

Reusing the NCBO BioPortal technology for agronomy to build AgroPortal

Clément Jonquet,^{1,2,6} Anne Toulet,^{1,2,}
Elizabeth Arnaud,³ Sophie Aubin,⁴ Esther Dzalé
Yeumo,⁴ Vincent Emonet,¹ John Graybeal,⁶ Mark A.
Musen,⁶ Cyril Pommier,⁴ Pierre Larmande^{2,5}

¹ Laboratory of Informatics, Robotics and
Microelectronics of Montpellier (LIRMM)
University of Montpellier & CNRS, France

² Computational Biology Institute (IBC) of
Montpellier, France

³ Bioversity International, Montpellier, France
⁴ INRA Versailles, France

⁵ UMR DIADE, IRD Montpellier, France

⁶ Center for BioMedical Informatics Research
(BMIR), Stanford, USA
jonquet@lirmm.fr

Abstract— Many vocabularies and ontologies are produced to represent and annotate agronomic data. By reusing the NCBO BioPortal technology, we have already designed and implemented an advanced prototype ontology repository for the agronomy domain. 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, food, environment (perhaps animal) sciences. We offer an ontology portal which features ontology hosting, search, versioning, visualization, comment, recommendation, enables semantic annotation, as well as storing and exploiting ontology alignments. All of these within a fully semantic web compliant infrastructure. The AgroPortal specifically pays attention to respect the requirements of the agronomic community in terms of ontology formats (e.g., SKOS, trait dictionaries) or supported features. In this paper, we present our prototype as well as preliminary outputs of four driving agronomic use cases. With the experience acquired in the biomedical domain and building atop of an already existing technology, we think that AgroPortal offers a robust and stable reference repository that will become highly valuable for the agronomic domain.

Keywords—ontology repository, ontology mapping, semantic annotation, agronomic sciences.

I. INTRODUCTION

Similarly to what happens in biomedicine, communities engaged in agronomic research need to access specific sets of ontologies for data annotation and integration. For instance, it has been established that the scientific challenges in plant breeding have switched from genetics to phenotyping and that standard traits/phenotypes vocabularies are necessary to facilitate breeder's data integration and comparison. In parallel of very specific crop dictionaries [1], important organizations have produced large reference vocabularies such as AGROVOC (Food and Agriculture Organization), NAL Thesaurus (National Agricultural Library) or the CAB Thesaurus (Centre for Agricultural Bioscience International) and are currently working on integrating them [2]. The more ontologies are being produced in the domain, the more the need to create, store and retrieve alignments between those

ontologies become important. In fact, there exists a need of a one-stop-shop for agronomical, environmental and plant sciences ontologies enabling to identify and select an ontology for a specific task as well as offering generic services to exploit them in search, annotation or other scientific data management processes. Therefore, our goal is to enable straightforward use of agronomic related ontologies, avoiding data managers and researchers the burden to deal with complex knowledge engineering issues.

In the biomedical domain, the NCBO BioPortal (<http://bioportal.bioontology.org>) [3] is a well-known open repository for biomedical ontologies originally spread out over the web and in different formats. The NCBO BioPortal functionalities have been progressively extended in the last 10 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 an overlap exists, the portal does not span to the agronomic, environment or animal domains. An important aspect is that NCBO technology is domain-independent and open source. A BioPortal virtual appliance¹ is available as a server machine embedding the complete code and deployment environment, allowing anyone to set up a local ontology repository and customize it.

In this paper we present an ontology repository advanced prototype to support these challenges in agronomy and plant sciences. The portal is built atop of the NCBO BioPortal technology. The main objective of the AgroPortal project is to develop and support a reference ontology repository for the agronomic domain.

II. RELATED WORK

In the biomedical or agronomic domains there exists several “knowledge organization systems” listings such as BioSharing (biosharing.org) or the VEST Registry (aims.fao.org/vest-registry). They usually register ontologies and provide a few metadata about them. However, because they are registries for different kind of resources, they do not

¹ www.bioontology.org/wiki/index.php/Category:NCBO_Virtual_Appliance

support the level of features that an ontology repository offers. More specifically to plant domain, the Crop Ontology web application (www.cropontology.org) [4] publishes online sets of ontologies & dictionaries 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. 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. The Planteome portal (www.planteome.org) [5], is reusing the Gene Ontology project AmiGO technology to build a database of searchable and browsable annotations for plant traits, phenotypes, diseases, genomes, gene expression data across a wide range of plant species. Although the portal hosts the reference ontologies in the plant biology (e.g., PO, TO, EO), the portal focus is on data (not ontologies) and the scope is not as large as the one we envision for AgroPortal.

