Towards an e-infrastructure for Open Science in Soils Security

Sérgio Manuel Serra da Cruz[1,2,3], Marcos Bacis Ceddia[1], Eber Assis Schmitz[3], Gabriel S. Rizzo[2], Renan C. T. Miranda[2], Sabrina O. Cruz[2], Ana Clara Correa[2], Felipe Klinger[2], Elton Marinho[3], Pedro Vieira Cruz[2]

[1] Universidade Federal Rural do Rio de Janeiro – PPGMMC/UFRRJ
[2] Programa de Educação Tutorial - PET-SI/UFRRJ
[3] Universidade Federal do Rio de Janeiro – PPGI/UFRJ

serra@ufrrj.br, ceddia@ufrrj.br

Abstract. *Soils Security is a critical and growing global concern. The OpenSoils´ objective is to host, connect and share large amounts of curated soil data and knowledge at the Brazilian and South America level. The e-infrastructure consists of several layers of services, a database of soil profiles, a cloud-based computational framework to compute and share soil data integrated with a map visualization tools. OpenSoils is open, elastic, provenance-oriented and lightweight computational e-infrastructure that collects, stores, describes, curates, harmonizes and directs to various soil resource types: large datasets of soils profiles, services/applications, documents, projects and external links. OpenSoils is the first open science-based computational framework of soils security in the literature.*

1. Introduction

Agriculture consists of a complex science from a data-centric point of view, with different disciplines (from genomics to soil sciences) and, different scales (from genes to geolocalisation). The ability to explore this complex dataset is a crucial issue to tackle new agricultural and societal challenges like food and soils security (WOLFERT *et al.*, 2017). To Koch *et al.* (2013), soils are probably the most important natural resource and biosystem that support the human and terrestrial life. It is a primary, finite natural resource which derives other resources, goods, and services.

Soils security is an emerging chief concept of soil sciences motivated by sustainable development and precision agriculture. It is related to the maintenance and improvement of the global soil resource to produce food, fibers and fresh water, human health, carbon sequestration, contribute to energy and climate sustainability, and to maintain the biodiversity and the overall protection of the ecosystem (KOCH et al., 2013). Soils security, like food security, has several dimensions (*e.g.*, capability, condition, capital, connectivity, and codification) that interact with environmental, social, and economic components (MCBRATNEY, FIELD & KOCH, 2014). Soils security is a data-intensive research domain which life-cycle starts at the harvest of new soils data in the field and finish at scientist's visualization workstation or decision maker´s desk (Figure 1). It is important to highlight that Figure 1 did not capture the complexity of soils security, once it does not encompass the interconnection of the five dimensions and the political, economic and sociological aspects of soil use and management. Figure 1 summarizes the life cycle of soil information at the research and academic level, which is the primary focus of this research.

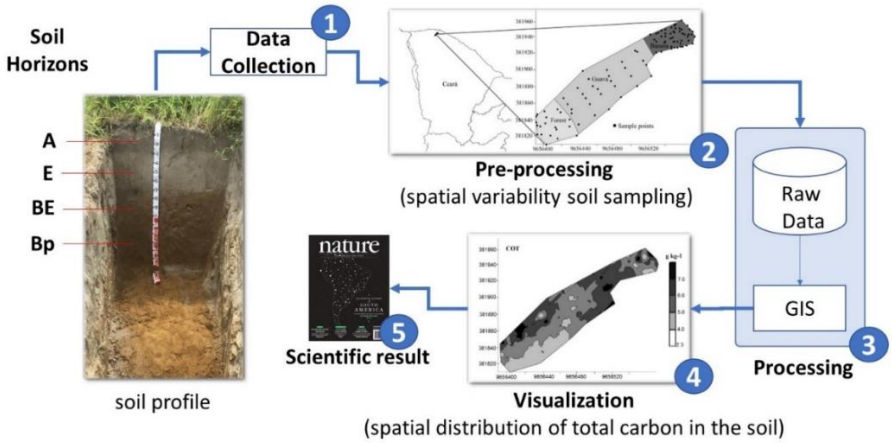

Figure 1 – Example of soil horizons and the main phases of the life-cycle of soils investigations (maps adapted from MELO *et al*., 2016).

