INRA
SCIENCE & IMPACT

Conception d'un corpus annoté BioNLP Shared Task

Extraction de régulations génétiques

chez *Arabidopsis thaliana*

Bibliome group

AgroParisTech Paris, In Ovive – 24 novembre 2015

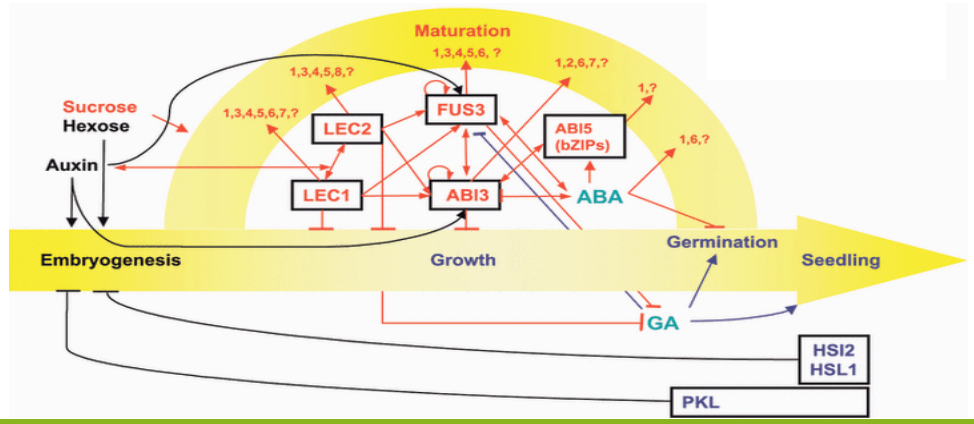
GRN, développement de la graine

de la plante modèle *Arabidopsis thaliana*

- **Focus**: construire un modèle des mécanismes de stockage de réserve et de maturation
- Une question majeure pour **l'amélioration des plantes** et pour la **recherche fondamentale**

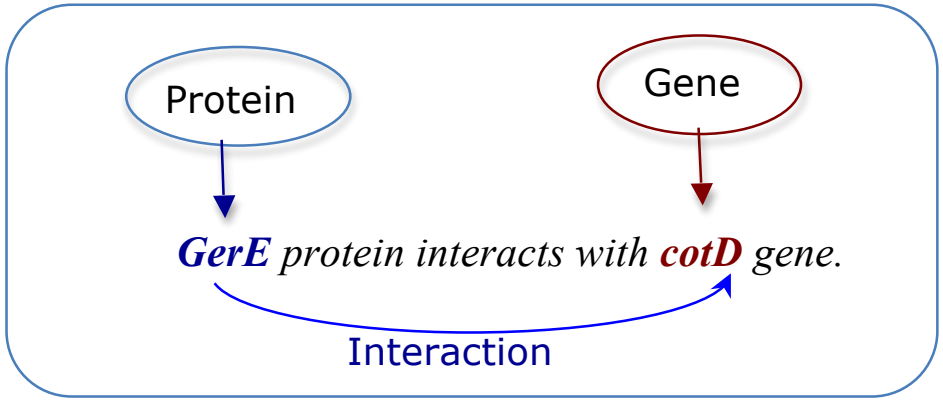
Résultats attendus de l'Extraction d'Information (EI), perspective de Biologie des Systèmes (IMSV)

- Multi-échelle: genetique, physiologique, phénotype and environment
- Un modèle riche est nécessaire

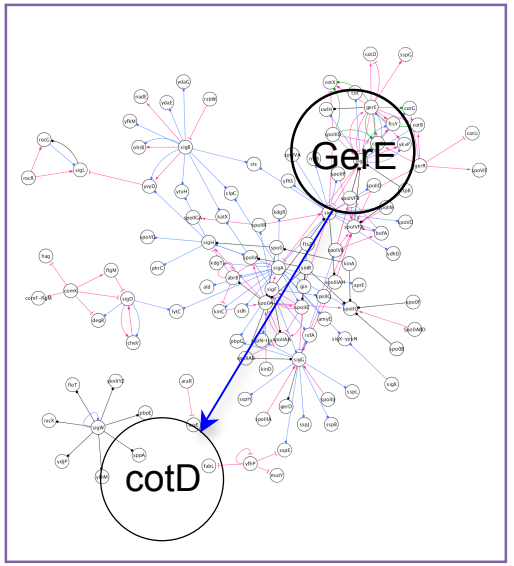


Gene Regulation Network (GRN)

Une connaissance clef en biologie, dispersée dans des milliers d'articles scientifiques.
L'extraction de réseaux de régulations, un des premiers buts de l'EI en Biologie
(challenges LLL, BioCreative, BioNLP-ST)

