AgroPortal : a proposition for ontology-based services in the agronomic domain

Clément Jonquet,[1,2] Esther Dzalé-Yeumo,[3]
Elizabeth Arnaud,[4] Pierre Larmande[2,5]

[1] Laboratory of Informatics, Robotics and Microelectronics of Montpellier (LIRMM)
University of Montpellier & CNRS
jonquet@lirmm.fr
[2] Computational Biology Institute (IBC) of Montpellier
[3] INRA Versailles
edzale@versailles.inra.fr
[4] Bioversity International, Montpellier
e.arnaud@cgiar.org
[5] UMR DIADE, IRD Montpellier
pierre.larmande@ird.fr

Abstract : Our project is to develop and support a reference ontology repository for the agronomic domain. By reusing the NCBO BioPortal technology, we have already designed and implemented a prototype ontology repository for plants and a few crops. We plan to turn that prototype into a real service to the community. The AgroPortal project aims at reusing the scientific outcomes and experience of the biomedical domain in the context of plant, agronomic and environment sciences. We will offer an ontology portal which features ontology hosting, search, versioning, visualization, comment, but we will also offer services for semantically annotating data with the ontologies, as well as storing and exploiting ontology alignments and data annotations. All of these within a fully semantic web compliant infrastructure. The main objective of this project is to enable straightforward use of agronomic related ontologies, avoiding data managers and researchers the burden to deal with complex knowledge engineering issues to annotate the research data. The AgroPortal project will specifically pay attention to respect the requirements of the agronomic community and the specificities of the crop domain. We will first focus on the outputs of a few existing driving agronomic use cases related to rice and wheat, with the goal of generalizing to other Crop Ontology related use cases. AgroPortal will offer a robust and stable platform that we anticipate will be highly valued by the community.

Keywords : ontology management, ontology mapping, semantic annotation, agronomic sciences.

1 Introduction

Agronomy, environmental and plant sciences are complementary disciplines that must combine the data they generate into meaningful information and interoperable knowledge potentially leading to innovation. Undeniably, data integration and semantic interoperability is necessary to enable new scientific discoveries that could be made by merging different available data. A key aspect in addressing semantic interoperability is the use of ontologies as a common denominator to describe data, make them interoperable and turn them into structured and formalized knowledge. Biomedicine has always been a leading domain for semantic interoperability [9] leading to the development of reference ontologies such as the Gene Ontology. This has served as model for the agronomic, environmental and plant sciences e.g., Plant Ontology [3], Crop Ontology [10]. However, there exist a need of a one-stop-shop for plant related ontologies offering generic services to exploit them in search, annotation or other scientific data management processes.

In the biomedical domain, the NCBO Bioportal (http://bioportal.bioontology.org) is a well-known open repository for biomedical ontologies originally spread out over the web and in different formats [8, 12]. There are 437 ontologies in this collection as of April 2015, but only a few are relevant for the agronomy community. By using the portal's features, health professionals and biologists can browse, search, visualize and comment on ontologies both interactively through a user web interface and programmatically, via web services. Within BioPortal, ontologies are used to develop an annotation workflow [6] that indexes several biomedical (text) data resources (in English) using the knowledge formalized in ontologies in order to provide semantic search features that enhance information retrieval experience [5]. The NCBO BioPortal functionalities have been progressively extended in the last 8 years, and the platform is fully semantic web compliant (ontologies, mappings and annotations are stored in an RDF triple store). However, the BioPortal is specific for health and biomedical ontologies and even if overlaps exist,[1] the portal does not span to the agronomic or plant domain. However, NCBO technology is domain-independent and open source. A BioPortal virtual appliance[2] is available as a server machine embedding the complete code and deployment environment, allowing anyone to set up a local ontology repository and eventually customize it.

