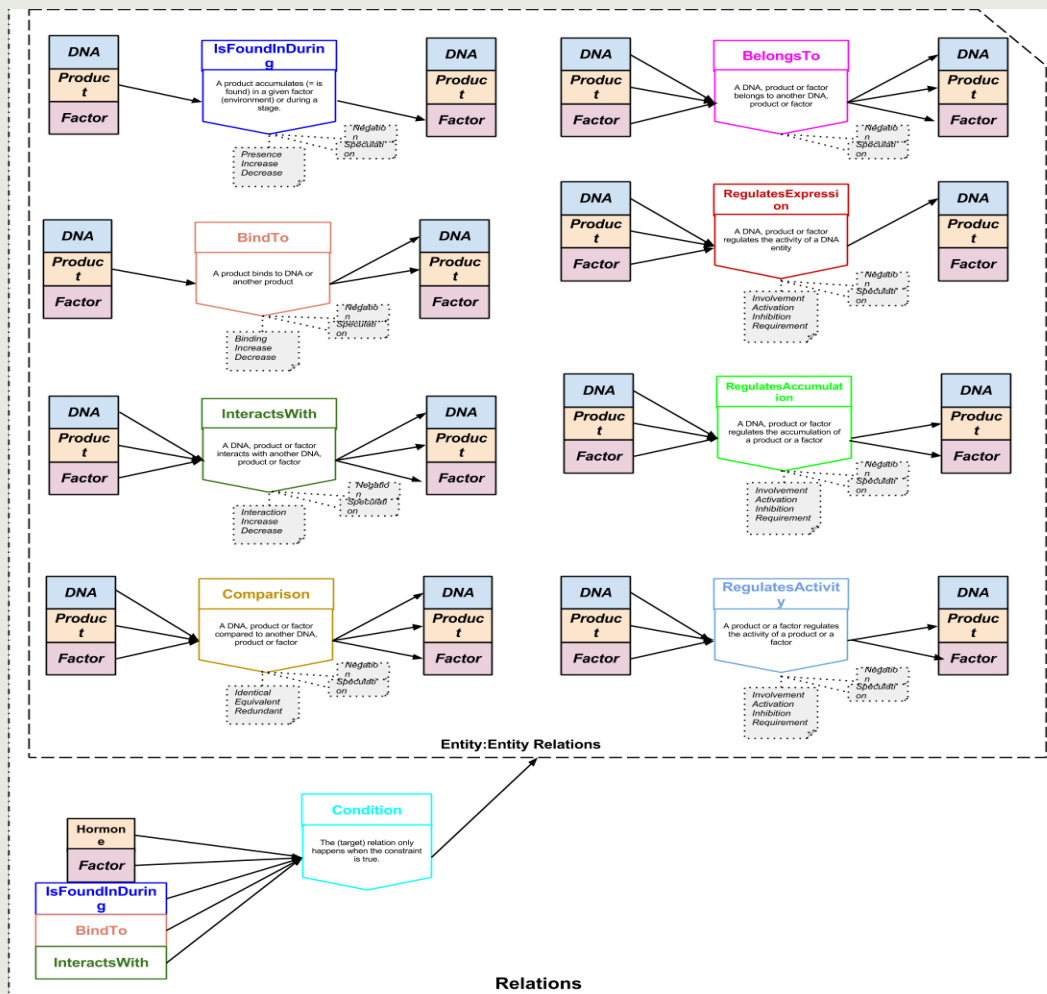
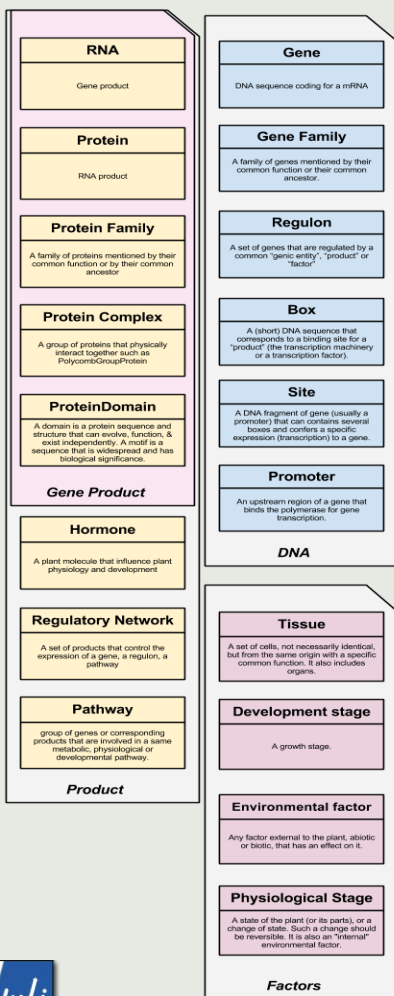
# Architecture du système d' EI