III. A PORTAL FOR AGRONOMIC RELATED ONTOLOGIES

We have clearly identified that the NCBO technology was the one that implements the most of the features (ontology & mapping repository, annotator, recommender, community support, etc.) the community would certainly be interested in, while being aware of the technical challenges of developing such a various and complex software. In addition, our vision is to adopt, as the NCBO did, an open and generic approach where users can themselves easily participate to the platform, upload and comment content (ontologies, mappings, projects). Plus, there are two major motivations for AgroPortal to reusing the outcomes of biomedicine: (i) to avoid re-developing technologies that have already been designed and extensively used; (ii) to offer the same tools, services and formats to both community 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.

We have developed and deployed an advanced prototype platform (v1.0 beta released in January 2016) <http://agroportal.lirmm.fr> – that currently hosts 49 ontologies – including 28 not originally present in BioPortal² – and we are working on 37 candidate ontologies. The platform counts already 38 registered users. The features offered by the portal are for example: (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 NCBO 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 (<http://data.agroportal.lirmm.fr>) and (ii) a

SPARQL endpoint (<http://sparql.agroportal.lirmm.fr>). While assuring the day to day maintenance and monitoring of the portal and keeping it up-to-date with the NCBO technology, we have started to work on customizations and specific services for the agronomic/plant community. For instances: organizing the content of the portal, working on multilingual support, interconnecting BioPortal and AgroPortal, scoring annotations, supporting different formats, adding new metadata.

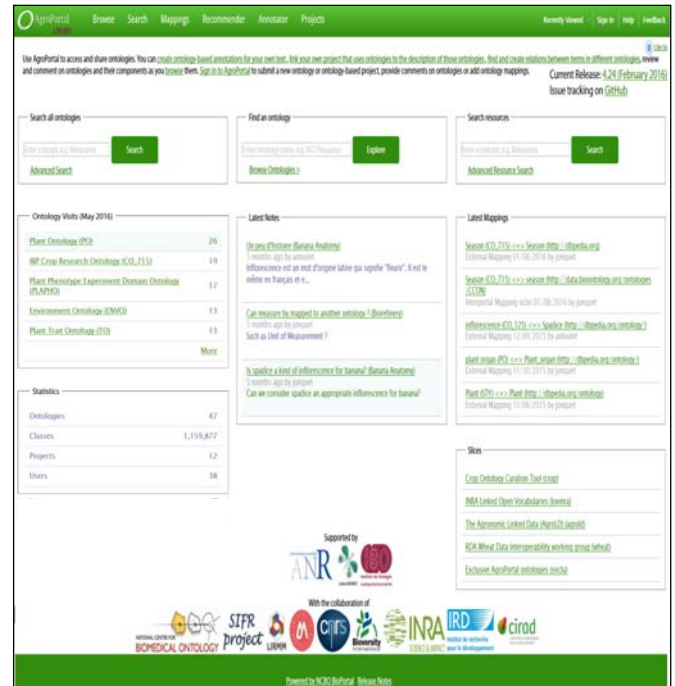


Fig. 1. Screenshots from the AgroPortal user interface

IV. DRIVING AGRONOMIC USE CASES

A. Agronomic Linked Data (AgroLD) within IBC

The Computational Biology Institute of Montpellier (IBC – <http://www.ibc-montpellier.fr>), develops methods for data integration and knowledge management within agronomic sciences to improve information accessibility and interoperability. The project is interested in identifying genes controlling roots and panicle branching as well as genes orthologous relationship for rice genes families. Using 8 ontologies for annotation, the project has built the AgroLD RDF knowledge base (<http://agrold.org>) that integrates data from a variety of plant resources (e.g., Gramene, SouthGreen, UniProtKB, OryGeneDB) and provides a portal for bioinformaticians to exploit the homogenized data models to efficiently build research hypotheses [6].

B. RDA Wheat Data Interoperability (WDI) working group

The WDI 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 needs identified by the group is to offer a repository of

² As of now, for technical reasons, we had to duplicate a few ontologies but the long term vision is an interconnected network of biportals that will enable anyone to access easily an ontology independently from where it's actually hosted.