Similarly to what happens in biomedicine, communities engaged in agronomic research need to access specific sets of ontologies for plant data annotation and integration. For instance, the Crop Ontology web application (www.cropontology.org) publishes online sets of ontologies required for describing crop germplasm, traits and evaluation trials. It contains 18 species-specific ontologies in addition to ontologies related to the crop germplasm domain.[3] In addition to its role as a repository, the Crop Ontology web application offers community-oriented features such as an Excel template for trait submission, addition and filtering of new terms, etc. A web API provides all necessary services to third party users like the Global Evaluation Trials Database that stores 35,000 trial records. Efforts have been made to structure and formalize the crop-specific ontologies following semantic web standards as well as offering collaborative ontology enrichment and annotation features. The current web application facilitates the complete ontology-engineering life cycle starting with collaborative construction, publishing, use and modification. However, it necessitates important improvements of the current versioning, curation, multilingual aspects, user interface as well as for data annotation and mapping features. For instance, it is important to support the alignment (or mapping) of terms within and across different ontologies: both within the Crop Ontology itself (in different crop branch) or with other top level ontologies commonly used in plant biology (Plant Trait Ontology, Plant Ontology, Pato, Environment Ontology) that will be maintained and extended within the Planteome project (www.planteome.org). Because of the custom requirements and the specificity of the domain, the Crop Ontology is referenced on the NCBO Bioportal aside other top level plant-related ontologies, but is not currently fully accessible and described through this portal (i.e., the ontology is listed but not uploaded and available through services).

2 Motivation

The main objective of the AgroPortal project is to develop and support a reference ontology repository for the agronomic domain, by reusing NCBO technology. There are two major motivations for AgroPortal to reusing the outcomes of the biomedical domain: (i) to avoid re-developing technologies and tools that have already been designed and extensively used; (ii) to offer the same tools, services and formats in both domains to facilitate the interface and interaction between the domains e.g., to enable a user to query the BioPortal or the AgroPortal without changing a line of code.

---

[1] BioPortal already host ontologies such as GO, ENVO, PATO, PO, PTO that are relevant for the domain of agronomy in addition of upper level ontologies such as SIO, BFO.

[2] http://www.bioontology.org/wiki/index.php/Category:NCBO_Virtual_Appliance

[3] Partners like the US Department of Agriculture (USDA), INRA and the Polish Genomic Network have uploaded ontologies. The International Wheat Initiative and data interoperability group have approached the Crop Ontology team to discuss best practices. In addition, the Crop Ontology web API is used by other project such as Agtrials, IBP, EU-SOL.

With the purpose of offering a customized semantic annotation workflow to the partners of the *Computational Biology Institute of Montpellier*, in order to annotate experimental data to enable interpretation, comparison, and discovery across databases [1]; we have already set-up a local instance of BioPortal on a LIRMM server. It is a prototype that currently hosts 15 plant related ontologies (including 10 ontologies not originally present in BioPortal). A specific group of ontologies has been also setup for ontologies produced or used by INRA. The new new annotation workflow currently relies on the same technology than the NCBO Annotator's [6] and can report same evaluation; however, we will be running soon new evaluations with some ontologies specific to AgroPortal. In addition, we have implemented within WebSmatch (an open environment for matching complex schemas from many heterogeneous data sources [2]) the possibility to call either the NCBO Annotator web service or the new annotator included in our virtual appliance. In parallel, we have also started the integration of multiple agronomic datasets (such as SouthGreen, Gramene, OryGeneDB) in RDF and we store the annotations within the local instance of BioPortal triple store, nearby the ontologies they reference [11].

Our goal now, is to develop and support a reference ontology repository for the agronomic domain and offer a robust and reliable service to the community that will feature ontology hosting, search, versioning, visualization, comment, but we will also offer services for semantically annotating data with the ontologies, as well as storing and exploiting ontology alignments and data annotations.

3 Proposition for a portal for agronomic related ontologies

The features offered by the portal are for examples: (i) to search across all the ontologies, (ii) to annotate a piece of text with all the ontologies, (iii) to store and serve mappings between ontologies within the portal and with the main BioPortal. All other features from BioPortal are generically available for the AgroPortal: ontology versioning, UI widget, ontology metrics, ontology recommender service, projects listing, community feedback (comment, subscription to ontology changes), users' management (and public or private access to ontologies). In addition, two endpoints allow automatic querying of the content of the portal: (i) a REST web service API and (ii) a SPARQL endpoint.

While assuring the day to day maintenance and monitoring of the portal and keeping it up-to-date with the NCBO technology, we will constantly implement new customizations and specific services for the agronomic/plant community. For instance, in the context of the SIFR project and in collaboration with the NCBO, we will implement specifications that have been proposed to make BioPortal multilingual [4]: handle multilingual ontologies that offer labels in different languages (either using basic xmllang tag or rich lexical enrichment vocabularies such as Lemon [7] and deal with monolingual ontologies and the translation mappings betweem them. We will make an inventory of the appropriate ontologies (Table 1) to host in the repository and develop the appropriate ontology wrappers to process and load in the portal specific formats not currently handled (e.g., XLS).