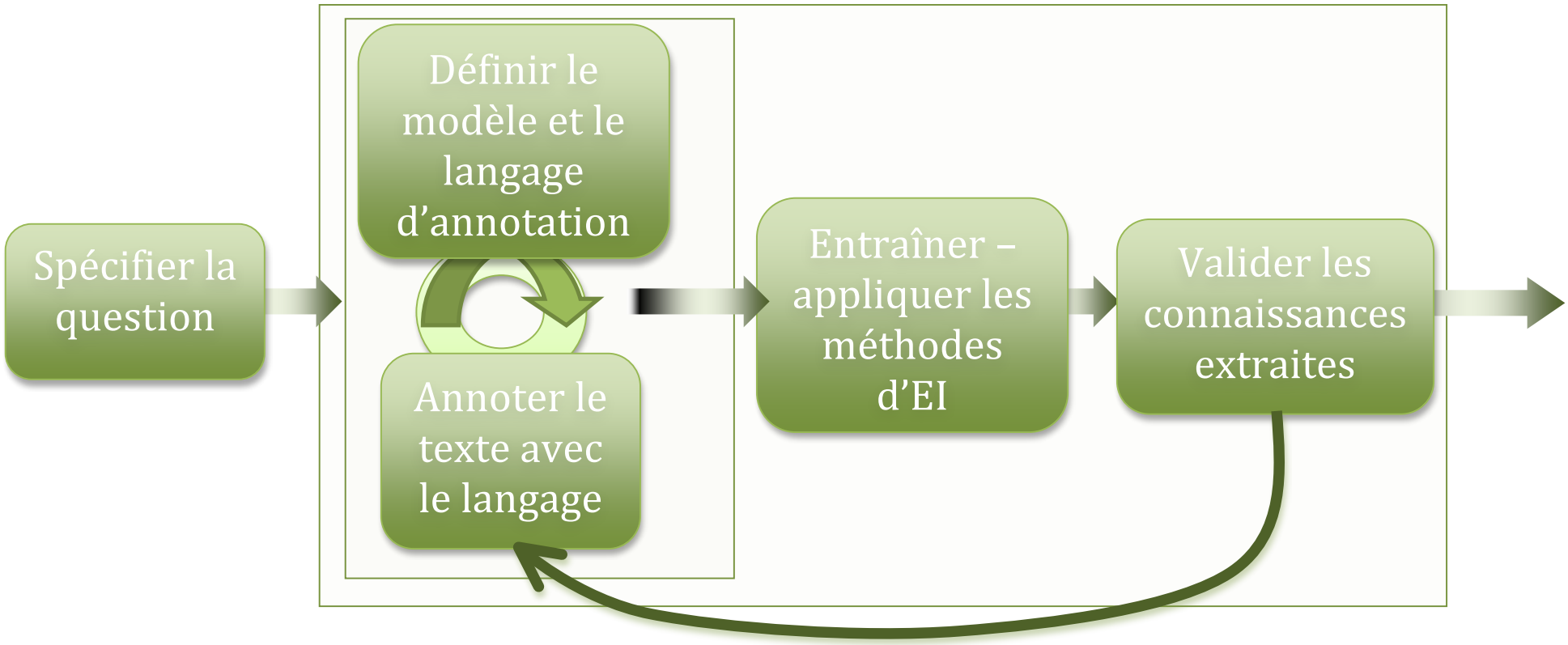


Extraction de l'Information à partir de textes



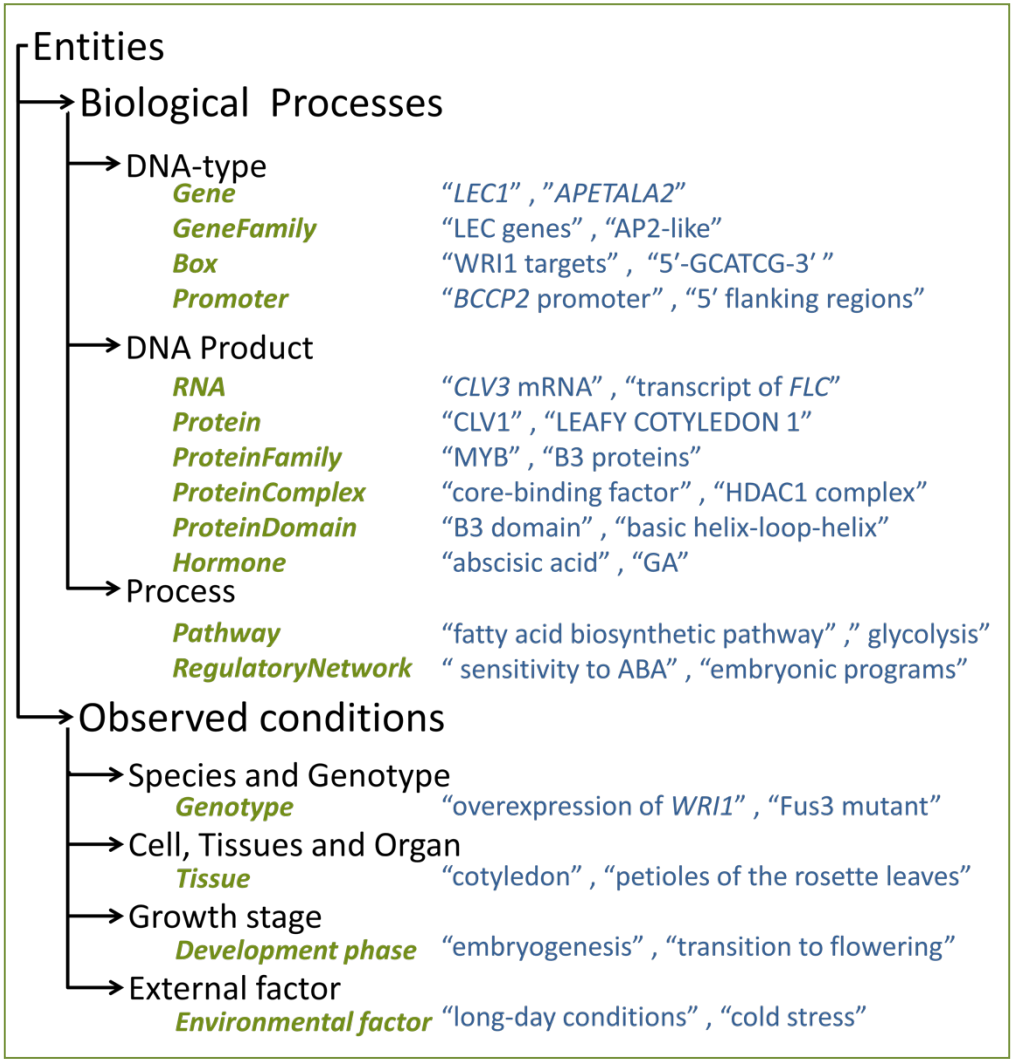
Réseau de régulation

Démarche de conception d'applications d'extraction d'information à partir de textes