Soils and food security investigations are in a rapid transformation. However, these disciplines did not draw the same degree of attention of other e-science subjects like bioinformatics, astronomy, computational chemistry. We advocate the utter necessity to do interdisciplinary research considering the roles of computer science, data governance, supply chain data integration and mathematical modeling in soils security to face the challenges. We foresee that several open data, semantic web, open science, big data, and data science approaches may aid the soils community to make wider investigations, do more accurate predictions in precision agriculture and deliver more knowledge to the society.

The goal of this paper is to present the big picture of OpenSoils. It was conceived to guide Brazilian policies by designing and laying the groundwork for a long-term effort aiming at achieving an e-infrastructure for open science in soils security that would position Brazil as a major global player at the forefront of research and innovation in this area. This paper is organized as follows. Section 2 presents the background. Section 3 presents OpenSoils conceptual architecture and uses. Section 4 the related work and Section 5 concluding remarks and future work.

2. Soil, Soils Data, and Open Science

The development of soil from inorganic and organic materials is a complex natural process. The soil is defined as the layer(s) of generally loose mineral and organic material that is affected by physical, chemical, and/or biological processes at or near the planetary surface and usually hold liquids, gases, and biota and support plants (VAN ES, 2017). The soil is considered an open system that interacts with other components of the geologic cycle. The characteristics of a soil are a function of Parent material, Climate, Relief, Organisms and Time. (PANSU & GAUTHEYROU, 2006). Soils are evaluated in the field through soil profiles, which is defined as a two-dimensional section composed of a vertical succession of horizons, commonly named O, A, B, C (beginning at the surface), that have been subjected to soil-forming processes (Figure 1). Each soil profile has very specific mineralogical, morphological, chemical, physical, biological and environmental properties. Soil investigations require actions in the field and wet scientific laboratories because soils properties are diverse and are hard to be collected, mapped, analyzed, stored and shared as soils data in databases.

Soils investigations, like any other scientific domain, has a life cycle and characteristics that deserves efforts to improve the long-term data management and use of

strategic the data assets (YAMSON *et al.,* 2016, ARROUAYS *et al*., 2017). Soil data has key features, for instance, there are lots of legacies unanalyzed raw data. However, either new or existing soils data are heterogeneous in its values and semi-structured in its formats.

Currently, there are many isolated data silos which store legacy soils data as (*e.g.,* scientific papers, spreadsheets, text, pdf files or web pages), having poor semantics and lacking metadata descriptors. Additionally, several soil databases are either inaccessible to structured queries or are presented as simple spreadsheets or text files, being hardly shared and reused by farmers and policymakers (ARROUAYS *et al.,* 2017). Lots of soil data and knowledge are still currently fragmented and at risk of getting lost in digital data silos or even in simple tables in scientific papers. Consequently, reproducing the results from scratch from several soils experiments is both time-consuming and error-prone at best, and sometimes impossible.

Recent evidence from meta-research studies suggests that problems with research integrity and reproducibility in several scientific domains (BAKER, 2016; NEVES *et al.,* 2017; FANELLI, 2018 & HUTSON, 2018; FREIRE & CHIRIGATI, 2018). Many scientists, journals, and funders are concerned about the biased, low reproducible and irreproducible scientific findings in soils security as well. Thus, one approach that may serve to expand the reliability and robustness of soils security investigations is the adoption of open science (MUNAFÒ, 2016), e-science (HEY *et al*., 2009) and data provenance (BUNEMAN *et al.,* 2000 & FREIRE *et al.,* 2008).

Open science is an umbrella term encompassing a multitude of assumptions about the future of knowledge construction (FECHER & FRIESIKE, 2013). It is a global movement to make scientific research, data, and dissemination accessible at all levels of an inquiring society. Nowadays, there are some open science infrastructures (*e.g.,* OpenAIRE, OSF, EOSC, among others) not experienced with features of soils security challenges. E-infrastructure is a computational tool that promotes open, centralized workflows by enabling capture of different aspects and products of the research life-cycle, including developing a research idea, designing an investigation, storing and analyzing collected data, and writing and publishing reports or papers. The e-infrastructures support a variety of scientific tools and services to assist in the research process (FOSTER & DEARDORFF, 2017).

3. OpenSoils e-infrastructure

It is useful to start from a theoretical e-infrastructure framing the complexity of challenges and demystifying the role of big data in soils security. OpenSoils is an open, elastic, provenance-oriented and lightweight computational open science e-infrastructure which rely on four overarching layers. Figure 2 illustrates the e-infrastructure, the layers and summarizes the data life-cycle of soil data (showed as arrows) (DEELMAN *et al.,* 2009; CRUZ, CAMPOS & MATTOSO, 2009; MATTOSO *et al.,* 2010).

