# Information Extraction Challenge

## *Gene Regulation Network*

## *in* Arabidopsis thaliana (GRNA)

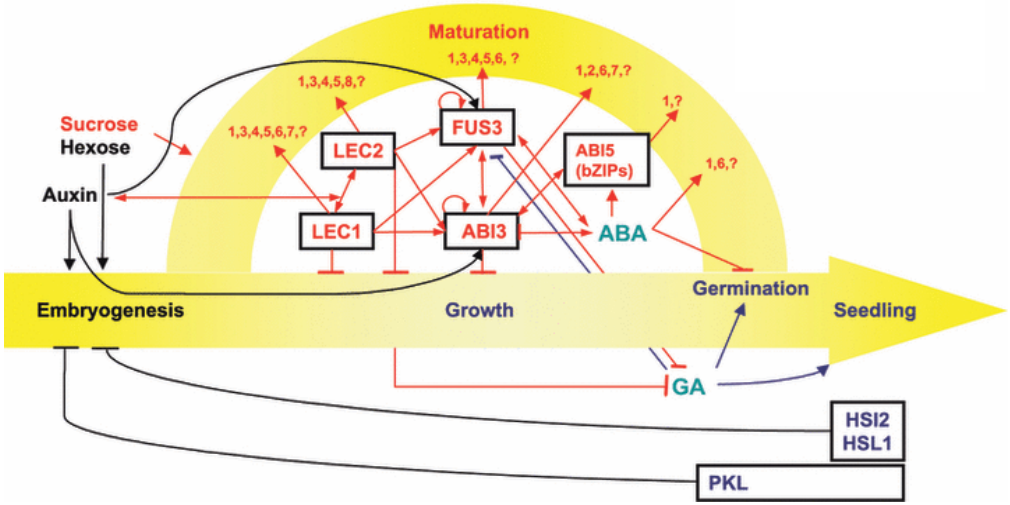

L. Lepiniec
B. Dubreucq
A. Fatihi
D. Valsamou

C. Nédellec
P. Bessières
R. Bossy
L. Deléger
E. Chaix
D. Valsamou

P. Zweigenbaum
D. Valsamou

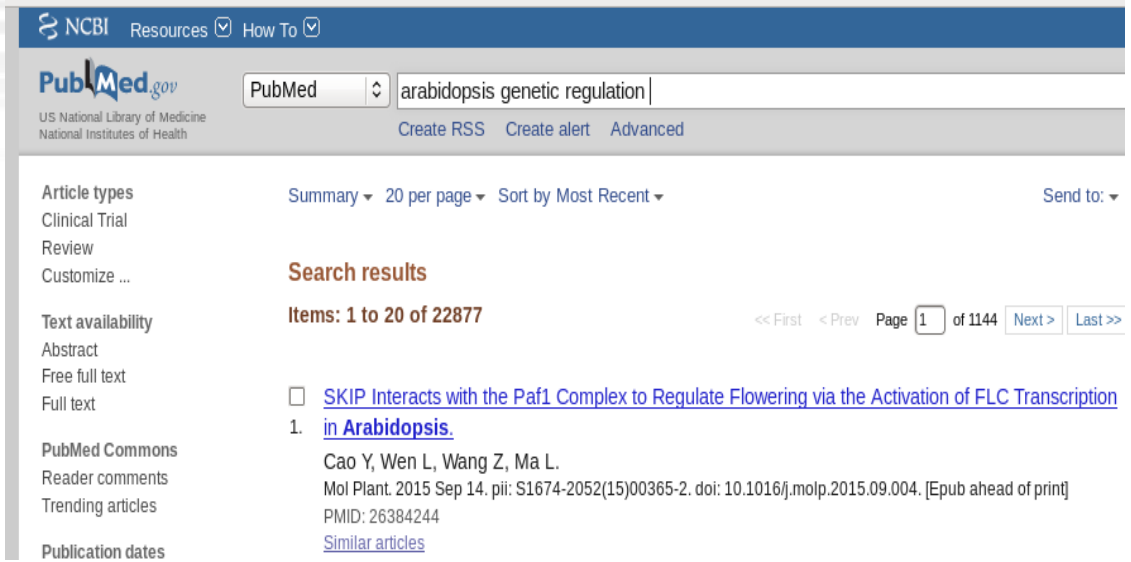
# Biological motivations

- *A.thaliana*, a model plant
  - ➢ From *A.th* to other species
- Seed development: industrial, agricultural and fondamental research interest
- Complex regulatory networks that involve genetic, molecular and physiological mechanisms