Modèle de connaissance pour le développement de la graine

- Modèle riche par rapport à l'état de l'art en Extraction d'Information



16 types d'entités biologiques

Un modèle relationnel simplifié pour l'annotation manuelle

→Regulation :
- *RegulatesActivityOf*
- *RegulatesAccumulationOf*
- *RegulatesExpressionOf*

→Interaction :
- *InteractWith*
- *BindTo*

→Localisation :
- *IsFoundIn*
- *IsFoundDuring*

→Similarity :
- *Comparison*
- *Belongs to*
- *Encodes*

1 relation to define n-ary events
- *Condition*

10 relations

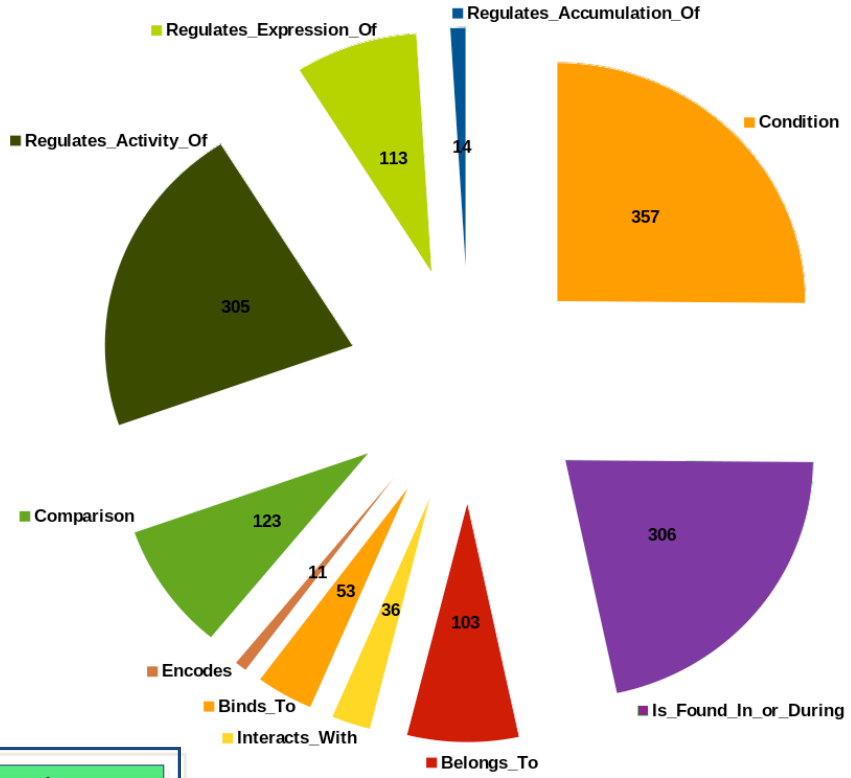
L'annotateur ne peut pas gérer beaucoup de relations

Compromis entre
- La simplicité pour l'annotation manuelle des données d'entraînement
- L'homogénéité des annotations pour l'apprentissage
- La précision de l'information