(i) The end-users layer (*e.g.,* soil specialists, data managers, policy makers) uses on the web portal and mobile applications. They are used to collect and ingest new soil data directly from the fields into OpenSoilsDB using OpenSoils app or query data through the web portal aiding policy-makers to make decisions (DSS), and urban planners do envision new soils usage (PSS).

The specialists and researchers use this layer to handle data. The first can use mobile, IoT and web applications (*e.g.,* OpenSoils App and Wet Lab tools) to collect the data directly in the fields and trace the route of each soil sample collected and sent to the chemistry and

physics laboratories (*i.e.,* wet labs) to be further analyzed. Usually, each soil sample is submitted *in situ* by the specialists to morphological analyses. Thus, OpenSoils app sends raw data to the database. After that, each soil sample is tagged and shipped to laboratories where the scientist does (*in vitro*) wet experiments and further execute (*in silico)* computational scientific experiments with SisGExp (CRUZ & NASCIMENTO, 2016) which evaluate specific physic-chemical properties of each soil horizon.

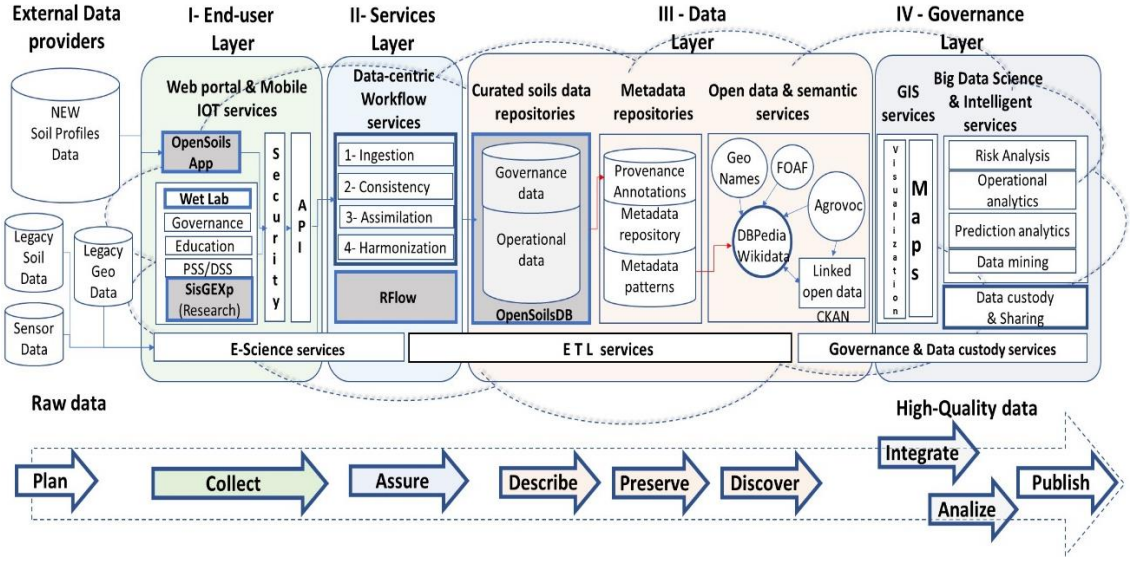

Figure 2 – Overview of the conceptual architecture of OpenSoils (the arrows describe data operations within the phases of life-cycle of soils investigations).

(ii) The services layer uses scientific and business models to generate curated data; they are composed of set data-centric scientific workflows (which ingest and analyses the consistency of the incoming of legacy soils data). RFlow is part of the layers (NASCIMENTO, 2015). It is a provenance-based approach that aid researchers to reproduce scientific experiments based on R scripts. RFlow manages, shares, and enacts the computational scientific workflows that encapsulate legacy R scripts it transparently captures provenance of R scripts and endows experiments reproducibility.

(iii) The data layer stores in the core of OpenSoils, it stores, describes, curates, various soils data sets, and metadata descriptors. The internal structure supports a diversified degree of data granularity and uses a relational database named OpenSoilsDB (former InfoSoilsBR, (RIZZO, CEDDIA & CRUZ, 2017). It can store new curated soils data annotated with provenance.