Table 1 - Ontologies of interest for the agronomic community, not present in the NCBO BioPortal and candidates (currently 2 are included) for the AgroPortal.

Title	Acronym	Group	#classes	Multilingual
Rice trait ontology	CO-RTO	CO	488	Partial
Wheat trait ontology	CO-WTO	CO	640	Partial
Wheat Plant Anatomy & Development Ontology	CO-WPA	CO	77	Partial
Multicrop passport ontology	CO-MPO	CO	87	Partial
ICIS germplasm methods ontology	CO-GMO	CO	166	Partial
Agrovoc	AGROVOC	FAO	~35K	Yes
Anaee thesaurus	ANAEE	ANAEE	3343	Partial
National Agriculture Library thesaurus	NALT	NAL	~58K	No
Bioversity molecular markers ontology	BMMO	-	173	No
Feature annotation location description ontology	FALDO	-	22	No

4 Driving agronomic use cases

4.1 Rice

The *Computational Biology Institute of Montpellier* (IBC – http://www.ibc-montpellier.fr), which aims to develop methods to aid data integration and knowledge management within the domain of agronomic sciences to improve information accessibility and interoperability will offer the first AgroPortal use case, related to rice. We will: (i) identify genes controlling roots and panicle branching; (ii) identify genes orthologous relationship for rice genes families. Data from 3k genomes or IRIGIN (International RIce Genomics Initiative) projects might be used.

4.2 Wheat

The *Wheat Data Interoperability working group* is part of the Research Data Alliance (RDA – https://rd-alliance.org). Its goal is to provide a common framework for describing, representing, linking and publishing wheat data with respect to open standards. One of the need identified by the group is to offer a repository of linked vocabularies and ontologies that are relevant for wheat. NCBO technology has been identified as suitable tool to address this need allowing one to search for terms across multiple vocabularies and ontologies, browse mappings between terms, receive recommendations on which vocabularies and ontologies are most relevant for a corpus and annotate text with terms. The group currently test the current AgroPortal prototype. This second use case is achieved in collaboration with INRA and in relation with the URGI platform (https://urgi.versailles.inra.fr).

4.3 INRA Linked Open Vocabularies (LovInra)

LovInra is an effort to publish vocabularies produced or co-produced by INRA scientists and foster their reuse beyond the original researchers. Many of such resources remain unknown to the research community despite of their value. To achieve this goal, there is a clear need to publish the vocabularies with respect to open standards and link them to existing resources. Here again, NCBO technology has been identified a suitable repository for this third used case.

4.4 The Crop Ontology

The *Crop Ontology project* (www.cropontology.org) of the Consultative Group on International Agricultural Research (CGIAR) will offer the fourth use case. The main goals of this project are: to publish online fully documented lists of breeding traits used for producing standard field books; and to support data analysis and integration of genetic and phenotypic data through harmonized breeders' data annotation. The project also offers a forum for scientists to discuss their variables, methods and scales of measurement, and fieldbooks. We will work on leveraging the backend of the cropontology.org application with the AgroPortal web service API, while keeping the current web application as the primary point of access. This will offer new functionalities to the crop ontology community such as versioning, SPARQL endpoint, notes, the annotation tool, but not break the uses of the current application. In addition, we will work on supporting the alignment (or mapping) of terms within and across different plant related ontologies: both within the crop ontologies themselves (in different crop branch) or with other global ontologies commonly used in plant biology.

5 Conclusions

Considering the position of the current NCBO BioPortal and the importance of such an equivalent repository of ontologies for the agronomic, environment and plant sciences, we therefore expect a broad adoption of the AgroPortal in the community. The implication of associated partners (IBC, IRD, CIRAD, INRA, Bioversity International) illustrate the impact and interest first in France, but also internationally (e.g., Planteome, iPlant Collaborative (www.iplantcollaborative.org) or Elixir (www.elixir-europe.org) projects). Making available such a portal will allow, we expect, the researchers to focus on the development of new ontologies and mappings between ontologies with the perspective of leveraging them in their

research and not being afraid of producing an additional piece in the big data cake. Exporting NCBO research results and technology will contribute to long term support of that technology while reinforce the connections with the biomedical domain. Besides, the AgroPortal project has also an important theoretical research dimension that will be addressed within the different involved research groups: semantic annotation, multilingual ontologies, gene annotations, phenotypic data acquisition, ontology engineering, etc.

The first version of AgroPortal will be released by the summer 2015 at the following URL: http://agroportal.lirmm.fr and the API and SPARQL endpoints will be respectively accessible at http://data.agroportal.lirmm.fr and http://sparql.agroportal.lirmm.fr.