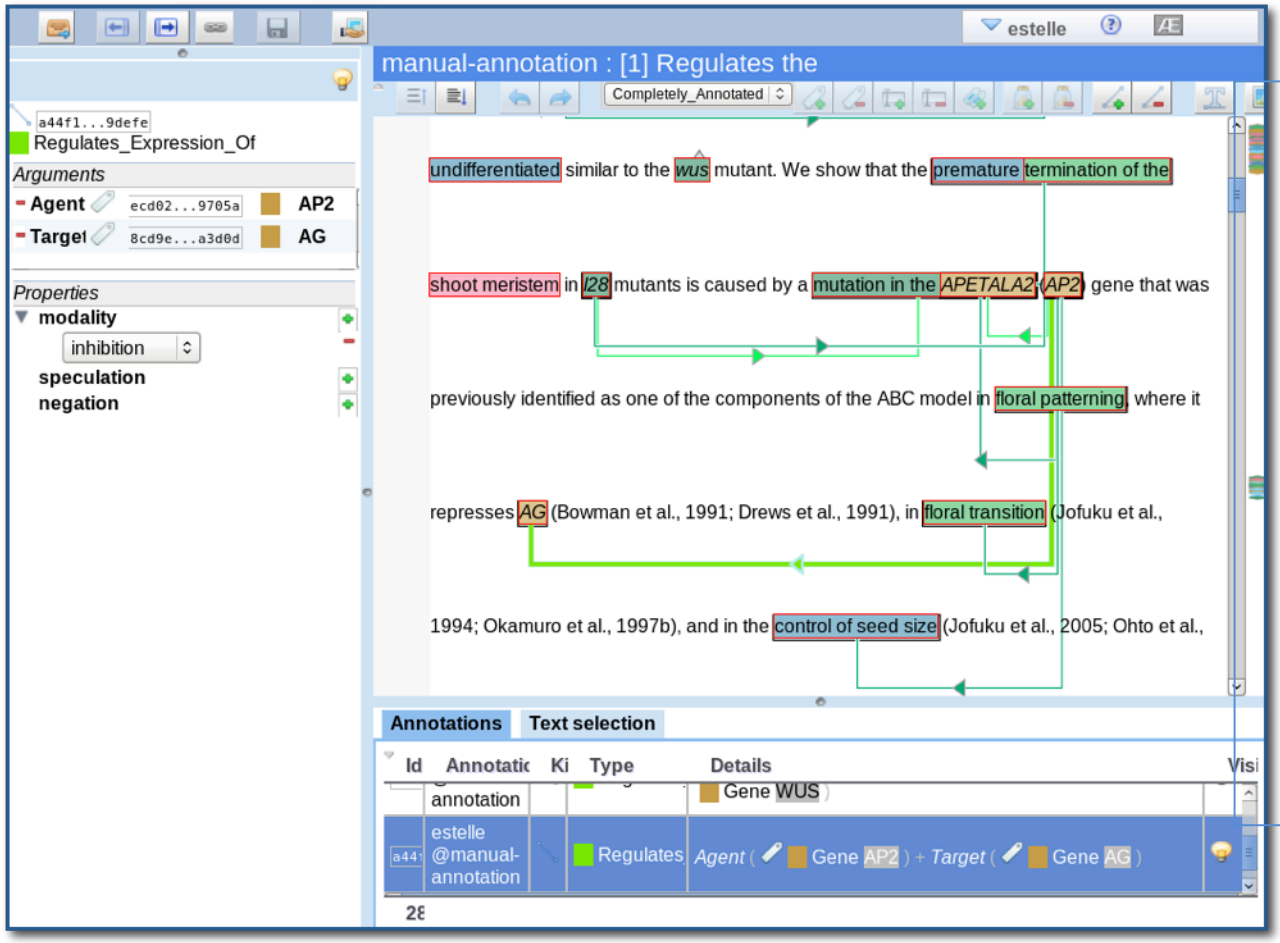
Schéma pour l'annotation manuelle

Solution

- 10 relations de haut niveau
- Fusion des relations sur le contexte (*i.e.* génotypes, phases de dévelopment et tissues) dans une seule relation *condition*
- Sont assignées à part,
  - Les modalités (spéculation et négation)
  - et les spécialisations de relation (increase, decrease, etc.)



Annotation manuelle



- Peuplement du modèle de connaissance à l'aide de l'éditeur d'annotation AlvisAE
- Préannotation par la chaîne AlvisNLP
-
- Annotation manuelle d'articles par quatre biologistes, 3 IJPB et 1 Bibliome.
- Aujourd'hui 4 444 entités et 1 421 relations