A model of genetic (framed) and molecular (in blue cyan) interactions involved in the control of seed development and maturation in *Arabidopsis thaliana* (from Santos-Mendoza *et al.,* 2008) (Numbers are biological processes).

# Biological motivations




- Crop improvement: improved selection of novel species with a better knowledge of genetic interactions
- Knowledge harvest of seed development from text data

# Information extraction motivations

- Text mining: extraction of scarce and critical information

  Genetic Regulation Network: some studies but not at this level of complexity

- Requirements
  - ✓ A rich knowledge model
  - ✓ Annotated corpus for relation extraction

➢ Towards Systems Biology

➢ International visibility

➢ Shared task: text-mining community building

# BioNLP Shared Task, information extraction in the biomedical domain

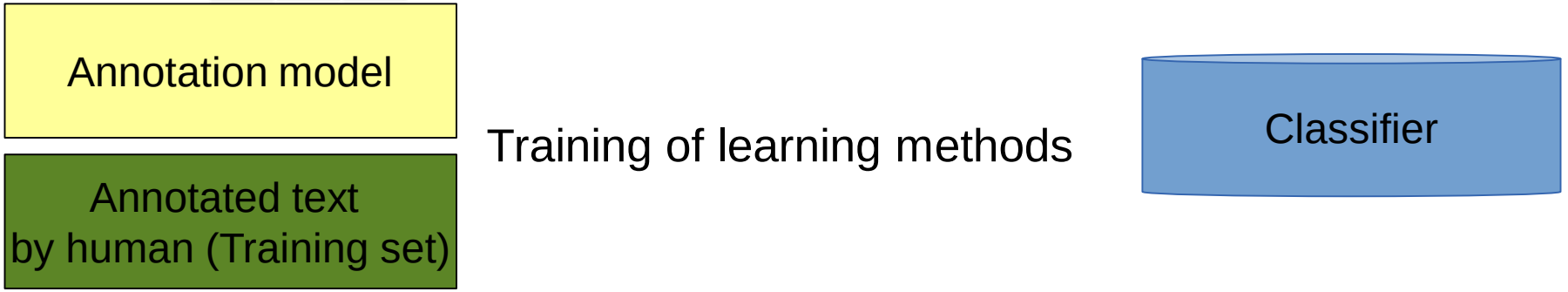
BioNLP-ST'09: 1 task (GENIA)
BioNLP-ST'11: 5 tasks (2 by INRA)
BioNLP-ST'13: 6 tasks (1 by INRA)

- BioNLP-ST'16: 4 tasks (2 by INRA)
  - *Bacteria Biotope*
  - *Plant Seed Development (SeeDev)*

