Information extraction from publications on bibliographic resources

Mihalis Papakonstadinou, AgroKnow
AgroHackathon, Montpellier, 2016

The problem

- Vast amounts of research on agriculture available (papers, bibliographic sources).
- These resources contain important metadata information describing it (topic, author, publisher, etc.)
- They sometimes also contain the pdf link to the actual resource.
- The pdf contains information not currently accessed by a system processing agricultural resources (textual, multimedia, etc.).

The suggestion

- Harvest data contained in the pdf files of (agricultural) resources.
- Focus on the extraction of textual information of these files.
- Use the extracted information against known endpoints to semantically the resource.
- Apply modular logic in the design of all the elements of the suggested hack.

Tools

- Dataset on viticultural resources having a pdf link. Available here. Currently this is manually annotated.
  - Can be used for testing and evaluation purposes.
- AgroPortal annotator.
- Crop Ontology JSON api, with the Vitis ontology, available here.
- MIMOS SPARQL endpoint, with the AGROVOC vocabulary.
- SPARQL endpoint with the grape varieties ontology.
- Freme API for semantically enriching text on known vocabularies.
- Any other proposition!

Benefits of Hack

- Having a generic mechanism for the extraction of (textual) data, regardless of the domain.
- Using known open endpoints for the semantic enrichments in an add-on fashion.
- Presenting end users with an intuitive way of uploading their pdf files and automating the rest of the process.