INTEGRATION OF ALVIS WITHIN THE OPENMINTED PLATFORM AS GALAXY UTILITIES

Jean-Baptiste Bohuon*, Mouhamadou Ba*, Estelle Chaix, Robert Bossy, Claire Nédellec
MaIAGE, INRA, Université Paris-Saclay, 78350, Jouy-en-Josas, France
firstname.lastname@inra.fr, *: corresponding authors

MOTIVATIONS

- Over 50 million scholarly publications are available
- Text and Data Mining (TDM) offers a set of mature softwares to extract knowledge from large corpora
- Domain-specific terminologies and ontologies enable normalization
- Alvis is an open-source modular TDM suite in Java
- Alvis performs named-entity recognition and normalization, and relation extraction

→ The integration of Alvis as a set of Galaxy tools will help the development of TDM workflows by non-experts

→ The integration of several TDM suites (UIMA, GATE) will make available state of the art TDM within the OpenMinTeD platform

TEXT MINING IN BIOLOGY

- Domain-specific text mining involves the use of TDM programs and domain resources
- A classifier is trained with manual annotations
- TDM example: named-entity recognition and relation extraction in a corpus about seed development of the *Arabidopsis thaliana* (Fig. 1, 2)

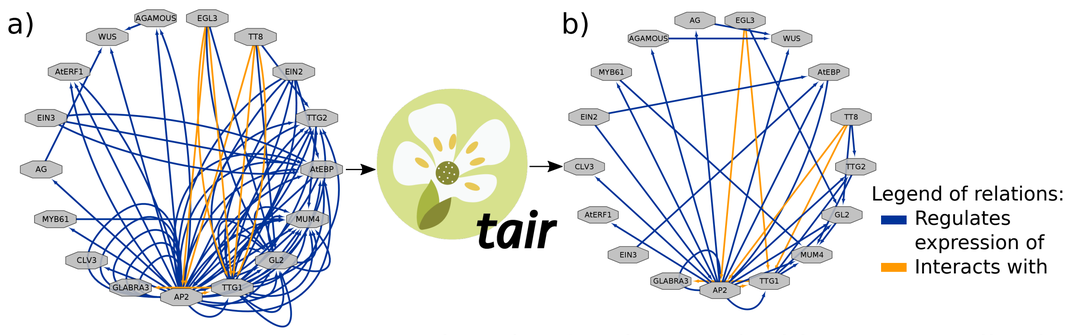

Fig. 1: a) A set of manual annotations (on 21 abstracts) with non-normalized entities. b) Entities and hence occurrences of relations are normalized using TAIR, a resource on *Arabidopsis thaliana*.

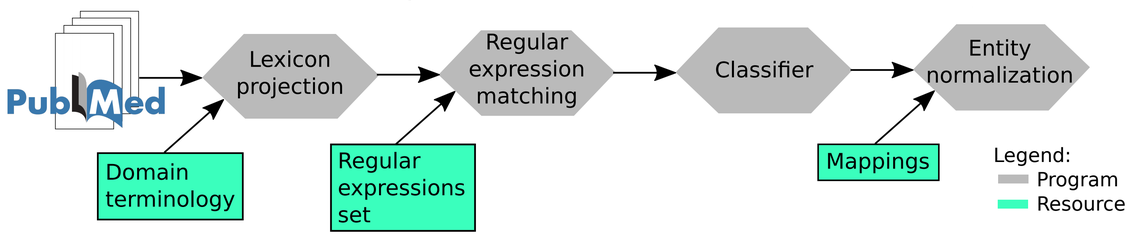

Fig. 2: Chained execution of TDM tools including the trained classifier predicts named-entities and relations on a relevant corpus (5000 abstracts).

- Manually annotated corpora are shared to train different models
- Successive tools constitute workflows, generic sequences of workflows are widely reused and domain specific ones are built *ad hoc*
- The extracted knowledge may be queried, *e.g.* a biologist performing sequence functional annotation of sequences may couple sequence similarity search and literature mining
- Workflow steps may be shared among applications and beyond the application domain

OpenMinTeD: A TDM COMMUNITY

- OpenMinTeD is a H2020 E-INFRA project (2016-2018)
- Major TDM platforms are involved: GATE, UIMA, Alvis, Argo, DKPro
- Covered domains are Agricultural Science, Biodiversity Science, Social Sciences, Scholarly Communications
- The main goals are 1) interoperability between TDM modules from different platforms, and 2) to enable the design of workflows made of tools from different platforms
- All developments will be open source

WORKFLOW ENGINES OF ALVIS AND GALAXY

- OpenMinTeD has choosen Galaxy as workflow engine
- The workflow engines of Alvis and Galaxy differ, each Alvis module can be wrapped as a Galaxy tool

Feature	Alvis	Galaxy
Execution of workflows	Sequential	Parallel
Workflow description	XML parameter-oriented	JSON with tool state
Data interoperability	Shared data structure	Native formats, shims
Data persistance	RAM, dedicated modules	Filesystem
Interfaces	CLI, REST API , Java API	GUI, REST API

ALVIS MODULES IN GALAXY

- We created a generic Alvis docker image
- Data interoperability within OpenMinTeD: RDF for documents and XMI for text annotations
- An RDF/XMI export module is under development
- Most Alvis workflows use a docker container giving standalone Galaxy tools
- Modules will be progressively wrapped as independent tools

COMMUNITY-DRIVEN WORKFLOWS

- Existing state of the art TDM workflows will be released on the OpenMinTeD platform as Galaxy workflows
- These workflows take domain-specific resources as input (corpora and semantic resources)
- Workflows in Agricultural Science and Biodiversity Science include:

Domain	Plant development	Wheat data	Microbial diversity
Client application	FLAGdb++	GnPIS, WheatIS	Florilège
Partner	IPS2 (INRA)	URGI (INRA)	E-infra partners
Main TDM resource	TAIR	WIPO	OntoBioTope

- The bioinformatics applications associated with workflows (FLAGdb++, GnPIS, Florilège) will integrate TDM results and bioinformatics data
- Bioinformatics applications will execute the preconfigured workflows
- Periodical execution: corpora and resource update and user feedback will trigger execution
- CORE, OpenAire, AgroPortal will provide access to data

DISCUSSIONS

- OpenMinTeD tools metadata are wider than Galaxy metadata
- Resources (terminologies and corpora) are widely reused

→ hence a registry of resources and components is under development

- The analysis of large corpora may exploit Galaxy's load balancing system or *ad hoc* preprocessing steps
- Several software packagers are considered: Docker, Maven or exposing programs as webservices

REFERENCES

- Ba *et al.* (2016). *Interoperability of corpus processing workflow engines: the case of AlvisNLP/ML in OpenMinTeD*. Presented at Meeting of working Group Medicago sativa, Portoroz, SVN.
- Chaix et al. (2016). *Overview of the regulatory network of plant seed development (SeeDev) task at the BioNLP shared task 2016.* In: *Proceedings of the 4th BioNLP Shared Task Workshop* (p. 12-22). Presented at BioNLP Shared Task, Berlin, DEU. Stroudsburg, USA : The Association for Computational Linguistics.
- Alvis git repository http://github.com/Bibliome/alvisnlp
- OpenMinTeD git repository http://github.com/openminted
- Alvis generic docker image http://github.com/openminted/alvis-docker

Permalink

www.openminted.eu