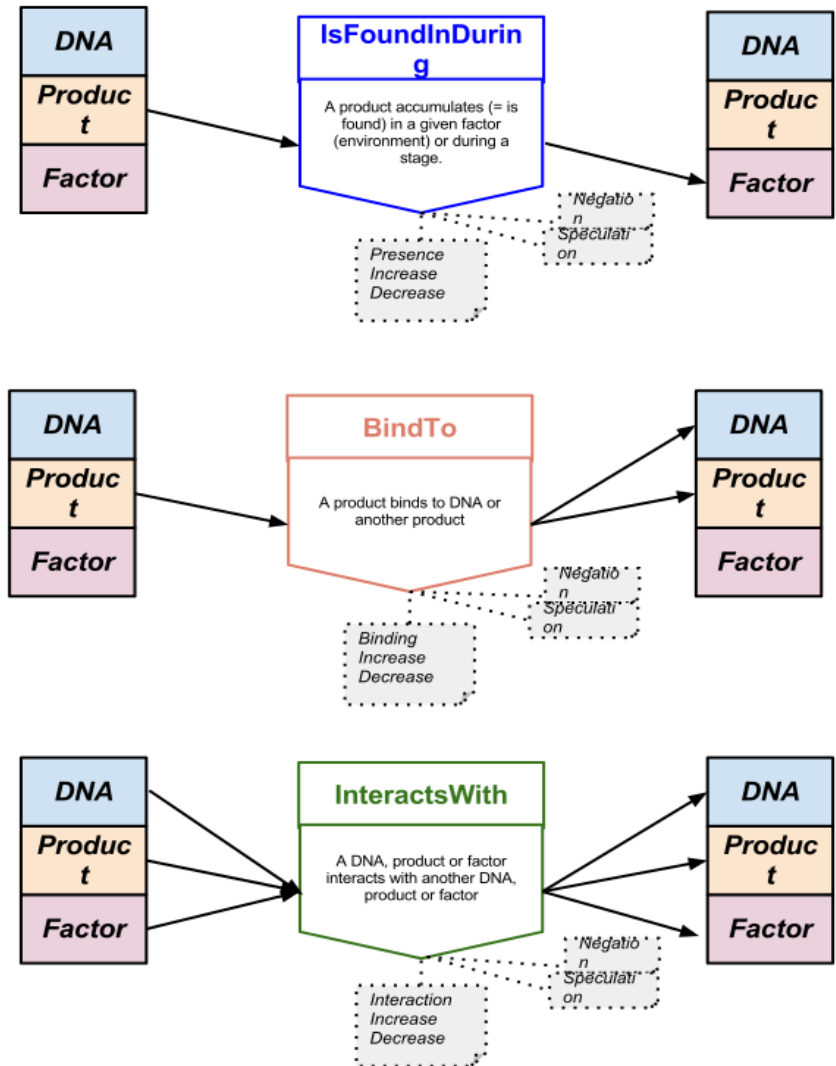
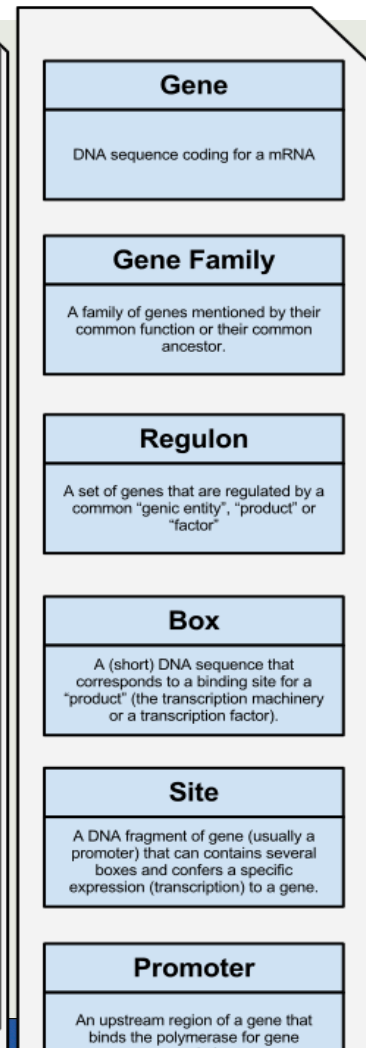
# Modélisation

- 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



# 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 1/2 page \*
- Deux experts
  - Accord entre deux annotateurs
  - Cohérence temporelle d'un annotateur

• \* © B. Dubreucq 2013

# Annotation : éditeur

Webpage Screenshot

manual-annotation : Regulates the Stem Cell Niche in the ...

laux[at]biologie.uni-freiburg.de; fax 49-761-203-2745.