Much of the information needed to assure the data quality and to allow researchers to reproduce soils security experiments can be obtained by systematically capturing its provenance. Provenance refers to the record trail that accounts for the origin of a piece of data (FREIRE *et al.,* 2008). OpenSoilsDB can store workflow and scripts provenance. Workflow provenance consists of the record of the derivation of a result (*e.g.,* a soil profile, an image, a map) by a computational process represented as scientific workflows. Script provenance is obtained by analyzing the source code of soils security experiments represented as R scripts (PIMENTEL *et al.,* 2017). OpenSoilsDB uses W3C PROV-DM recommendation to store prospective and retrospective provenance for workflows and scripts (MOREAU & MISSIER, 2013). Besides, OpenSoilsDB supports FAIR guidelines (Findable, Accessible, Interoperable, and Reusable) for scientific data management and sharing (WILKINSON *et al.,* 2016).

The database also supports the ingestion of legacy soils data imported through ETL tools (*e.g.,* Pentaho/Kettle). The layer can store operational and governance data. Besides, to support open data we use CKAN (http://ckan.org/) which stores curated open data sets. Besides, CKAN is an international open data standard provides a streamlined way to make curated soils data publishable, usable, discoverable and interoperable by third-part soils applications. CKAN support data annotation with thesaurus ensuring semantic interoperability between computer systems, research teams or community users to exchange data with unambiguous meaning.

The thesaurus is used to semantically annotate soils data, allowing us to link it as RDF triples in DBpedia (2018), as depicted in Figure 2. The thesaurus used in the e-infrastructure is Agrovoc (CARACCIOLO *et al.,* 2013). Currently, Agrovoc is a SKOS-XL concept scheme published as Linked Open Data which covers all areas of interest of the Food and Agriculture Organization (FAO), including food, agriculture, environment. FAO publishes it; it is edited by a community of experts and consists of over 34,000 concepts available in 29 languages. It is used by researchers, librarians and information managers for indexing, retrieving and organizing data in agricultural information systems.

OpenSoilsDB database has two abstraction layers (*e.g.,* operational and governance). The lower operational layer 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 40 entities, and 250 attributes to stores the soil properties and soil experiments (*e.g.,* mineralogical, morphological, chemical, physical and environmental data). Furthermore, the database support data versioning, data provenance, and stores georeferenced soil data as text and images about physic-chemical analytical data from each horizon and soil samples analyzed in wet laboratories.

The upper layer of the OpenSoilsDB improves the accessibility and reuse of soil data and knowledge. Data governance and data literacy are two important building blocks in the knowledge base of information professionals involved in supporting data-intensive research, and both address data quality and research data management. Adopting data governance in OpenSoils is advantageous because it is a service based on standardized, repeatable processes and is designed to enable the transparency of data-related processes and cost reduction. It refers to rules, policies, standards; decision rights; accountabilities and methods of enforcement.

(iv) The governance layers are composed by data management, data license, analytical and visualization tools and map generation services that can be connected to other software (*e.g.,* QGIS, ArcGIS, R, Tableau or sci-kit-learn) to generate analytical reports, soils prediction, raster maps to name a few.

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 quality of the soils dataset; thus, to be successful at governance, quality must be continuously measured, and the results continuously fed back by the data and services layers. We stress that this layer has roles of individuals. For instance, these individuals are the application owners, data custodians and application data architect, they are responsible for compliance with data standards, resolve data-related issues, share the soil datasets, and support enforcement of data/soil standards.

3.1 Daily uses of OpenSoils

OpenSoils was conceived as an e-infrastructure because refers to a combination and interworking of digitally-based software technologies, resources (data, services, digital repositories), communications (protocols and data access rights), and the people and organizational structures needed to support modern and collaborative research in soils security. OpenSoils has three primary uses:

(i) Offer diverse, integrated, timely and trustworthy digital repositories to researchers (*e.g.,* statistical studies of the quality of soils, soils mapping, evaluation of contamination by heavy metals and organic waste management system).

(ii) Offer tools to city planners, agronomists, farmers to make better decisions using high-quality harmonized open data (*e.g.,* studies to erosion, risk of landslides, risk of flooding, potential for agricultural use of soils; environmental and economic and ecological zoning, insurance of agronomic entreprises, land classification for irrigation; support in the recommendation of fertilizers and limestone).

(iii) Help students to increase their knowledge and skills about soils, the e-infrastructure is connected to the Brazilian Soils Museum at UFRRJ, where users can explore the collection of soil monoliths, soil artifacts, pictures and browse the data.

