Data Provenance in Agriculture

Sérgio Manuel Serra da Cruz[1] [0000-0002-0792-8157], Marcos Bacis Ceddia[1] [0000-0002-8611-314X], Renan C. T. Miranda[1] [0000-0002-2668-4266], Gabriel Rizzo[1] [0000-0002-8988-9955]1, Filipe Klinger[1] [0000-0002-4751-6587], Renato Cerceau[1,2][0000-0003-3953-4715] Ricardo Mesquita[4] [0000-0002-0267-6886] , Ricardo Cerceau[1] [0000-0003-3016-229X] Elton C. Marinho[5] [0000-0003-0117-0610], Eber A. Schmitz [5] [0000-0002-4839-4606] Elaine Sigette[3] [0002-1139-1356], Pedro Vieira Cruz[1] [0000-0001-6476-3865]

[1] Federal Rural University of Rio de Janeiro, Seropédica, RJ, Brazil
[2] National Agency of Supplementary Health, Rio de Janeiro, RJ, Brazil
[3] Federal Fluminense University, Volta Redonda, RJ, Brazil
[4] SENAI-RJ, Rio de Janeiro, RJ, Brazil
[5] Federal University of Rio de Janeiro, Cidade Universitária, RJ, Brazil

serra@ufrrj.br, ceddia@ufrrj.br

Abstract. Soils are probably the most critical natural resource in Agriculture, and soils security represents a critical growing global issue. Soils experiments require vast amounts of high-quality data, are very hard to be reproduced, and there are few studies about data provenance of such tests. We present OpenSoils; it shares knowledge about data-centric soils experiments. OpenSoils is a provenance-oriented and lightweight e-infrastructure that collects, stores, describes, curates and, harmonizes various soil datasets.

Keywords: Reproducibility, Soil Security, Open Data, Data Quality, Big Data.

1 Introduction

According to Food and Agriculture Organization (FAO)[1], an agency of the United Nations, the world's population is expected to grow to about 9,6 billion by 2050. Thus, there is widespread concern about the challenges to soil and food systems in meeting the demand of populations for sufficient, affordable, and nutritious food. There are similar concerns about meeting those challenges in ways that agriculture would benefit hugely from common shared global agronomic data spaces.

The modern Agriculture is a data-centric interdisciplinary domain, with the integration of different subjects (from genomics to soil sciences), different scales (from genes to geolocalisation) and, different markets (from local farmers to multinational research teams). The ability to manage and explore these datasets is a crucial issue to tackle the current sustainability challenges. A wide variety of datasets underpin prod-

[1] http://www.fao.org/about/what-we-do/en/

ucts and processes, which vary in size, complexity, structure, semantics, subject matter and in how they are updated and used.

Soils are probably the most critical natural resource in Agriculture; they generate environmental, health and socio-economic benefits that are vital to sustaining life on Earth [1]. Soil experiments are indispensable sources of knowledge. Researchers conduct several kinds of soils experiments which are characterized as long-term field experiments (LTE) and short-term (*in vitro* and *in silico*) lab experiments (STE). The LTE have been running for years in many parts of the world for the last 175-years-old (*e.g.* Rothamsted) and need more time to execute the research procedures. On the other hand, STE experiments can be performed in a few weeks or months and have the potential to contribute to the improve LTE. Thus, it is essential to deliver to the agronomic community a novel computing infrastructure that can share raw and curated data and the provenance of STE and LTE and augment the reproducibility of soil experiments. This paper presents a multi-layer e-infrastructure which bring innovations to Soils Science using FAIR principles (Findable, Accessible, Interoperable, and Reusable) [2], W3C PROV-DM[2], open data and semantic web standards.