## ABSTRACT

Postembryonic organ formation in higher plants relies on the activity of stem cell niches in shoot and root meristems where differentiation of the resident cells is repressed by signals from surrounding cells. We searched for mutations affecting stem cell maintenance and isolated the semidominant *z28* mutant, which displays premature termination of the shoot meristem and differentiation of the stem cells. Allele competition experiments suggest that *z28* is a dominant-negative allele of the *APETALA2* (*AP2*) gene, which previously has been implicated in floral patterning and seed development. Expression of both *WUSCHEL* (*WUS*) and *CLAVATA3* (*CLV3*) genes, which regulate stem cell maintenance in the wild type, were disrupted in *z28* shoot apices from early stages on. Unlike in floral patterning, *AP2* mRNA is active in the center of the shoot meristem and acts via a mechanism independent of *AGAMOUS* which is a repressor of *WUS* and stem cell maintenance in the floral meristem. Genetic analysis shows that termination of the primary shoot meristem in *z28* mutants requires an active *CLV* signaling pathway, indicating that *AP2* functions in stem cell maintenance by modifying the *WUS-CLV3* feedback loop.

## INTRODUCTION

**Annotations** | **Text selection**

Id	Annotation Set	KI	Type	Details	Vis
5e3e1...f	loic @manual-annotation		Belongs_To	Element (  Gene AP2 ) + Set (  Regulatory_Network ABC )	
8e9d9...c	loic @manual-annotation		Regulates_Expression_Of	Agent (  Gene AP2 ) + Target (  Gene AG )	
f85c5...2f	loic @manual-annotation		Regulates_Activity_Of	Agent (  Gene AP2 ) + Target (  Development_Phase floral transition )	
0f54c...da	loic @manual-annotation		Regulates_Activity_Of	Agent (  Gene AP2 ) + Target (  Regulatory_Network seed size )	
f221a...0c	loic @manual-annotation		Regulates_Activity_Of	Agent (  Gene AP2 ) + Target (  Regulatory_Network A-function )	

903

<http://bibliome.jouy.inra.fr/test/alvisae/arabido/AlvisAE/#docView:u=4&c=9&d=333&o=0&t=20>