4. Related Work

Traditionally, soils security has operated along disciplinary lines in using and applying its data and analytical tools. Soils data management, curation, and governance is an issue that is still underestimated in soils sciences, with data being analyzed for isolated applications and with small groups of researchers working with isolated data silos on their personal computers and not properly sharing them (LOKERS et al., 2016). Today, there are no open science software platforms to support the full cycle of research in soils security. Thus, we conceived OpenSoils as an open e-infrastructure that than be used by the researcher, decision maker, data curator, city planner, farmer and students.

The investigations of soils security in Brazil and Latin America are still beginning. They are depicted as several isolated investigations and data silos about legacy soils data. For instance, BDSolos (BDSOLOS, 2018) is a relational database developed by EMBRAPA Solos that stores about 9.000 soils profiles. The database has no provenance nor metadata descriptors, besides there are no public interfaces to allow researchers to insert new soils data. Furthermore, the interfaces to query data are hard to be used even by soil specialists. Last but not least, there are no concerns about soils security nor map visualization facilities. We can point out the same limitations are shared by Fe.BR (FEBR, 2018). It is a single HTML website that stores the same type of data of BDSolos. The dataset is presented as a set of google docs resting in a virtual drive on the Web; their authors claim that it is open data. However, we stress that it fails to fulfill the eight Open Data principles (OGP, 2018), has no governance policies and unfortunately does not commit with the best web semantic practices (GYRARD *et al.,* 2015).

Fortunately, OpenSoils is entirely different from related works; it was conceived to adopt the open science, e-science, open data and data provenance emerging trends. First, it is a multi-disciplinary, community and data integrative e-infrastructure. Second, it supports the movement to make scientific experiments more reproducible and the publications and scientific data available as open access. Third, it can handle large amounts of data of soils security investigations. Fourth, it based on web, workflows services, and clouds infrastructure

which offer access to elastic and abundant resources that can be provisioned and de-provisioned on-demand.

5. Concluding Remarks

Conditions are now ripe for a comparable step change in the interplay between soils science and computer science, a change that will not only spur economic growth and competitive advantage, but also will help scientists to develop solutions to our societal challenges, understand climate change, and explore new frontiers of knowledge.

The soil has an integral part to play in the global environmental sustainability challenges. Nevertheless, there is still a lot of computational work needed to be fully developed in soil sciences. The growth of open science and the curated open soils databases may aid scientist to increase the reliability, robustness, and reproducibility of soils security experiments.

In this paper we presented OpenSoils, a novel e-infrastructure which provide knowledge about soils security to different kinds of users and not only researchers. The infrastructure enhances reproducibility and delivers high-quality soils datasets, knowledge and maps based on curated open data. OpenSoils is being developed; the mobile apps can be found at PET-SI Google Play and the further information about Wet Labs applications, the scientific workflows or ETL components can be found at http://www.opensoils.org.

As future work, we plan to finish the implementation of the e-infrastructure and investigate the alternative semantic relationships between soils data, digital objects and related domains to enhance solutions and improve data sharing, data curation, and long-term data stewardship policies.

Acknowledgments

This work was supported in part by the Brazilian funding agencies FNDE, PIBIC/CNPq and Petrobras. The author's thanks, PET-SI/UFRRJ, MEC/SESU, Reds CYTED – BigDSSAgro and SmartLogistcs@IB.

References

Arrouays, D. et al., Soil legacy data rescue via GlobalSoilMap and other international and national initiatives. GeoResJ 14, pages 1-19, 2017.

Baker, M., 1,500 scientists lift the lid on reproducibility. Nature. 533:7604, 2016.

Buneman, P., Khanna, S., Tan, W-C. Data Provenance: Some Basic Issues. In: Kapoor S., Prasad S. (eds) FST TCS 2000: Foundations of Software Technology and Theoretical Computer Science. FSTTCS 2000. Lecture Notes in Computer Science, vol 1974. Springer, Berlin, Heidelberg.

BDSolos, Banco de Dados de Solos. https://www.bdsolos.cnptia.embrapa.br/consulta_publica.html, (acessado em 9.3.2018).

Caracciolo, C. et al. The AGROVOC Linked Dataset. Semantic Web, 4, 3, pages. 341-348. 2013.

Cruz, S.M.S, Campos, M. L. M and Mattoso, M. Towards a Taxonomy of Provenance in Scientific Workflow Management Systems. In: SERVICES I, pages. 259-266, USA, 2009.