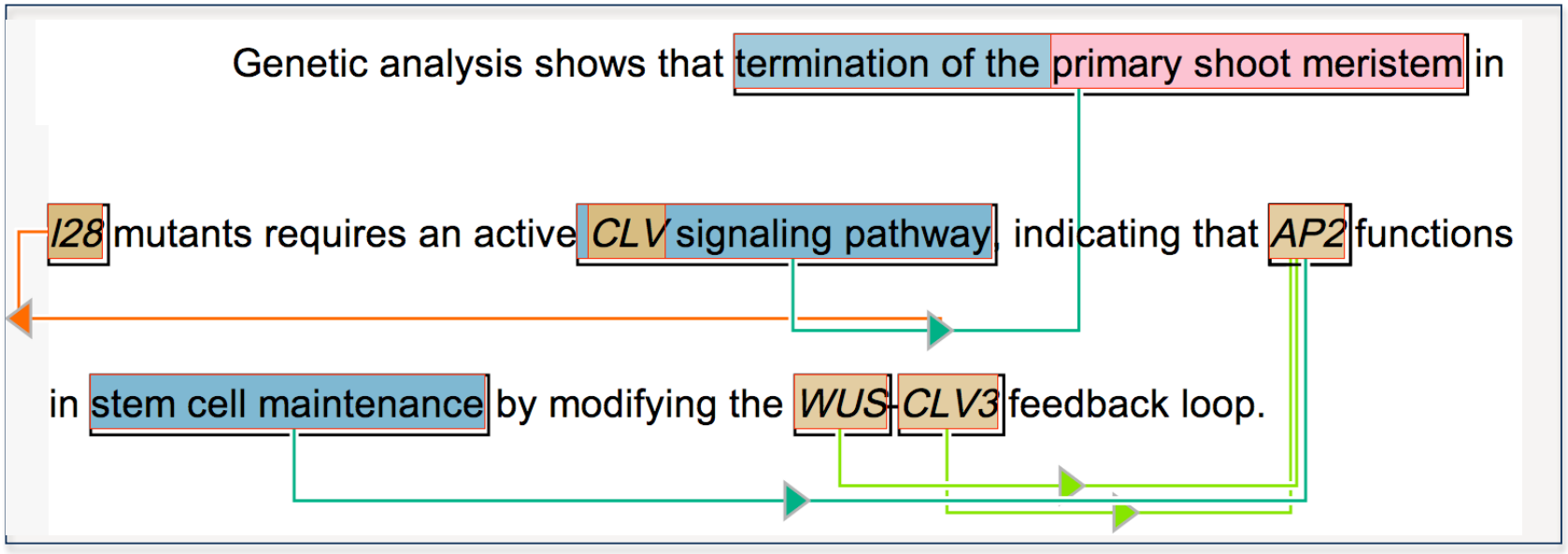
Exemples d'annotation

Entités

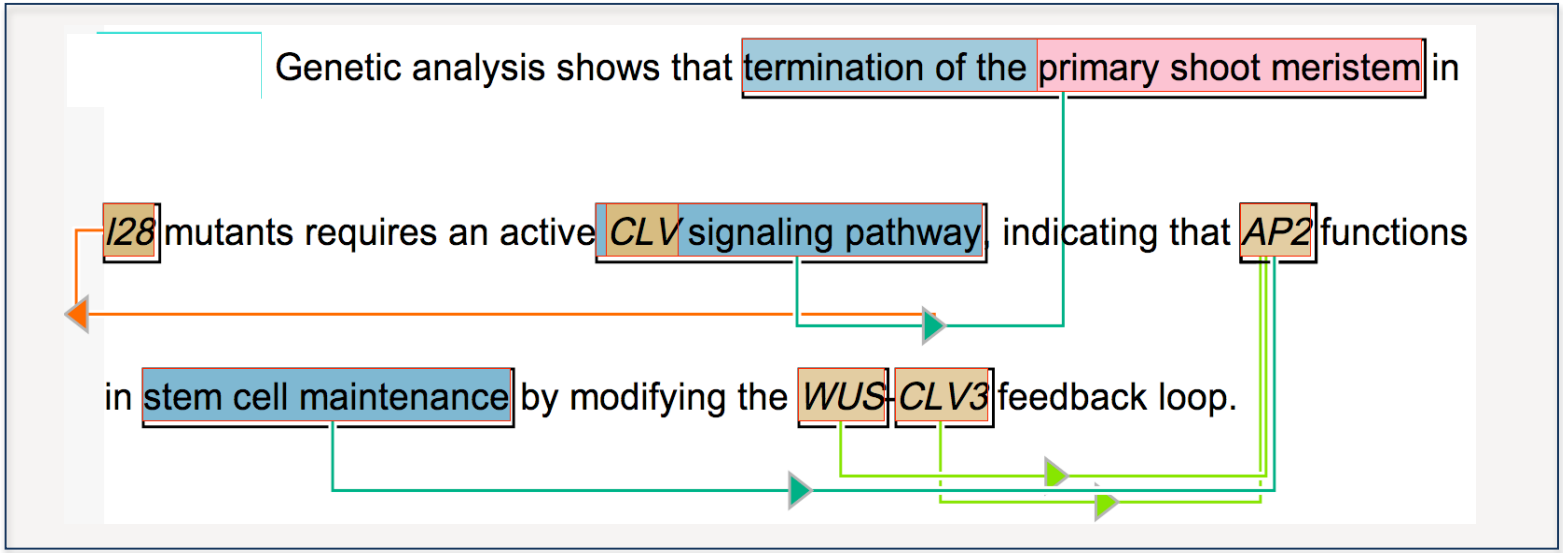
Gene
Tissue
Metabolic patway
Regulation network

Relations

Agent **(Met. pathway)** RegulatesActivityOf Target **(Gene).**
Agent **(Gene)** RegulatesExpressionOf Target **(Gène).**
...



Exemple d'annotation formelle de texte



Gènes. *I28, AP2, WUS, CLV3*
Tissus. *primary shoot meristem*
Voie métabolique. *CLV signaling pathway*
Réseau de régulation. *termination of the primary shoot meristem, stem cell maintenance*

Entités

(**Reg net.** *stem cell maintenance*)	RegulateActivityOf	**(Gène** *AP2***).**
(Metab. pat. *CLV signaling pathway***)**	RegulateActivityOf	**(Réseau** *termination of the primary shoot meristem***).**
(Gene *WUS***)**	RegulatesExpressionOf	**(Gène** *AP2***).**
(Gene *CLV3***)**	RegulatesExpressionOf	**(Gène** *AP2***).**

Relations

Méthodologie d'annotation

1. **Définition du modèle**, préparation du document de consigne initial

2. Annotation avec AlvisAE

- Annotation des entités
- Annotation des relations en double-aveugle
- Adjudication
- Réunion des participants pour discuter les révisions

Itérations entre

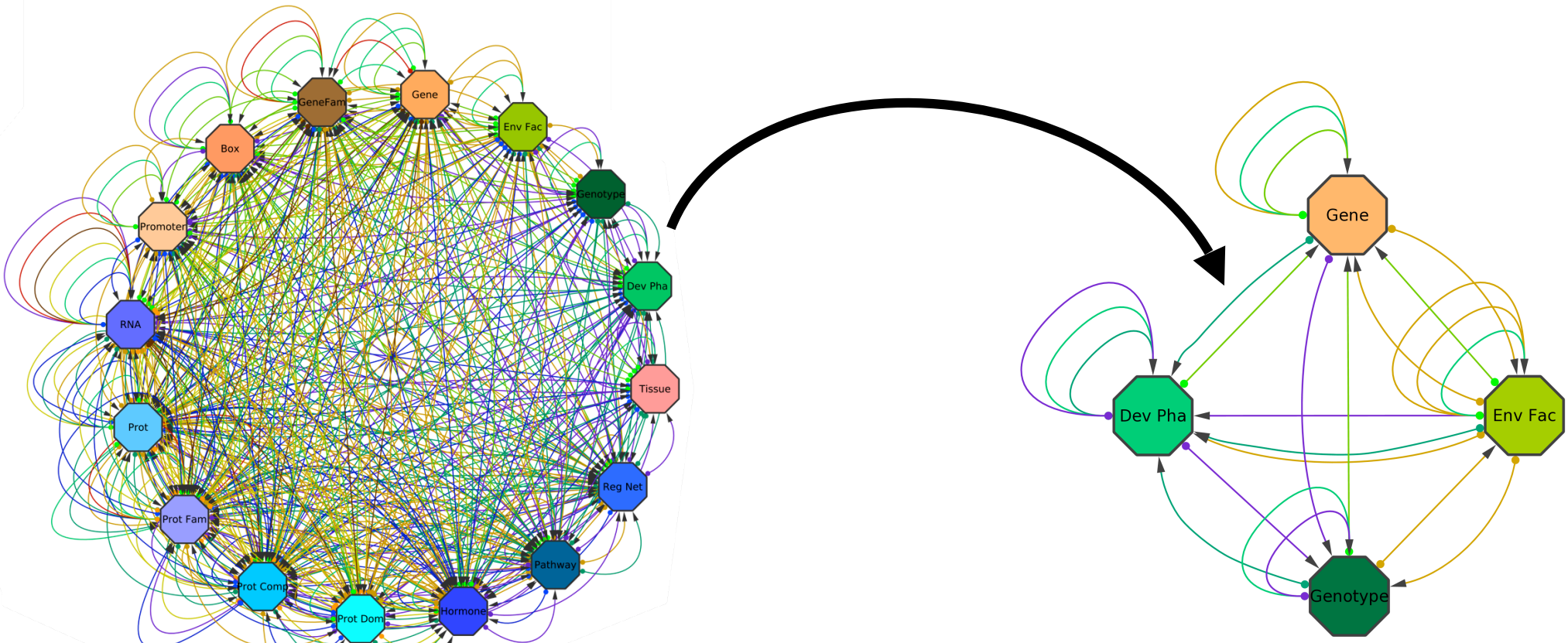
- annotations
- modification du modèle
- révision des guidelines

