Data-driven Agricultural Research for Development

A Need for Data Harmonization Via Semantics

Medha DEVARE[1], Céline AUBERT[2], Marie-Angélique LAPORTE[2], Léo VALETTE[2], Elizabeth ARNAUD[2], Pier Luigi BUTTIGIEG[3]

[1]CGIAR Consortium Office ; [2]Bioversity International
Montpellier, France

[3]Alfred-Wegener-Institute, Helmholtz-Zentrum für Polar- und Meeresforschung
Bremerhaven, Germany

***Abstract*— Addressing global challenges to agricultural productivity and profitability increasingly requires access to data from a variety of disciplines, and the ability to easily combine and analyze related data sets. Innovation in agricultural research for development must therefore be mediated by reliable and consistently annotated information resources across disciplinary domains. Leveraging semantics ensures this consistency and ease of reuse, and the global CGIAR Consortium that includes 15 agricultural research for development Centers is attempting to harness this promise through efforts such as its Open Access, Open Data Initiative. CGIAR's Crop Ontology project plays a key role in this, and will soon be enhanced by an Agronomy Ontology (AgrO). AgrO is being built to represent traits identified by agronomists and the simulation model variables of the International Consortium for Agricultural Systems Applications (ICASA). Further, it will coordinate its semantics with existing ontologies such as the Environment Ontology (ENVO), Unit Ontology (UO), and Phenotype And Trait Ontology (PATO). Once stable, it is anticipated to address one of the domains temporarily represented in the Sustainable Development Goals Interface Ontology (SDGIO), pertaining to multiple SDGs such as the elimination of hunger and poverty. AgrO will complement existing crop, livestock, and fish ontologies to enable harmonized approaches to data collection, facilitating data sharing and reuse. Further, AgrO will power an Agronomy Management System and fieldbook, similar to the Crop Ontology-based Integrated Breeding Platform (IBP) and fieldbook. There is substantial interest from agronomists and modelers in such a fieldbook to standardize agronomic data collection, and the ontology itself as a means of facilitating hitherto missing linkages with breeding and other data, and enabling wider sharing and reuse of agronomic research data.**

Keywords—ontology; agronomy; fieldbook; standardization; semantics

I. INTRODUCTION

CGIAR is a consortium of 15 Centers around the globe, focused on agricultural research for development. These 15 Centers and other organizations nvolved in agricultural research and development are charged with tackling challenges at a variety of scales from the local to the global; however, all too often, research outputs are difficult to discover or are not sufficiently described to be truly accessible and usable. CGIAR Centers have made strong progress implementing publication and data repositories that meet minimum interoperability standards; however, barriers to information discoverability, access and integration still persist.

CGIAR information products, like those throughout the agricultural domain, are often characterized by their distribution across diverse databases, variation (particularly true for field data), differing scales (from the micro, through the landscape, to the global), disciplinary and geographic diversity, and the fact that much of the data is not born digital (except in the agricultural genetics/genomics domains). All of this contributes to substantial complexity in research outputs, making it difficult to relate and cross-link them to provide relevant, accurate, and complete information. This situation is further exacerbated by scientists typically not being habituated to thinking of data as a shared resource, and therefore not annotating it well or in standardized ways. Standard metadata and semantics (vocabularies and ontologies) can help the agricultural community address these challenges: an essential step towards integrating related information and fully realizing the value of agricultural information resources.

II. SEMANTICS IN AGRICULTURAL DATA MANAGEMENT

We envision the establishment of a semantics-based agricultural cyberinfrastructure and are building the Agronomy Ontology to facilitate its growth. A well-developed semantic layer will enhance the coordination of the diverse data resources handled by CGIAR and its partners. Indeed, understanding the consequences of varying factors within any cropping system involves the synthesis of disparate data types, including management practices, crop phenotypes, and socioeconomic data. However, integration of pre-breeding, breeding, socioeconomic, agronomy and related data does not typically happen, largely because these data are often collected, described, and stored in inconsistent ways, impeding data comparison, mining and interpretation for meta-analysis, as well as data reuse in models and decision-support tools. Comprehensive standardization at the level of data and information is, generally speaking, unlikely in this varied domain; thus, ensuring broad use of minimal, standardized metadata and variables unambiguously represented in a reference ontology will provide a more stable foundation for future synthesis.

To support semantic coherence, we are developing AgrO as a reference agronomy ontology which will represent key variables from three sources: the International Consortium for Agricultural Systems Applications (ICASA)[1], the Crop Research Ontology[2], and traits developed by Medha Devare at the International Maize and Wheat Improvement Center (CIMMYT). The variables selected out of these three lists are gathered into an Excel template that resolves a measured variable into three components: the real-world parameter, the method used to generate information, and scale or units of the information artifact produced by the method (Table 1).