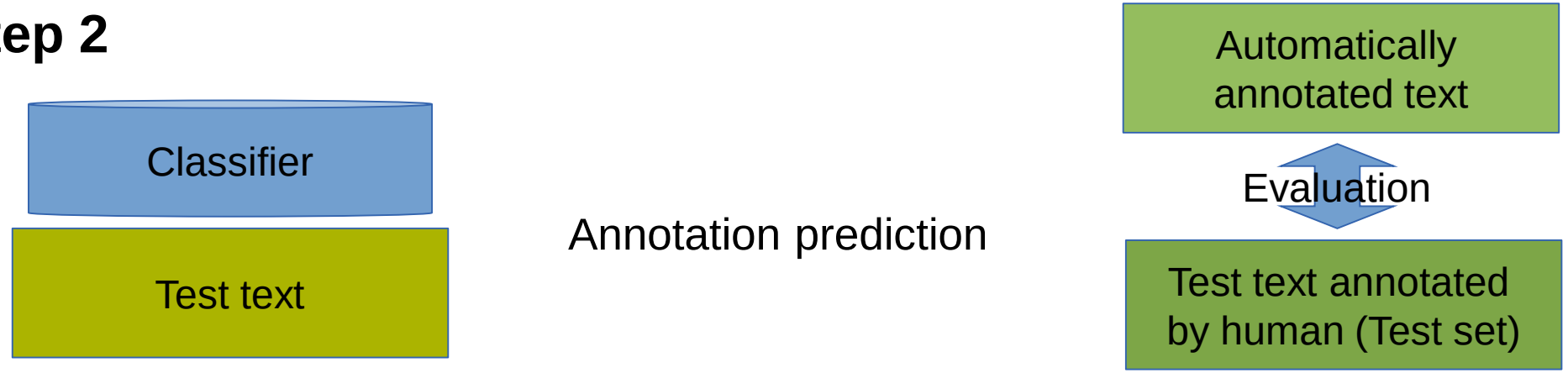

- Knowledge-based contruction from text
- More direct application to Life Science

# Information Extraction Challenge

**Step 1**



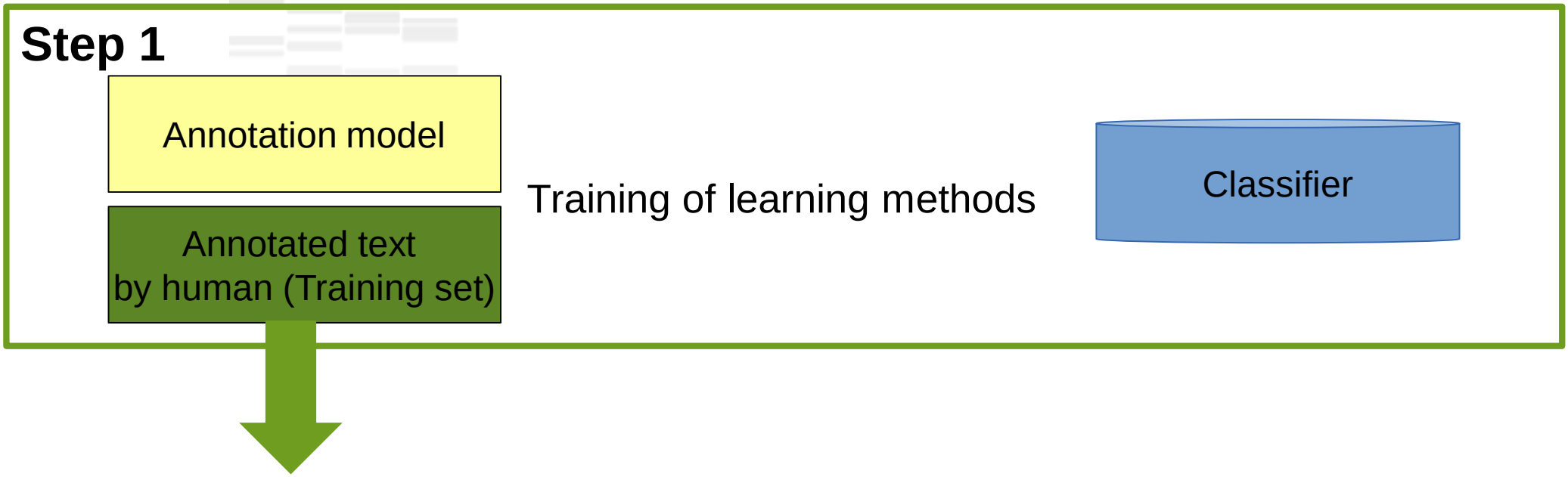

**Step 2**




**Step 3**

Community workshop: 11-12 August 2016
(Association for Computational Linguistics)

# Information Extraction Challenge




- Critical stage, long and complex
- Involved biology experts
- Assisted by knowledge engineer (CDS funding)
- Guidelines