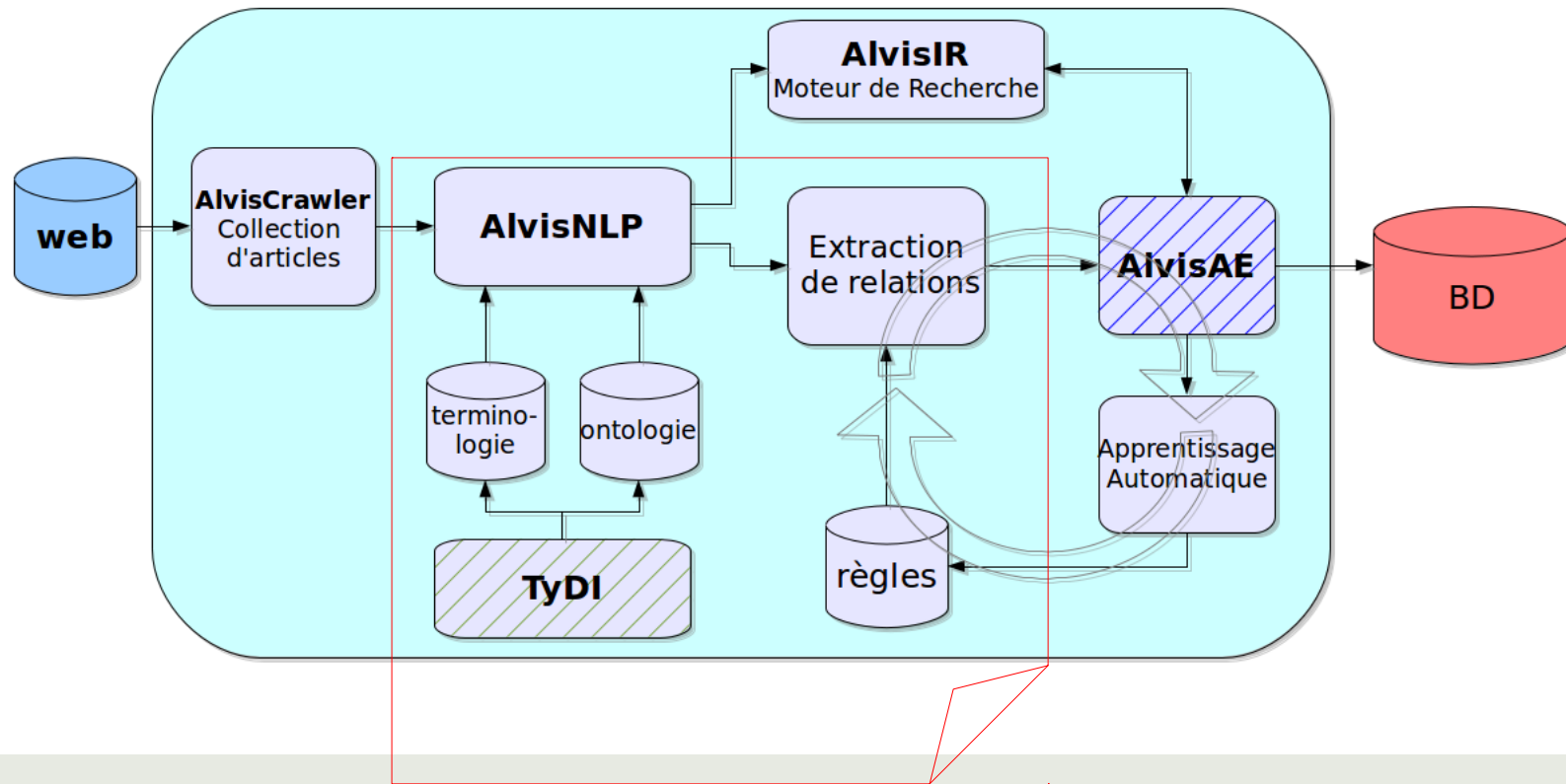
# Défis linguistiques

- 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-aires
  - relation syntaxique ↔ relation sémantique