2 Experiments in Soils Science

Soil Science represents the area that studies the soil (and its properties) as a natural resource, including soil formation, composition, classification, mapping, management and use [1,3], these properties could be about physical, chemical, biological, and fertility. Soils experiments are costly because the soils are incredibly diverse, and it is necessary to treat them in a specific manner [3]. Any recommendation fits specific soil and weather conditions. Besides, the soil properties have high spatial and time variability. Finally, changes in soil properties can often be proved and quantified only after decades.

The LTE is essential in monitoring and understanding the changes in soil physics or fertility occurring because of long-term agrotechnical operations. Their scientific and practical value is immeasurable and keeps improving over the years. The information about the soils use cannot be replaced by any other means [3]. Additionally, the STE produced much of the data that built the sciences of soil physics, chemistry, and biology [1,3]. STE often explore soil processes subject to change over decades, topics such as aggregation, weathering, microbial activity, and soil fertility itself. Although STE enriches soil models, most tend to be reductionist, isolating individual components, and do not study the whole soil, with its high-order interactions that become apparent only with time.

3 Open Soils

Data and provenance are the primary and permanent assets in OpenSoils (www.opensoils.org). The architecture is an open, provenance-oriented, and light-

[2] https://www.w3.org/TR/prov-dm/

weight computational e-infrastructure which rely on layers to store, compute and share curated data of (STE and LTE) soils experiments [5]. Fig. 1 illustrates a conceptual view and the flow of information in the architecture.

Layer 1 (*End-users layer)* - hosts on the OpenSoils Web portal; it collects soil data directly from the LTE into OpenSoils database. The specialists can use mobile and web applications (*e.g.*, OpenSoils App, API and Wet Lab tools) to collect the data directly in the fields (LTE experiments) and trace the route of each soil sample sent to chemistry and physics laboratories to be analyzed. Usually, the morphological properties of the soil are analyzed *in situ* by the specialists. OpenSoils app sends raw data to the cloud-based database thought the API. After that, each soil sample is tagged and sent to laboratories where the scientist does wet experiments and execute STE which evaluate specific physic-chemical properties of each soil horizon and selected soil samples are shipped to the UFRRJ´s soils museum.

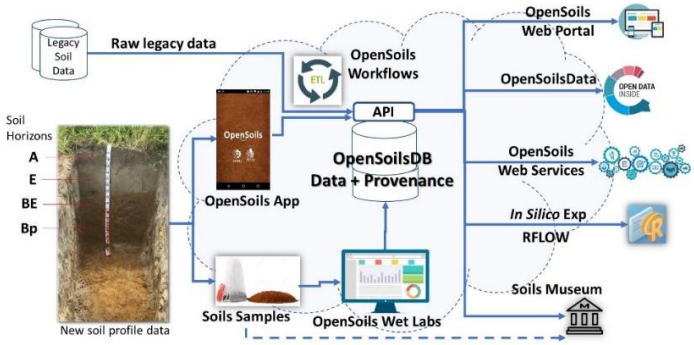

Fig. 1. Overview of the conceptual data-flow in OpenSoils.

Layer 2 (*Services layer*) - hosts soil models and data-centric scientific workflows which ingest large amounts of legacy data and analyses the consistency of the incoming data [3].

Layer 3 (*Data layer*) - stores and describes various soils datasets with metadata. The internal structure supports a diversified degree of data granularity and uses a database named OpenSoilsDB [5,6] which can store new curated soils data annotated with provenance metadata. Much of the information needed to assure the data quality and to allow researchers to reproduce STE experiments can be obtained by systematically capturing data provenance [4]. OpenSoilsDB can store provenance from ETL workflows and scripts. ETL Workflow provenance consists of the record of the derivation of a result (*e.g.*, a soil experiment, an image, a map) by a computational process represented as scientific workflows. Script provenance is obtained by running the source code of scripts (*e.g.* R, Pyhton). OpenSoilsDB used W3C PROV-DM recommendation to store provenance and was designed to support the FAIR principles for scientific data management and data stewardship [2]. The principles ensure transparency, reproducibility, and reusability of the experiments, facilitating data sharing more systematically.