Importance du document de *Guidelines* dans la préparation des données

- Définition du modèle
- Consignes
- Exemples typiques et cas limites
- Guidelines partagées sous GoogleDoc

Longue mise au point du modèle, complexité biologique

Modèle riche : de nombreux *types* d'argument par relation



Relation binaire et ternaires, 116 types, combinaison de

+ Modalités : *speculation, negation*

+ Hiérarchies de relations: *presence, increase, decrease, involvement. activation. inhibition. reauirement*

Exemple de représentation dans AlvisAE

DNA product regulates the activity of DNA type or process in genotype

Une relation à 3 arguments: the agent that controls the target in given genotype.

Représentations pour l'utilisateur

WRI1 protein is able to regulate in planta the activity of the *BCCP2* promoters.

RegulatesActivityOf

Agent (protein [*WRI1*]) + *Target* (promoter [BCCP2]) + *Condition* (genotype [planta])

Représentation formelle

Schéma

RegulatesActivityOf(X, Y, Z)

Has_role(X,Agent),

Has_role(Y,Target), Has_role(Z,Condition)

Protein(X), Promoter(Y), Phenotype(Z)

Instance

RegulatesActivityOf(p1,pr1,v1)

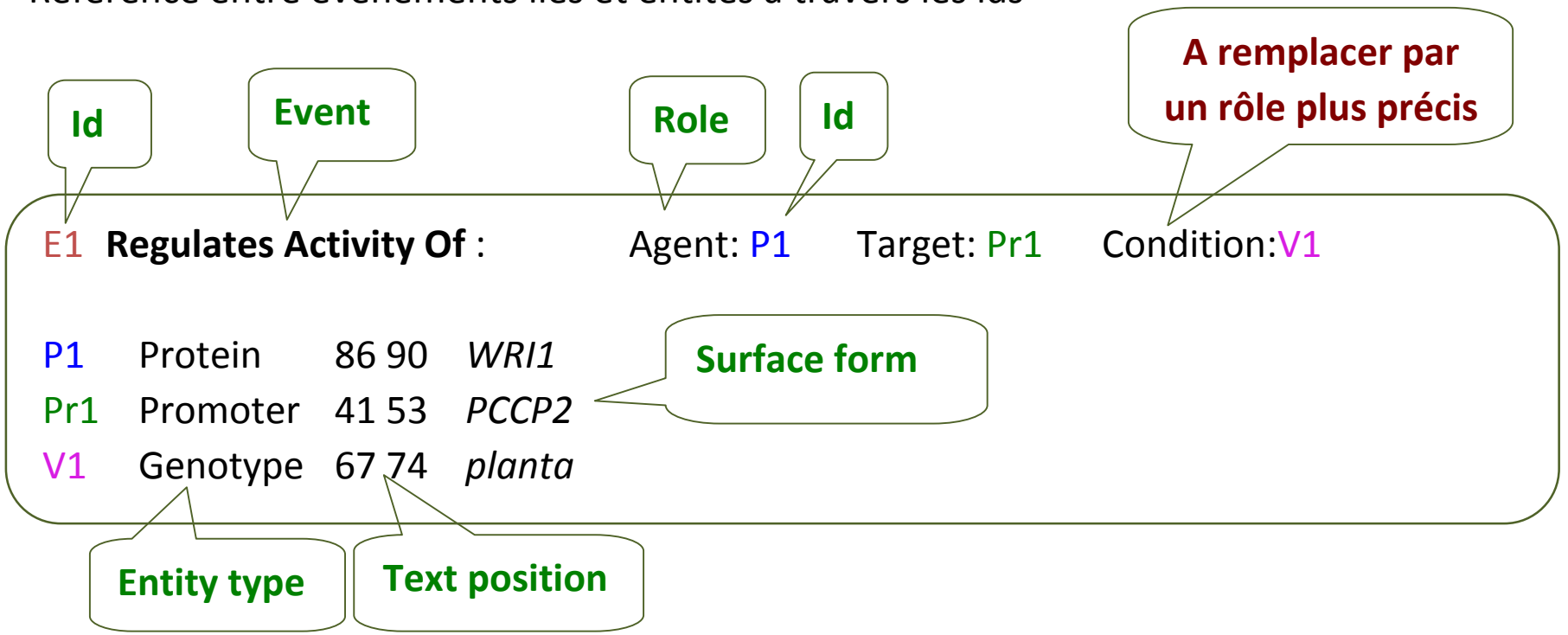
Surface_form(p1, *WRI1*)

Surface_form(pr1, BCCP2)

Surface_form (v1, planta)

Langage pour le copus de référence : BioNLP-ST

- **Entités**: id, type, forme de surface (discontinuité possible) et position dans le texte
- **Evénements**: Type, id, les arguments sont des entités ou des événements avec leurs *rôles*
- Reference entre événements liés et entités à travers les ids



Changements de représentation

de l'annotation manuelle au corpus de référence

Le principe de la représentations de BioNLP ST est similaire à celui de AlvisAE editor

La réécriture dans les faits, pas si simple,
en raison des modalités et des spécialisations de relation fusionnées (condition)