# BioNLP Shared Task Schedule

| | |
|---|---|
| November 2015 | 1st announcement (with sample data set and detailed plans) |
| January 2016 | Training data release and on-line evaluation service open |
| (mid) March 2016 | Evaluation on test data open |
| (end) March 2016 | Evaluation done, notification to participants |
| 8 May 2016 | Workshop paper due |
| 5 June 2016 | Notification of acceptance |
| 22 June 2016 | Camera-ready papers due |
| 11-12 August 2016 | Workshop |

# Knowledge model

## Fined-grained and complex model
## 16 biological entities and 10 events

- Entities
  - Biological Processes
    - DNA-type
      - ***Gene*** "*LEC1*" , "*APETALA2*"
      - ***GeneFamily*** "LEC genes" , "AP2-like"
      - ***Box*** "WRI1 targets" , "5'-GCATCG-3' "
      - ***Promoter*** "*BCCP2* promoter" , "5' flanking regions"
    - DNA Product
      - ***RNA*** "*CLV3* mRNA" , "transcript of *FLC*"
      - ***Protein*** "CLV1" , "LEAFY COTYLEDON 1"
      - ***ProteinFamily*** "MYB" , "B3 proteins"
      - ***ProteinComplex*** "core-binding factor" , "HDAC1 complex"
      - ***ProteinDomain*** "B3 domain" , "basic helix-loop-helix"
      - ***Hormone*** "abscisic acid" , "GA"
    - Process
      - ***Pathway*** "fatty acid biosynthetic pathway" ," glycolysis"
      - ***RegulatoryNetwork*** " sensitivity to ABA" , "embryonic programs"
  - Observed conditions
    - Species and Genotype
      - ***Genotype*** "overexpression of *WRI1*" , "Fus3 mutant"
    - Cell, Tissues and Organ
      - ***Tissue*** "cotyledon" , "petioles of the rosette leaves"
    - Growth stage
      - ***Development phase*** "embryogenesis" , "transition to flowering"
    - External factor
      - ***Environmental factor*** "long-day conditions" , "cold stress"

10 relation types were defined :

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

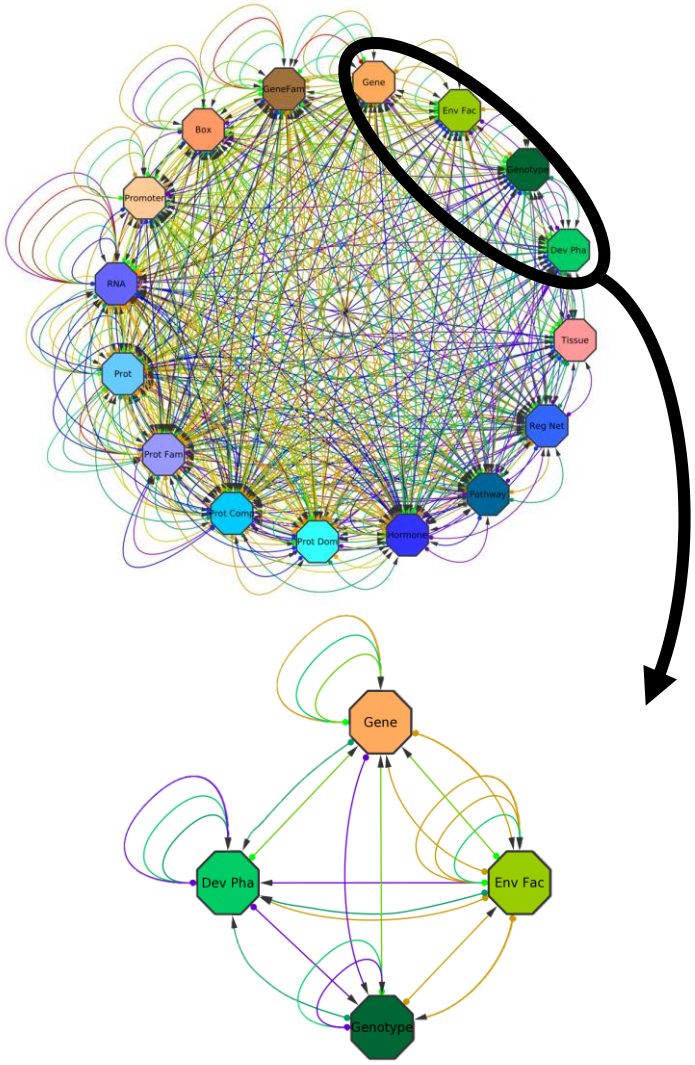
→Interaction :
- *InteractWith*
- *BindTo*