Acknowledgements

This work is partly achieved within by Semantic Indexing of French biomedical Resources (SIFR – www.lirmm.fr/sifr) project funded by the French National Research Agency, grant ANR-12-JS02-01001, the NUMEV Labex (www.lirmm.fr/numev), grant ANR-10-LABX-20, the Computational Biology Institute of Montpellier (IBC – www.ibc-montpellier.fr), grant ANR-11-BINF-0002 as well as by University of Montpellier and the CNRS.

References

[1] Castanier, E., Jonquet, C., Melzi, S., Larmande, P., Ruiz, M., and Valduriez, P. Semantic Annotation Workflow using Bio-Ontologies. In *Workshop on Crop Ontology and Phenotyping Data Interoperability* (Montpellier, France, April 2014), CGIAR.

[2] Coletta, R., Castanier, E., Valduriez, P., Frisch, C., Ngo, D., and Bellahsene, Z. Public data integration with Websmatch. In *1st International Workshop on Open Data, WOD'12* (Nantes, France, May 2012), G. Raschia and M. Theobald, Eds., ACM, pp. 5–12.

[3] Cooper, L., Walls, R. L., Elser, J., Gandolfo, M. A., Stevenson, D. W., Smith, B., Preece, J., Athreya, B., Mungall, C. J., Rensing, S., Hiss, M., Lang, D., Reski, R., Berardini, T. Z., Li, D., Huala, E., Schaeffer, M., Menda, N., Arnaud, E., Shrestha, R., Yamazaki, Y., and Jaiswal, P. The Plant Ontology as a Tool for Comparative Plant Anatomy and Genomic Analyses. *Plant and Cell Physiology 54*, 2 (December 2012), e1.

[4] Jonquet, C., Emonet, V., and Musen, M. A. Roadmap for a multilingual BioPortal. In *4th Workshop on the Multilingual Semantic Web, MSW4'15* (Portoroz, Slovenia, June 2015).

[5] Jonquet, C., LePendu, P., Falconer, S., Coulet, A., Noy, N. F., Musen, M. A., and Shah, N. H. NCBO Resource Index: Ontology-Based Search and Mining of Biomedical Resources. *Web Semantics 9*, 3 (September 2011), 316–324. 1st prize of Semantic Web Challenge at the 9th International Semantic Web Conference, ISWC'10, Shanghai, China.

[6] Jonquet, C., Shah, N. H., and Musen, M. A. The Open Biomedical Annotator. In *American Medical Informatics Association Symposium on Translational BioInformatics, AMIA-TBI'09* (San Francisco, CA, USA, March 2009), pp. 56–60.

[7] McCrae, J., Spohr, D., and Cimiano, P. Linking lexical resources and ontologies on the semantic web with lemon. In *8th Extended Semantic Web Conference, ESWC'11* (Heraklion, Crete, Greece, May 2011), no. 6643 in Lecture Notes in Computer Science, Springer, pp. 245–259.

[8] Noy, N. F., Shah, N. H., Whetzel, P. L., Dai, B., Dorf, M., Griffith, N. B., Jonquet, C., Rubin, D. L., Storey, M.-A., Chute, C. G., and Musen, M. A. BioPortal: ontologies and integrated data resources at the click of a mouse. *Nucleic Acids Research 37*, (web server) (May 2009), 170–173.

[9] Rubin, D. L., Shah, N. H., and Noy, N. F. Biomedical ontologies: a functional perspective. *Briefings in Bioinformatics 9*, 1 (2008), 75–90.

[10] Shrestha, R., Arnaud, E., Mauleon, R., Senger, M., Davenport, G. F., Hancock, D., Morrison, N., Bruskiewich, R., and McLaren, G. Multifunctional crop trait ontology for breeders' data: field book, annotation, data discovery and semantic enrichment of the literature. *AoB Plants 2010* (May 2010).

[11] Venkatesan, A., Larmande, P., Jonquet, C., Ruiz, M., and Valduriez, P. Facilitating efficient knowledge management and discovery in the Agronomic Sciences. In *4th Plenary Meeting of the Research Data Alliance* (Amsterdam, The Netherlands, September 2014).

[12] Whetzel, P. L., Noy, N. F., Shah, N. H., Alexander, P. R., Nyulas, C., Tudorache, T., and Musen, M. A. BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. *Nucleic Acids Research 39*, (web server) (June 2011), 541–545.