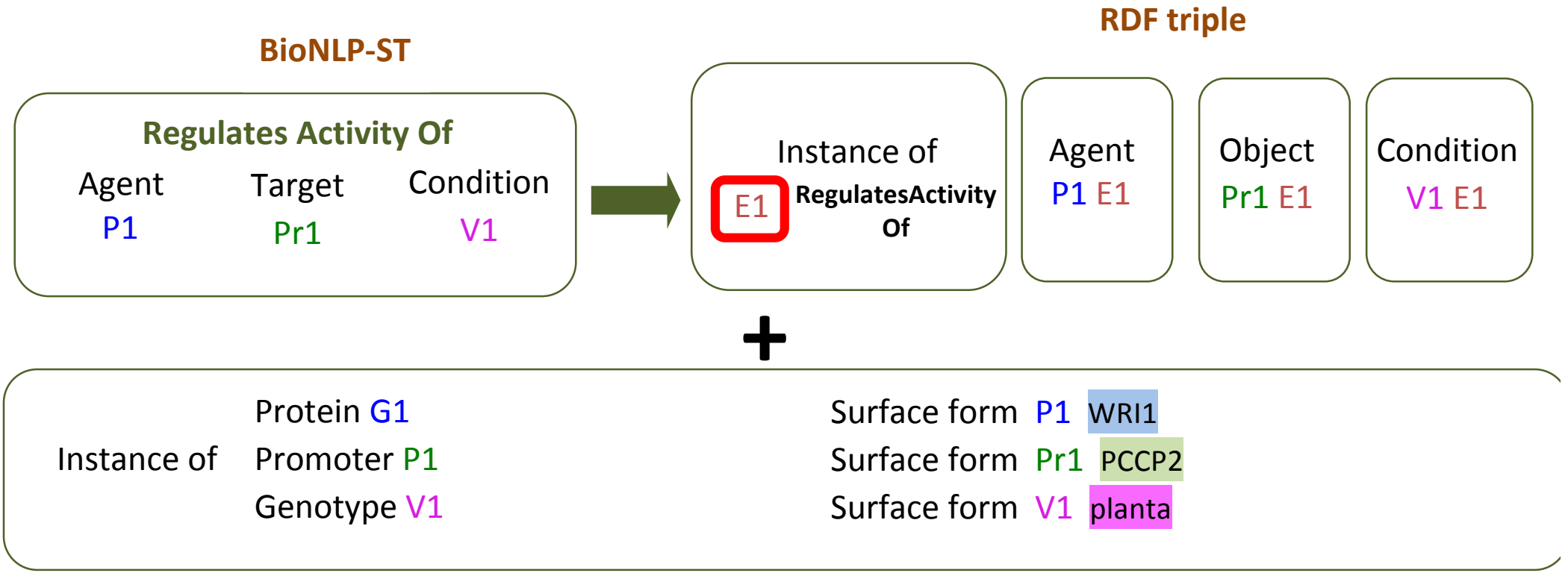
- Assignation d'un label précis en fonction des types des arguments
- Modalités representés par des arguments

Dépôt RDF

Relations binaires, entre paires d'argument, avec un label

Réécriture de chaque relations ternaire en quatre relations binaires

Sans perte d'information : création d'une variable additionnelle.



Organisation de BioNLP-ST'16

4è édition, après 2009, 2011 et 2013

4 tâches prévues

Pharmacologie (Nactem), Genia (DBCLS),

Bacteria Biotope, ***Gene Regulation Network in Ath*** (Inra)

Calendrier

- Diffusion des exemples et format disponibles : fin novembre
- Diffusion des données d'entraînement : mi-décembre
- Test : mars
- Rédaction des articles du workshop ACL: mars-avril
- Workshop ACL BioNLP (joint avec BioASQ) à Berlin : début août
- Préparation d'un numéro spécial pour un journal

Exploitation du corpus GRNA

Paris-Saclay Center for Data Science

INRA SCIENCE & IMPACT

cnrs

Modèle et corpus de référence
Réseaux de régulation
dans le développement de la graine d'*Ath.*

Données d'entraînement et de test pour

La mise au point de méthodes

- **d'extraction de relations**
  - thèse de Dialekti Valsamou — thèse IDI UPSAY
  - BioNLP Shared Task

d'annotation de termes par des ontologie

- thèse d'Arnaud Ferré — thèse IDI UPSAY

Intégration de la connaissance extraite dans les modèles de la biologie des systèmes — IMSV

Use case *plant* du projet européen d'infrastructure de text-mining openMINTED

Extension à d'autres espèces d'intérêt agronomique (blé, maïs)

Extraction de relations par apprentissage supervisé, premiers résultats

Méthode choisie : *shortest dependency path (AlvisGrammar) - global alignment kernel (AlvisRE)*
Représentation : chemin syntaxique entre les arguments candidats

Principe

- Utilise un alignement global entre les chemins pour comparer les exemples
- Utilise les similarités entre chemins comme des noyaux de SVM

RegulatesActivityOf(WRI1, PKp-β1)
Oui, connu

RegulatesActivityOf(LEC1, WRI1)
Vrai ???

The WRINKLED1 (**WRI1**) *protein regulates the activity of* ***PKp-β1*** *in planta*

Regulates

Expression Of

WRI1	MOD_ATT	WRINKLED1	MOD_ATT	protein	SUBJ	regulate	OBJ	activity	COMP_OF	PKp-β1
LEC1	-	-	-	-	SUBJ	up-regulate	OBJ	expression	COMP_OF	WRI1

The expression of ***WRI1*** *is up-regulated by* ***LEC1***

??