→Localisation :
- *IsFoundIn*
- *IsFoundDuring*

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

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

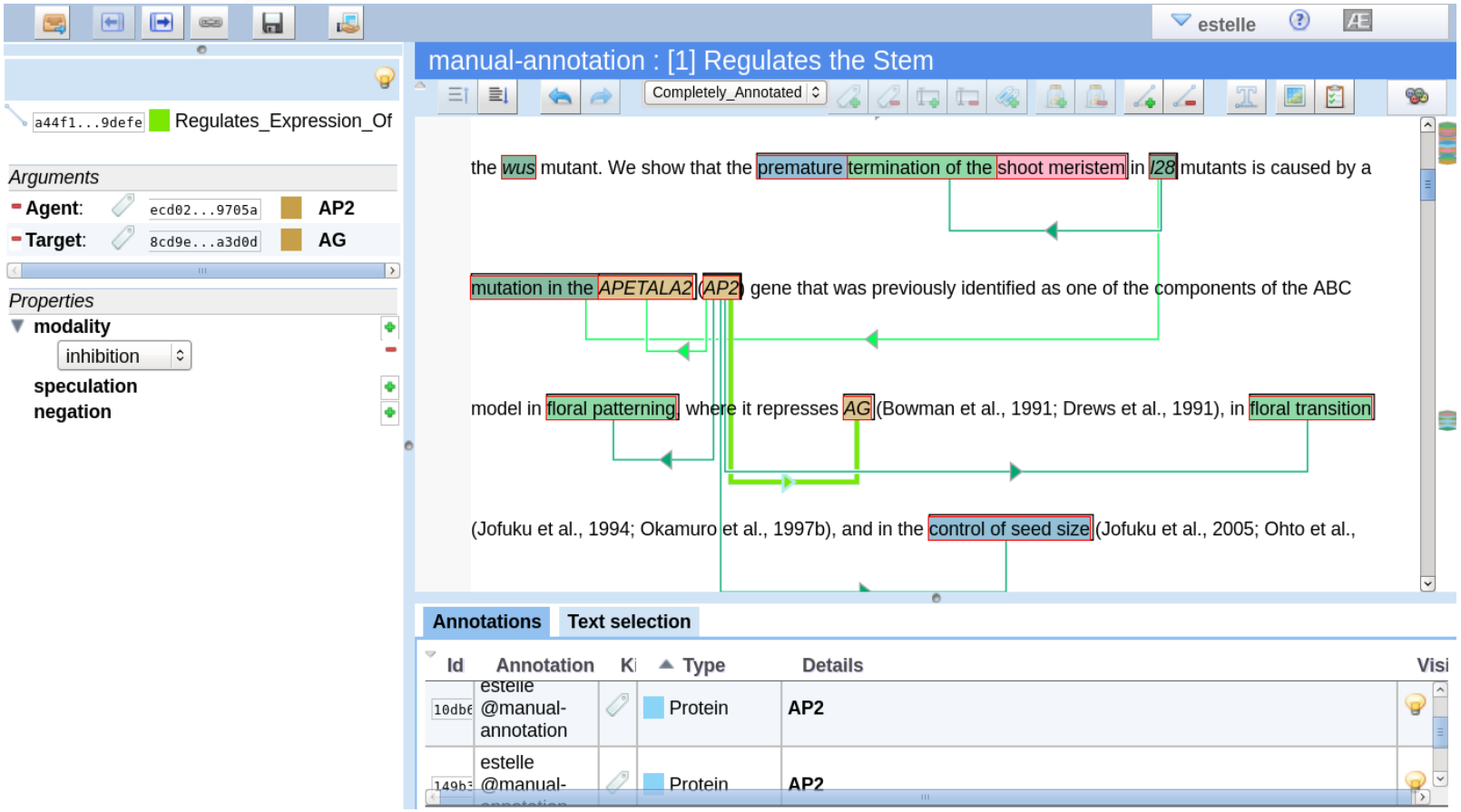
# Gene Regulation Network in *A.th*




- Integration of the knowledge model in the corpus annotation editor, AlvisAE
- Annotation of a corpus of scientific articles by biology experts (on-going)
  Currently 4,444 entities and 1,421 n-ary events.

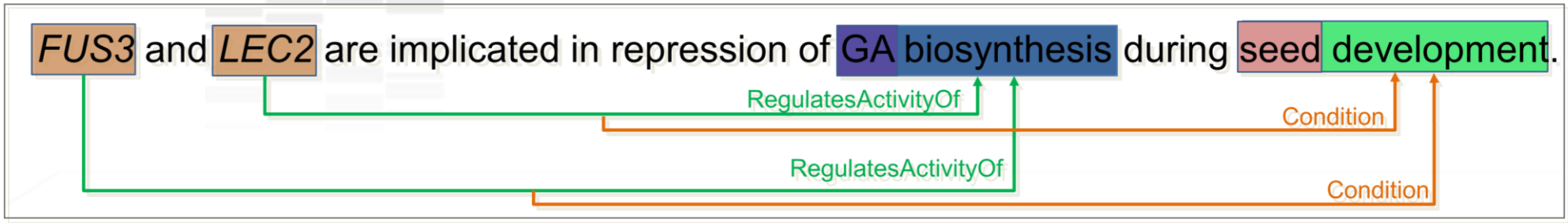
# AlvisAE : corpus annotation editor

- A module of Alvis suite, an integrated software library that provides all the necessary tools for semantic analysis and terminology and ontology acquisition for Ontology-based Data Integration.




(Papazian *et al.*, 2012)

# Example of annotation




**Entities**

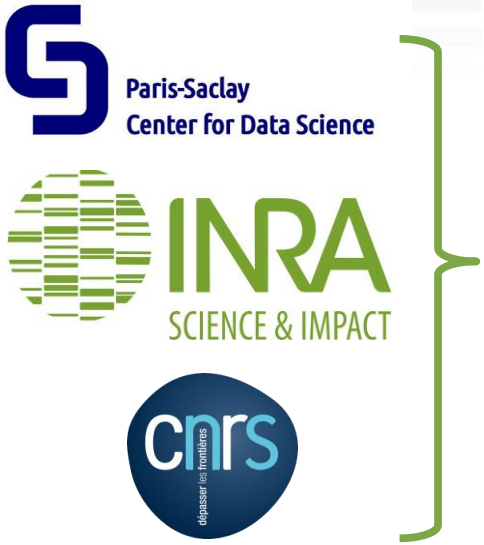
Gene *FUS3* , *LEC2*
Hormone GA
Pathway GA biosynthesis
Tissue seed
Developmental Phase seed development

**Relation**

*Agent* (Gene) regulates_activity_of *Target* (Pathway)
*Agent* (Event) has_as_condition *Target* (Developmental Phase)

- ***FUS3*** *regulates activity of* ***GA biosynthesis*** *with condition* ***seed development***
  (-)

# Further use of *Arabidopsis* corpus




**Model and reference corpus of regulation network for *A.th* seed development**

BioNLP Shared Task 2016 :
*Plant seed development*

Training data and evaluation data for
- relation extraction (D. Valsamou, PhD)
- ontology based annotation (A. Ferré, PhD)

Idex Paris-Saclay grants

Use case plant in text-mining Europeean project
openMINTED

Extension to species of agronomic interest (wheat)

Integration of knowledge extracted in plant models of system biology (on-going project with IMSV)