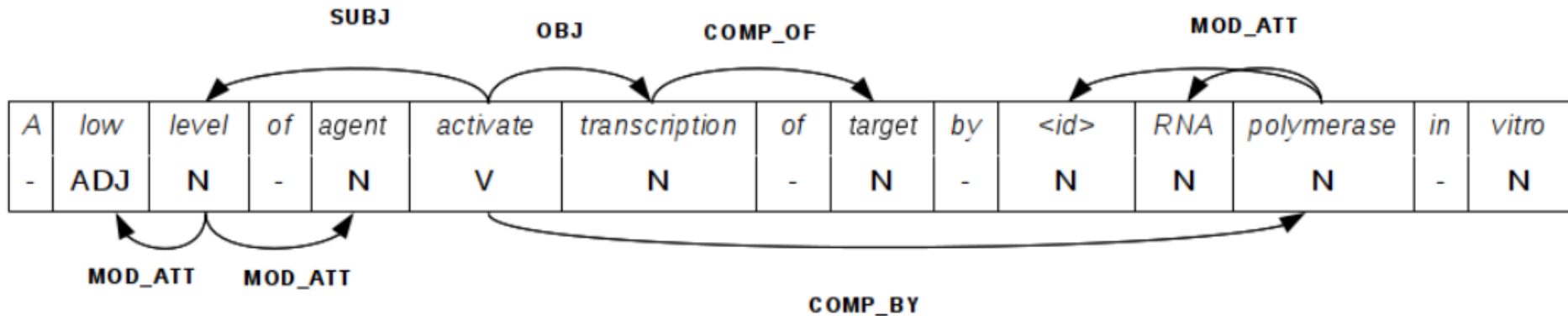
# Architecture du système d' EI



# 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>

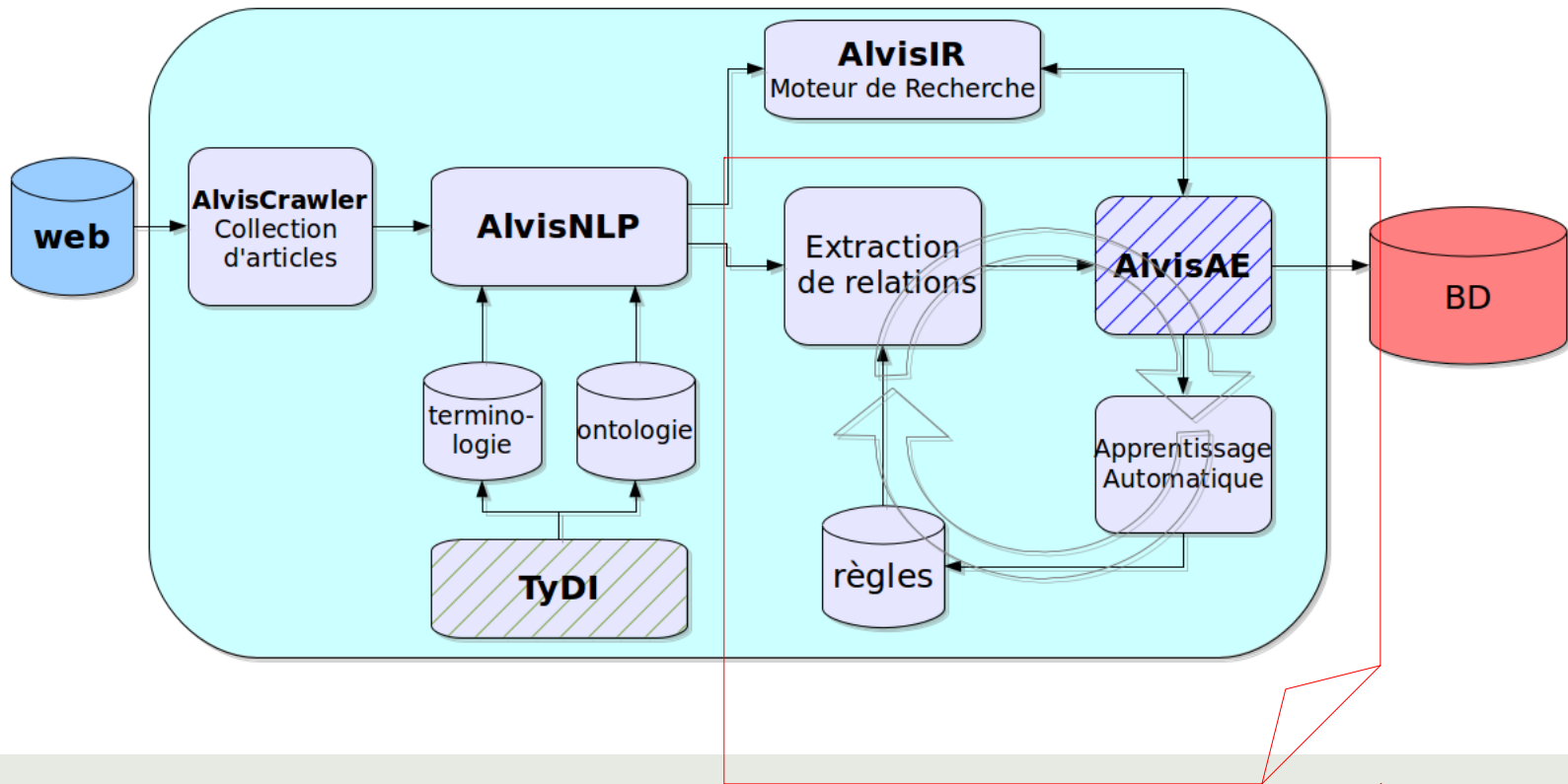


*<AGENT> appears to control <TARGET> expression*

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

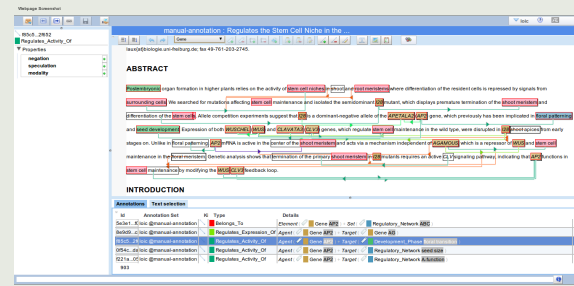
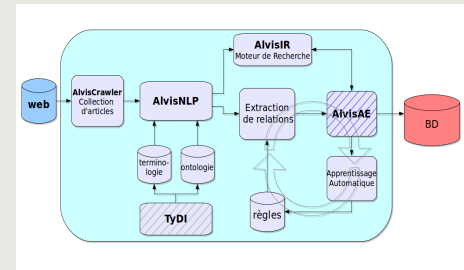
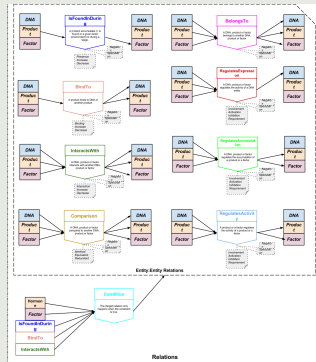
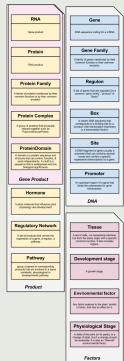
# À venir

- 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