Text-mining methods used for information extraction in plant scientific papers

4. Entity categorization

Categorization

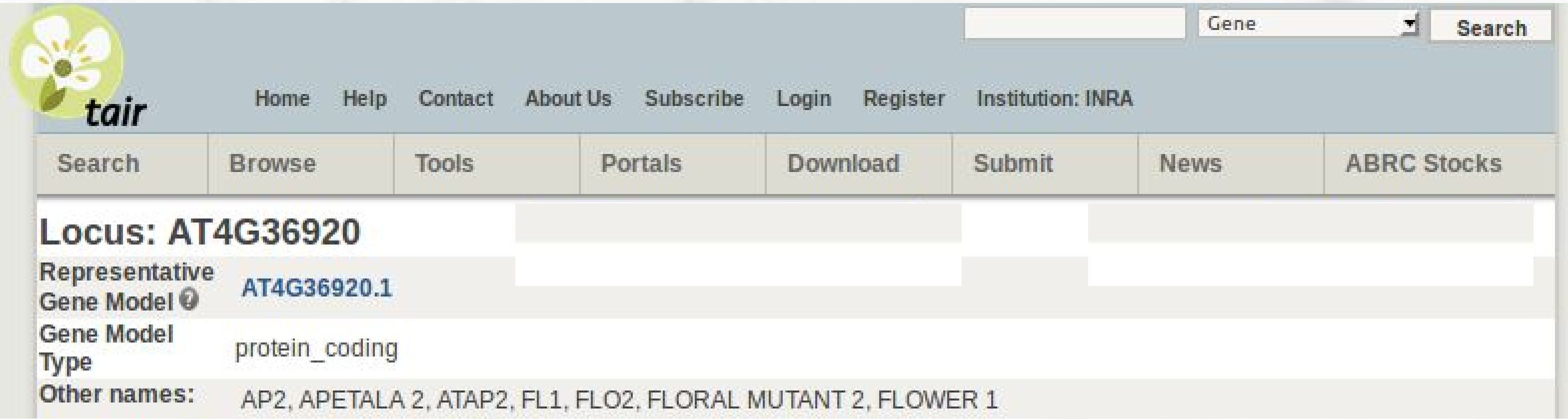
Different entities predicted could be similar to others, and be linked to a single category.

Categorization allows to extract the information and to go back to a superior level of conceptualization compared to the form of the raw text.

Categorization

In the case of synonyms: explains that AP2 and APETALA 2 are the same concept, and could be linked to a unique category such as AT4G36920 (locus of the gene)

- ➢ Improves the query of TDM results
- ➢ Allows link with external resources



Gene
Search
tair
Home Help Contact About Us Subscribe Login Register Institution: INRA
Search Browse Tools Portals Download Submit News ABRC Stocks

Locus: AT4G36920

Representative Gene Model	AT4G36920.1
Gene Model Type	protein_coding
Other names:	AP2, APETALA 2, ATAP2, FL1, FLO2, FLORAL MUTANT 2, FLOWER 1

Ontology for Entity recognition & Categorization

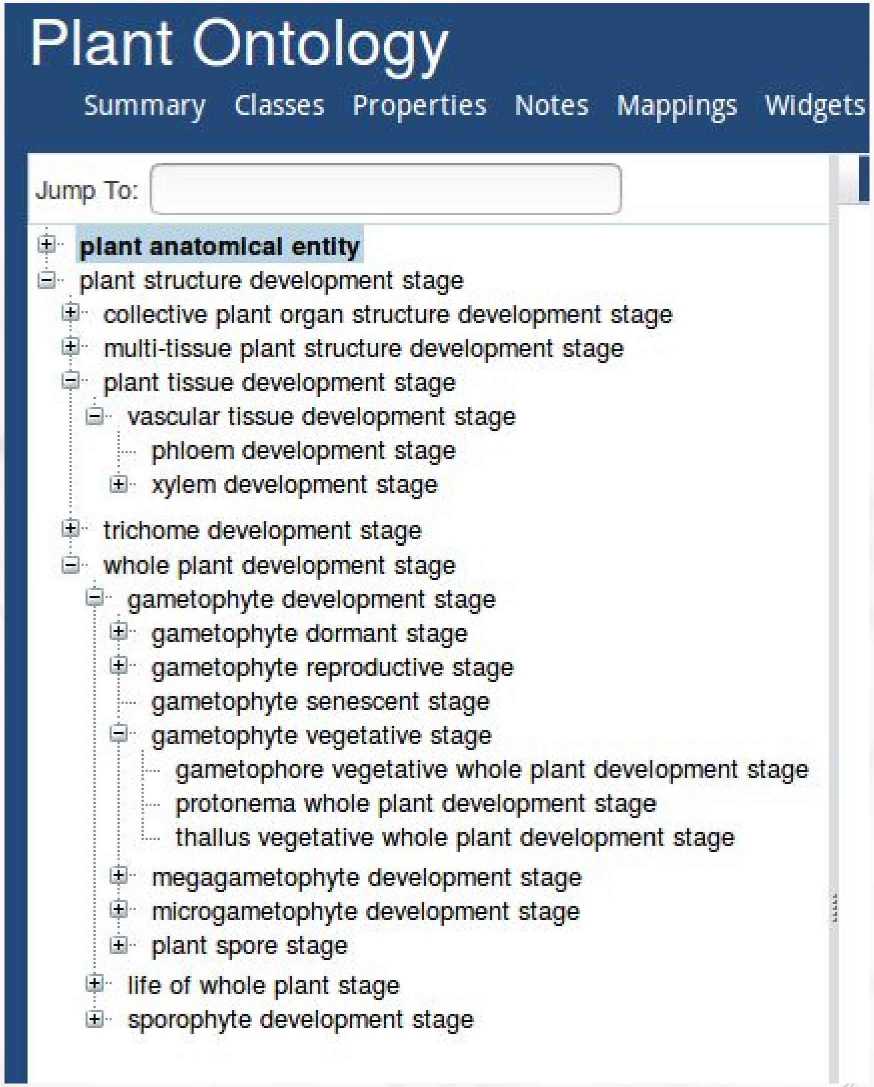
To detect and categorize some complex entities such as Development Phase or Tissue, we use ontologies.

An ontology is a semantic representation of knowledge. Concepts are organized and linked together by different relations, for example "is_a", "part_of", "has_part" ...

Example of an ontology

Plant Ontology

We can reason
and infer knowledge
from ontologies.



Plant Ontology

- **plant anatomical entity**
- plant structure development stage
  - collective plant organ structure development stage
  - multi-tissue plant structure development stage
  - plant tissue development stage
    - vascular tissue development stage
      - phloem development stage
      - xylem development stage
  - trichome development stage
  - whole plant development stage
    - gametophyte development stage
      - gametophyte dormant stage
      - gametophyte reproductive stage
      - gametophyte senescent stage
      - gametophyte vegetative stage
        - gametophore vegetative whole plant development stage
        - protonema whole plant development stage
        - thallus vegetative whole plant development stage
      - megagametophyte development stage
      - microgametophyte development stage
      - plant spore stage
    - life of whole plant stage
    - sporophyte development stage

openMINTED

A plant ontology for TDM use

It will be built to allow detections of complex entities, such as Tissue, Developmental Phase, Pathway ...

It will be used for :

- Entity detection
- Categorization of predicted entities
- The interrogation of the final application