The database also supports the ingestion of legacy soils data imported through ETL workflows. The layer can store scientific and governance data. Besides, to support

open data, we can use general-purpose data repositories (*e.g.*, CKAN, Dataverse, DSpace, Dryad, DataHub).

A specific thesaurus is used to add semantics and annotate soils data, allowing us to link it as RDF triples in WikiData. The thesaurus used in the e-infrastructure is Agrovoc [7], which is a SKOS-XL (Simple Knowledge Organization System eXtension for Labels) concept scheme published as LOD (Linked Open Data). It covers several areas of interest of the FAO including food, agriculture and, environment. This thesaurus is used by researchers, librarians, and information managers for indexing, retrieving, and organizing data in agricultural information systems.

Data management is not a target in itself, but a key conduit leading to knowledge discovery and innovation in soil sciences. OpenSoilsDB database stores scientific and governance data. The scientific data aims to serve high quality-assessed, georeferenced soils profiles database to the Brazilian and international communities upon their standardization and harmonization. Each soil profile description recorded in the database has more than 43 entities, and 250 attributes to stores the soil properties and soil experiments (mineralogical, morphological, chemical, physical, and environmental data). Furthermore, the database support data versioning and provenance; stores georeferenced soil data (text and images) about physic-chemical analytical data from each horizon and soil samples analyzed in wet laboratories.

Data governance is an essential block in the knowledge base of information professionals involved in supporting data-intensive research. Its adoption is advantageous because it is a service based on standardized, repeatable processes, designed to enable the data discovery and the transparency of data-related transformation processes.

Layer 4 (*Governance layer*) - hosts data licenses, re-use rights, analytical tools, visualization and map generation services that can be connected to other software (*e.g.*, ArcGIS, R or Jupyter) to generate analytical reports, prediction and raster maps. Although received little attention in soils research communities, this layer is foundational for soils security. The prime function of the layer is to improve and maintain the citations and quality of the soils dataset; thus, to be successful at governance, quality must be continuously measured, and the results continuously retrieved by the data and services layers.

4 Concluding remarks

Maintaining healthy soils is a key to modern agriculture. However, there is still much computational work needed to be developed in soil sciences and more in-depth studies to understand the role of data provenance in Agriculture. We introduced OpenSoils; it is an e-infrastructure which share knowledge about STE and LTE in soils security using FAIR, PROV, and semantic web approaches. The infrastructure is being developed and aims to enhance reproducibility of experiments and deliver high-quality datasets, knowledge and maps based on curated data.

Acknowledgments

This work was supported in part by the Brazilian agencies FNDE/MEC/SESU, PIBIC/CNPq, Petrobras and CYTED networks BigDSSAgro and SmartLogistcs@IB.

References

1. Koch, A. et al.: Soil Security: Solving the Global Soil Crisis. Global Policy 4(4) 434-441, (2013).
2. Wilkinson, M. D. et al.: The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, vol. 3 (2016).
3. Körschens, M. The importance of long-term field experiments for soil science and environmental research – a review. Plant Soil Environ., 52, 1–8 (2006).
4. Cruz, S.M.S, Nascimento, J.A.P.: SisGExp: Rethinking Long-Tail Agronomic Experiments. In: IPAW 2016, LNCS, vol. 9672, 214-217. Springer, Heidelberg (2016).
5. Cruz, S.M.S. et al.: Towards an e-infrastructure for Open Science in Soils Security. In: XII Proceedings on BRESCI, 8pp. SBC, Natal-RN (2018).
6. Rizzo, G. S. C., Ceddia, M. B., Cruz, S. M. S.: Banco de Dados Pedológico: Primeiros Estudos. In: 5th Proceedings on V RAIC, pp.1-2. UFRRJ, Seropédica (in portuguese) (2017).
7. Caracciolo, C. et al.: The AGROVOC Linked Dataset. Semantic Web, 4(3), 341-348 (2013).