TABLE 1: EXAMPLE OF A VARIABLE OF THE AGRONOMY ONTOLOGY

Variable	Parameter	Method of measurement	Scale
SoiBul_measure_ g/cm3	Soil bulk density is the weight of soil in a given volume	Collect a known volume of soil using a metal ring, and determine the weight after drying. Bulk density = Dry soil weight /Soil volume	g/cm^3

The template derives from the Trait Dictionary template designed by the Crop Ontology project (Shrestha et al., 2012). A literature review supports the identification of the methods of measurement and the scales. A total of 435 variables have been selected for AgrO, based on their relevance to agronomic trials generally conducted to develop and assess new technologies and practices that can help farmers sustainably improve system productivity and profitability. Variables are validated with existing datasets to ensure they represent the scope of a typical trial. The first dataset is a multi-locational agronomic wheat trial conducted in 1988 by the US Department of Agriculture – Agricultural Research Service (USDA-ARS)[3]. Besides the identification of standard variables and their components, relevant external ontologies such as ENVO (Buttigieg et al., 2013), UO and PATO (Mabee et al., 2007) will be integrated to complete AgrO.

A stable AgrO is envisioned to be of utility to the Sustainable Development Goals Interface Ontology (SDGIO), representing agricultural and agronomic entities of direct relevance to SDGs 1-3, 12, 15. From this perspective, AgrO seeks to capture the insights of agronomists and field researchers and connect them with the global development agenda through an interoperable semantic layer. This will be invaluable in promoting accurate representation of agricultural, socioeconomic, and ecological realities and harmonizing the data that describe them. A key use case is the creation of an agronomy fieldbook underpinned by AgrO to standardize data collection and annotation by diverse field agents. AgrO will also be employed by the envisioned agricultural infrastructure to make agricultural data discoverable, accessible, interoperable, and reusable.

III. OUTLOOK AND CONCLUSIONS

Development of the Agronomy Fieldbook requires a good understanding of the process of agronomy trial design, captured in a semantically coherent fashion. Regular consultations with a Community of Practice composed of agronomists and data managers from CGIAR, CIRAD, INRA and other institutions support the design of a prototype fieldbook based on AgrO, along with mock-ups of the user interface to help these potential users clearly visualize the nature and utility of the fieldbook, and provide feedback without being overwhelmed by the ontology itself.

The first stage of the development of AgrO is well advanced, with agronomists and data managers expressing strong support and interest in AgrO, the Agronomy Fieldbook, and the proposed agricultural cyberinfrastructure. It is anticipated that this infrastructure and the attendant building blocks (such as AgrO and the Crop Ontology) involved in its construction will greatly enhance knowledge sharing, innovation and impact in the agriculture domain writ large, while enriching the external reference ontologies it draws from.

ACKNOWLEDGMENTS

The authors gratefully acknowledge Dr. Jeffrey White, USDA-ARS, and Dr. Cheryl Porter, University of Florida, who provide regular expert advice to the development of AgrO.

This work is supported by a Bill and Melinda Gates Foundation grant.

P.L. Buttigieg is supported by the ERC Advanced Grant “Abyss” (no. 294757) to Antje Boetius.

[1] http://research.agmip.org/display/dev/ICASA+Master+Variable+List

[2] http://www.cropontology.org/ontology/CO_715/Crop%20Research

[3] Data set provided by Dr. Jeffrey White, US Department of Agriculture

REFERENCES

[1] Rosemary S., Matteis L., Skofic M., Portugal A., McLaren G., Hyman G., Arnaud E.: 2012. Bridging the phenotypic and genetic data useful for integrated breeding through a data annotation using the Crop Ontology developed by the crop communities of practice. Frontiers in Physiology, vol. 3.

[2] Database Systems. Liu L., Tamer Özsu M. (Eds.), Springer-Verlag

[3] Buttigieg, P. L., Morrison, N., Smith, B., Mungall, C. J., & Lewis, S. E. (2013). The environment ontology: contextualising biological and biomedical entities. *J Biomed Semant*, *4*(1), 43. http://doi.org/10.1186/2041-1480-4-43

[4] Mabee, P. M., Ashburner, M., Cronk, Q., Gkoutos, G. V, Haendel, M., Segerdell, E., Mungall, C., & Westerfield, M. (2007). Phenotype ontologies: the bridge between genomics and evolution. Trends in Ecology & Evolution, 22(7), 345–50. http://dx.doi.org/10.1016/j.tree.2007.03.013

[5] The Sustainable Development Goals Interface Ontology (SDGIO) Code Repository. https://github.com/SDG-InterfaceOntology/sdgio