

AgroLD indexing tools with ontological annotations

Stella Zevio^{1,2}, Nordine El Hassouni^{2,3}, Manuel Ruiz^{2,3} and Pierre Larmande^{2,3}

¹ Université de Montpellier, Montpellier, France,

² Institut de Biologie Computationnelle, Montpellier, France,

³ South Green Bioinformatics Platform, CIRAD, IRD, Montpellier, France

`pierre.larmande@ird.fr`

Abstract. The Agronomic Linked Data project (AgroLD) is a Semantic Web knowledge base designed to integrate data from various publicly available plant centric data sources. The aim of AgroLD project is to provide a portal for bioinformaticians and domain experts to exploit the homogenized data towards enabling to bridge the knowledge. Here we present new tools that enable "full text search" functionalities with Elastic clusters and enhance data annotation with ontologies.

Keywords: Plant Molecular Biology, Linked Data, Elastic

1 Introduction

Agronomy is an overarching field constituting various research areas such as genetics, plant molecular biology, ecology and earth science. The last several decades has seen the successful development of high-throughput technologies that have revolutionized and transformed agronomic research. The application of these technologies have generated large quantities of data and resources over the web. In most cases these sources remain autonomous and disconnected. The Agronomic Linked Data project (AgroLD) is a Semantic Web knowledge base designed to integrate data from various publicly available plant centric data sources. These include Gramene, Oryzabase, TAIR and resources from the South Green platform among many others. The conceptual framework for the knowledge in AgroLD is based on well-established ontologies: Gene Ontology, Plant Ontology, Plant Trait Ontology (TO) and Plant Environment Ontology (EO). The current phase (phase one) covers information on genes, proteins, ontology associations, homology predictions, metabolic pathways, plant traits, and germplasm. Information on the integrated databases, ontologies and identifiers can be found in the documentation page (<http://www.agrold.org/documentation.jsp>).

The RDF knowledge bases are accessed via SPARQL endpoints but these endpoints are more suitable for programmatic access. This requires at the minimum

a moderate knowledge of SPARQL which is not usually suitable to the non-technical users (biologists). Consequently, the Semantic Web resources not being exploited completely. Alternatively, the AgroLD website provides four entry points to access the underlying knowledge:

1. **Quick Search**, a faceted search plugin made available by Virtuoso, that allows users to search by keywords to browse the knowledge contained in AgroLD.
2. **SPARQL Query Editor**, that provides an interactive environment to formulate SPARQL queries. The SPARQL editor is based on YASQE and YASR tools [5].
3. **Explore Relationships visualizer**, an implementation of RelFinder [1] that allows the user explore and visualize existing relationships between entities.
4. **Advanced Search**, a query form providing entity (e.g. gene) specific information retrieval. The Advanced Search query form is based on the REST API suite developed under the AgroLD project.

To further enable the Quick Search functionality by retrieving more textual information and hiding the technical details, we developed an indexation tool that enables to communicate easily with Elastic clusters. Thus, this tool enables to index Json files and to manage indexes (i.e. update, delete) on Elastic clusters without using cURL. Starting from RDF graphs in AgroLD, we used this tool to automate the indexation task and set up Elastic clusters. Furthermore, we developed a generic annotation tool that enables communication with NCBO Annotator [2] to annotate JSON files with ontologies available from such portals (i.e. BioPortal [4] and AgroPortal [3]). We use this tool with AgroPortal Annotator to enrich and index the Json files with additional information from ontological terms such as labels, synonyms, parent and child terms, etc.

References

1. P. Heim *et al.* 2009. RelFinder: Revealing Relationships in RDF Knowledge Bases. In Lecture Notes in Computer Science, 5887 LNCS:18287.
2. C. Jonquet *et al.* 2009. The Open Biomedical Annotator, in American Medical Informatics Association Symposium on Translational BioInformatics, AMIA-TBI09, (San Francisco, CA, USA), pp. 5660.
3. C. Jonquet *et al.* 2015. AgroPortal : a proposition for ontology-based services in the agronomic domain. IN-OVIVE'15, Jun 2015, Rennes, France.
4. N. F. Noy *et al.* 2009. BioPortal: ontologies and integrated data resources at the click of a mouse, Nucleic Acids Research, vol. 37, pp. 170173.
5. L. Rietveld, and R. Hoekstra. 2015. The YASGUI Family of SPARQL Clients. Semantic Web Journal 0: 110.