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 WDI is maintaining a list of vocabularies and ontologies within an AgroPortal specific slice (<http://wheat.agroportal.lirmm.fr>) which has been reported in the WDI's set of guidelines for wheat data description (<http://datastandards.wheatis.org>). More recently, two other RDA working groups (Rice Data Interoperability and AgriSemantics) have expressed interest in using AgroPortal as a backbone for data integration and/or standardization.

C. 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 developed within specific focus projects 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. A specific group of ontologies has been setup in AgroPortal for ontologies produced or used by INRA and we are helping ontology editors to follow the semantic web standards when making their ontologies sharable and available.

D. The Crop Ontology project

The *Crop Ontology project* (www.cropontology.org) of the Consultative Group on International Agricultural Research (CGIAR) is AgroPortal's 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 field-books. We work on leveraging the backend of the cropontology.org web application with the AgroPortal web service API, while keeping the current web application as the primary point of access. We actually offer new functionalities to the Crop Ontology community such as versioning, SPARQL endpoint, notes, the annotation tool, while not breaking the uses of the current application. In addition, we 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 reference ontologies commonly used in plant biology.

V. CONCLUSION

In this paper we have briefly introduced AgroPortal, an open ontology repository for the agronomy domain. We have discussed four use cases that are already using the portal to support their work on data interoperability. The thematic boundaries of the portal are not precisely defined yet, (e.g., agriculture also includes animals) and it will be to the users to express what they expect to find into such a repository.

Although multiple questions need to be addressed, we do believe that the NCBO technology is a good candidate for this project and we see here an opportunity to capitalize technology and scientific outcomes of the ten last years.

Considering the position of the current NCBO BioPortal and the importance of having 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) illustrates the impact and interests first in France, but also internationally (e.g., Planteome, Elixir, BioSharing, EBI, FAO). Making available such a portal allows 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 contributes to long term support of that technology while reinforce the connections with the biomedical domain.

In the future we will identify more potential users for the portal and support new research scenarios. For each ontology available in the portal, we will go through an extensive description of its metadata in order for the portal to facilitate the comprehension of the landscape of agronomical ontologies.

ACKNOWLEDGMENT

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 (www.ibc-montpellier.fr), grant ANR-11-BINF-0002 as well as by University of Montpellier and the CNRS. We also thank the National Center for Biomedical Ontologies for help and time spent with us in deploying the AgroPortal.

REFERENCES

- [1] R. Shrestha, E. Arnaud, R. Mauleon, M. Senger, G. F. Davenport, D. Hancock, N. Morrison, R. Bruskiewich, and G. McLaren, "Multifunctional crop trait ontology for breeders' data: field book, annotation, data discovery and semantic enrichment of the literature," *AoB Plants*, vol. 2010, May 2010.
- [2] T. Baker and O. Suominen, "Global Agricultural Concept Scheme (GACS): A multilingual thesaurus hub for Linked Data." Report, 2014.
- [3] N. F. Noy, N. H. Shah, P. L. Whetzel, B. Dai, M. Dorf, N. B. Griffith, C. Jonquet, D. L. Rubin, M.-A. Storey, C. G. Chute, and M. A. Musen, "BioPortal: ontologies and integrated data resources at the click of a mouse," *Nucleic Acids Research*, vol. 37, pp. 170–173, May 2009.
- [4] L. Matteis, P. Chibon, H. Espinosa, M. Skofic, R. Finkers, R. Bruskiewich, and E. Arnaud, "Crop ontology: vocabulary for crop-related concepts," in *1st International Workshop on Semantics for Biodiversity* (P. Larmande, E. Arnaud, I. Mougenot, C. Jonquet, T. Libourel, and M. Ruiz, eds.), vol. CEUR Workshop Proceedings, (Montpellier, France), pp. 37–46, May 2013.
- [5] P. Jaiswal, L. Cooper, J. L. Elser, A. Meier, M.-A. Laporte, C. Mungall, B. Smith, E. K. Johnson, M. Seymour, J. Preece, X. Xu, R. S. Kitchen, B. Qu, E. Zhang, E. Arnaud, S. Carbon, S. Todorovic, and D. W. Stevenson, "Planteome: A resource for Common Reference Ontologies and Applications for Plant Biology," in *24th Plant and Animal Genome Conference, PAG'16*, (San Diego, USA), January 2016.
- [6] A. Venkatesan, P. Larmande, C. Jonquet, M. Ruiz, and P. Valduriez, "Facilitating efficient knowledge management and discovery in the Agronomic Sciences," in *4th Plenary Meeting of the Research Data Alliance*, (Amsterdam, The Netherlands), September 2014.