Résultats préliminaires (entraînement sur 5 articles)

Relation	Nb occ	Précision	Rappel	F-mesure
Regulates_Activity_Of	132	0.39	0.61	0.48
Encodes	6	0.6	0.5	0.55
Regulates_Expression_Of	94	0.43	0.48	0.45
Belongs_To	33	0.32	0.46	0.38
Comparison	55	0.36	0.69	0.47
Total	320	0.39	0.53	0.45

Au niveau de l'état de l'art sur l'extraction de réseau de régulation chez *B. subtilis* (BioNLP ST 13)

Données complétées depuis

Interacts_With (1 occ)

Regulates_Accumulation_Of (2 occ)

Données hétérogènes, bon rappel mais précision faible

Is_Found_In_or_During (139 occ, Préc = 0,17)

Binds_To (19 occ, préc=0,1)

Perspectives

Représentation des connaissances

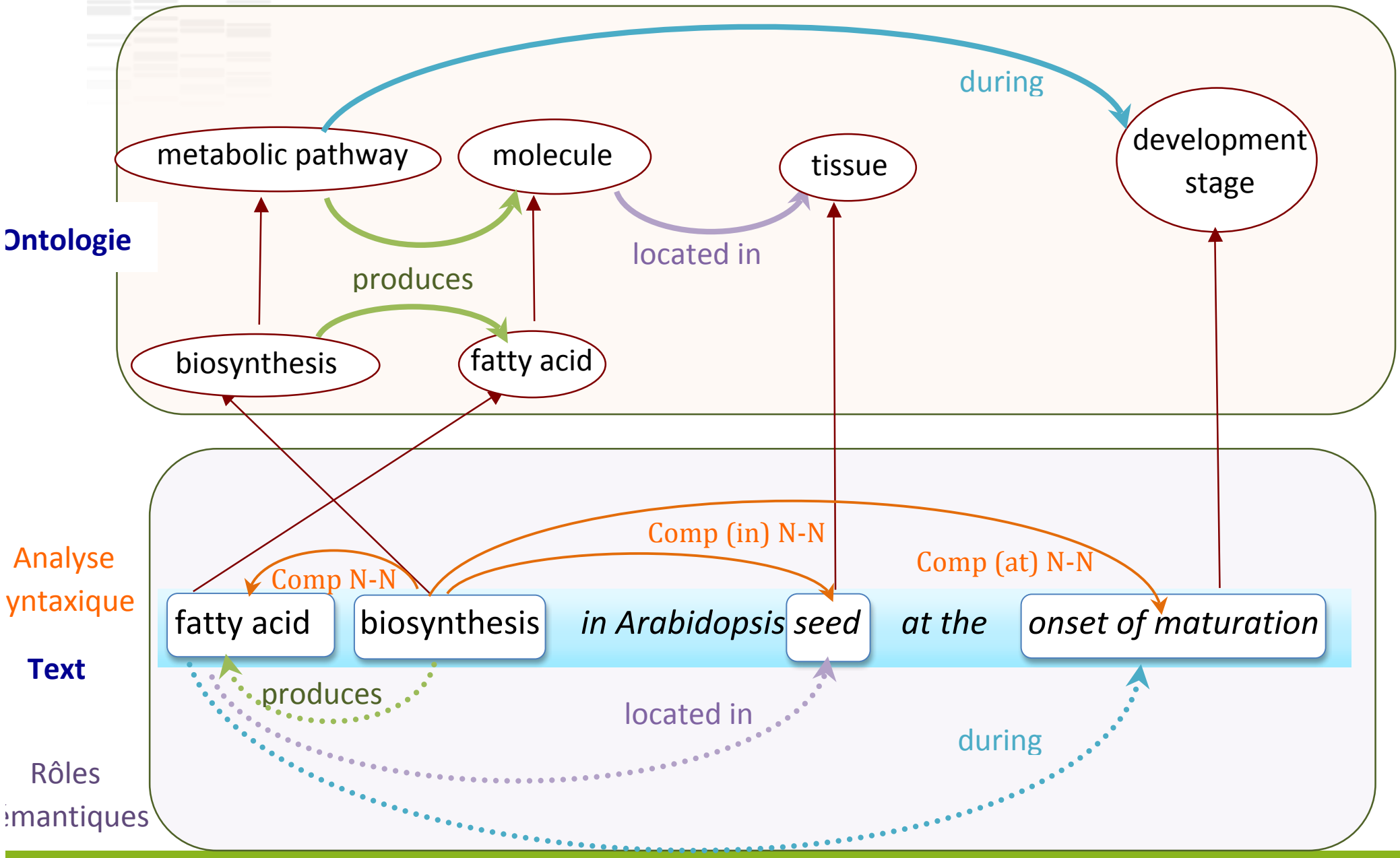
Exploitation des informations du texte dans un modèle commun avec
les informations des images, les données expérimentales, les réseaux inférés à partir de données d'expression

Recherche en extraction d'information

- Pour une analyse sémantique et conceptuelle plus fine des entités désignées par des termes complexes (ex. phases de développement, voies métaboliques, anatomie)
- Méthode hybride ToMap (terminologie, inférence dans l'ontologie, analyse distributionnelle).
- Thèse IDI d'Arnaud Ferré en cours (MaIAGE-LIMSI-IJPB)

Développement d'application

Projet H2020 OpenMinTeD *use case* sur le développement des plantes
En préparation : BioGemma, URGI, IFB, DIST, BioSys (MaIAGE), ...


Ontologie
metabolic pathway
molecule
tissue
development stage
during
produces
located in
biosynthesis
fatty acid
Analyse syntaxique
Text
Rôles sémantiques
Comp (in) N-N
Comp N-N
Comp (at) N-N
fatty acid
biosynthesis
in Arabidopsis seed
at the
onset of maturation
produces
located in
during

Projet H2020 OpenMinTed infrastructure europénne de text-mining

La création d'une plateforme et d'une infrastructure d'extraction d'information à partir de textes

- Agrégés à partir de différentes sources (éditeurs, archives)
  Qui correspondent aux besoins des utilisateurs (chercheurs).

- Incluant les **spécifications techniques** pour assurer l'interopérabilité, des métadata standard et l'agrégation de nouveaux services)

- **Validation** sur des besoins de grandes communautés (SHS, Sciences de la Vie, Agriculture & Biodiversité, etc.)

- **Démarrage** du projet au 1er juin 2015 pour 3 ans.
  Participation de l'Inra MaIAGE et DIST. Coordination par l'Université d'Athènes.