Scientific Symposium FAIR Data Sciences for Green Life Sciences
12 December 2018 | Wageningen | the Netherlands

Conference Proceedings | DOI: https://doi.org/10.18174/FAIRdata2018.16284

# CocoaSoils data interoperability vision

Turdukulov, U.[1], R. Knapen[1], S. Janssen[1], H. Boogaard[1], L. Woittiez[2], K. Giller[2]

*[1] Wageningen Environmental Research, Wageningen University & Research, Droevendaalsesteeg 3, 6708 PB Wageningen, the Netherlands*
*[2] Plant Production Systems Group, Wageningen University & Research, Droevendaalsesteeg 1, 6708 PB Wageningen, the Netherlands*
*Corresponding author's e-mail: ulan.turdukulov@wur.nl*

Data-generative approaches are becoming increasingly common in modern life science research. Agronomy, food, plant sciences, and biodiversity are examples of complementary scientific disciplines that can greatly benefit from the integration and re-sue of the data that they produce. For instance, at WENR Earth Informatics group we focus on integrating datasets for calibration and validating methods in the crop modelling and monitoring domain. These datasets are scattered in many repositories in various formats often with missing metadata and column headings. Our recent pilot to integrate CGIAR data repositories revealed that even within a single institution these data sets are not harmonised to allow easy aggregation: varying templates for each experiment, missing metadata fields or in separate codebook / research papers, unclear column headings are just few examples of integration barriers. This illustrates how the current agricultural scientific community is fragmented in its data management and lacks commonly available reference data on (benchmarking of) agricultural production.

The sharing and publication of good quality data for wider use requires the integration of several working steps in a single data curation workflow. It cannot solely rely on one solution, e.g. a data-upload button or discoverability platform. Several efforts need to be carried out in parallel: quality data collection and archiving, data discovery and FAIRnisation, tailored visualisation and storage for specific domains and problems (i.e. crop monitoring).

Recently we have joined with CocoaSoils project (cocoasoils.org) to facilitate data collection and sharing of Integrated Soil Fertility Management (ISFM) of cocoa. The CocoaSoils project focuses on understanding the nutrient demand of cocoa trees through well-controlled randomised fertiliser experiments, and on improving soil fertility management and productivity in cocoa farms using on-farm experiments and extension. The first phase of the project will last for five years and runs across several countries in West Africa, Latin America and Southeast Asia, with plans to continue beyond the first phase and to include other countries. Currently the trials are being designed and implemented, and a baseline study among more than 2000 African farmers is being set up.

First step in this process is quality data collection. In CocoaSoils we are currently implementing this stage using open-source software (Open Data Kit, PostgreSQL, R and Python) on Red Hat OpenShift Container Platform. ODK app will allow enumerators and field workers to collect data more efficiently with fewer errors using scanners, and researchers to evaluate data shortly after data collection. Structural integrity and standardised naming of parameters makes the system generally applicable to various crops and experimental setups. We also intend to standardise, harmonise, and RDFise agronomic data at the collection stage by joining CGIAR's efforts to develop Agronomy Field Information Management System (AgroFIMS). AgroFIMS reuses existing reference ontologies like Environmental Ontology, mapping to reference plant ontologies like Planteome or creating concepts when missing. The Crop and Agronomy Ontologies are used for creating fieldbooks, store data in breeding databases, describe variables for the analytical platforms, provide accurate keywords for the Metadata of Dataverse repositories. The format of the observation variable has been adopted by other standards like MIAPPE (http://www.miappe.org/) or BrAPI (https://brapi.docs.apiary.io/).

Once data is collected and ontology enriched, CocoaSoils plans to share the field experiment data across and beyond its partners. CocoaSoils also intends to consolidate many of the external cocoa related efforts across the globe. This will pose additional challenges if the elements of the external data sets cannot be mapped into common ontologies. However with the cocoa specific ontology and agronomy ontology still being developed, it might be right time to link to various initiatives like, the Global Open Data in Agriculture Network (GODAN, http://www.godan.info), The Crop Ontology project (http://www.cropontology.org/) and CGIAR'S Big Data (https://bigdata.cgiar.org/). Or even to form Cocoa Data Interoperability group (similar to Wheat Data Interoperability group, https://www. rd-alliance.org/groups/wheat-data-interoperability-wg.html, to work on a common metadata schema and ontologies and form the basis of allied efforts to make CocoaSoils datasets interoperable in order to enable value addition and data-driven innovation in cocoa research.