Cruz, S.M.S, Nascimento, J.A.P. SisGExp: Rethinking Long-Tail Agronomic Experiments. IPAW 2016.

DBPedia, http://wiki.dbpedia.org/ (acessado em 24.3.2018).

Deelman, E. et al., Workflows and e-Science: An overview of workflow system features and capabilities. Future Generation Computer Systems, 25:5, pages. 528–540, 2009.

Fanelli, D. Opinion: Is science really facing a reproducibility crisis, and do we need it to? Proceedings of the National Academy of Sciences of the USA, March 2018.

FeBR, Repositório de dados de solos. http://coral.ufsm.br/febr/, (acessado em 9.3.2018).

Freire, J., Koop, D., Santos, E., Silva C.T. Provenance for Computational Tasks: A Survey. Computing in Science and Engineering, 10:3, pages 11–21, 2008.

Freire, J. Chirigati, F. Provenance and the Different Flavors of Computational Reproducibility. IEEE Data Engineering Bulletin, 41(1), pages. 15-26, 2018.

Fecher, B. and Friesike, S. Open Science: One Term, Five Schools of Thought. Opening Science, pages. 17-47, 2013.

Foster, E. D., Deardorff, A. Open Science Framework (OSF). J Med Libr Assoc. 105:2, pages. 203–206. 2017.

Gyrard, A., Serrano, M., Atemezing G. A. Semantic Web Methodologies, Best Practices and Ontology Engineering Applied to Internet of Things. 2nd IEEE World Internet of Things, 2015.

Hey, T., Tansley, S., Tolle, K. The Fourth Paradigm: Data-Intensive Scientific Discovery. Microsoft Research, 2009.

Hutson, M., Artificial intelligence faces reproducibility crisis. Science, 359: 6377, pp. 725-726, 2018.

Koch, A. et al., Soil Security: Solving the Global Soil Crisis. Global Policy, 4:4 ages 434-441. 2013.

Lockers, R. et al., Analysis of Big Data technologies for use in agro-environmental science. Environmental Modelling & Software, 84, pp. 494-504, 2016.

Mattoso, M. et al., Towards supporting the life cycle of large-scale scientific experiments. International Journal of Business Process Integration and Management, 5:1, pages 79-92, 2010.

McBratney, A., Field, D. J and Koch, A. The dimensions of soil security. Geoderma. 213, pages 203-213, 2014.

Melo, A. A. B. et al., Spatial distribution of organic carbon and humic substances in irrigated soils under different management systems in a semi- Arid zone in Ceará, Brazil. SEMINA: CIENCIAS AGRARIAS, 37:4, pages 1845-1856, 2016.

Moreau. L., Missier, P. PROV-DM: The PROV Data Model. https://www.w3.org/TR/prov-dm/ (acessado em 24.3.2018).

Munafò, M. Open Science and Research Reproducibility. Ecancer medical science. 10, ed56. 2016.

Nascimento, J. A. P. RFLOW: uma arquitetura para execução e coleta de proveniência de workflows estatísticos. Dissertação de Mestrado, UFRRJ, 2015.

Neves V. C. et al., Managing Provenance of Implicit Data Flows in Scientific Experiments. ACM ACM Transactions on Internet Technology. Volume 17 Issue 4, Article No. 36, 2017.

OGP, Open Government Partnership, 2017. https://www.opengovpartnership.org/countries/brazil (acessado em 9.3.2018).

Pansu, M., Gautheyrou, J., Handbook of Soil Analysis, Springer, 2006.

Pimentel, J. F et al., noWorkflow: a Tool for Collecting, Analyzing, and Managing Provenance from Python Scripts 2017. Proceedings of the VLDB. vol 10:12, pages 1841-1844, 2017.

Rizzo, G.S.C, Ceddia, M. B., Cruz, S. M. S. Banco de Dados Pedológico: Primeiros Estudos. V RAIC – UFRRJ, 2017.

Wilkinson, M. D. et al., The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3, Article number: 160018, 2016.

Worlfert, S. et al., Big Data in Smart Farming – A review. Agricultural Systems, v. 153, pages 69-80, 2017.

Yamson, D. O., et al., Putting Soils Security on the Policy Agenda: Need for a Familiar Framework. Challenges. 4:2 15 pages. 2016.

van Es, H., A New Definition of Soil CSANews 62:20-21, 2017.