Provenance, Blockchain, and Smart Contracts as a Traceable Soil Mapping Solution

Élton Carneiro Marinho1, Eber Assis Schmitz[1], Sérgio Manuel Serra da Cruz[1,2]
[1] Graduate Program in Informatics – Federal University of Rio de Janeiro (PPGI/ UFRJ) – Rio de Janeiro – RJ – Brasil
[2] Computer Science Department – Federal Rural University of Rio de Janeiro (DECOMP/UFRRJ) – Seropédica– RJ – Brasil

{elton.marinho, serra}@ppgi.ufrj.br, eber@nce.ufrj.br

Level: **Doctorate**
Ful Name: Élton Carneiro Marinho
Advisors: Sérgio Manuel Serra da Cruz and Eber Assis Schmitz

Entry into 01/09/2017
Defense Forecast: 03/2023
Qualification exam: 08/06/2021

Publications

Prêmio - ***Selo de Inovação da SBC*** - Premiação de Melhores Trabalhos de Inovação na CSBC 2022 - Projeto OpenSoils, Sociedade Brasileira de Computação.

Marinho, É. C., Pontes, A., Da Silva, M. F., Da Cruz, S. M. S., & Schmitz, E. A. (2021). ***Uso de blockchain na gestão sustentável de dados do solo***. Anais Da VII Escola Regional de Sistemas de Informação Do Rio de Janeiro. www.opensoil.org

Marinho, É. C., Pinheiro, A. G., Bessa, A. C. F., Cruz, S. M. S. da, & Schmitz, E. A. (2020). ***Proteção de Dados: Proposta de gerenciamento de dados de solos usando os princípios FAIR e a tecnologia blockchain***. Proceedings of TICAL.

Marinho, É. C., Pinheiro, A. G., Bessa, A. C. F., da Cruz, S. M. S., & Schmitz, E. A. (2020). ***Agricultura 4.0: uma abordagem a partir dos princípios FAIR e tecnologia blockchain para gestão de dados de levantamentos pedológicos***. CAI.

Marinho, É. C., Cruz, S. M. S. da, & Schmitz, E. A. (2020). ***An approach for sustainable and efficient soil data management***. Anais Do BigDssAgro.

Cruz, S. M. S. da, Klinger, F., Duarte, A., Mendes, J., Cruz, P. V., **Marinho, E. C.,** & Schmitz, E. A. (2019). ***OpenSoils: Uma Plataforma de Apoio à Ciência do Solo***. SBIAgro, November. https://www.researchgate.net/publication/337305016_OpenSoils_Uma_Plataforma_de_Apoio_a_Ciencia_do_Solo

Cruz, S., Klinger, F., Cruz, P., Vieira, A., Schmitz, E., & **Marinho, E.** (2019). ***Desenvolvendo Sistemas Agrícolas de Próxima Geração: Um Estudo em Ciência de Solos***. Anais Do Workshop de Computação Aplicada à Gestão Do Meio Ambiente e Recursos Naturais (WCAMA), 135–144. https://doi.org/10.5753/wcama.2019.6428

Cruz, S. M. S, Ceddia, M. B., Tàvora Miranda, R. C., Rizzo, G., Klinger, F., Cerceau, R., Mesquita, R., Cerceau, R., **Marinho, E. C**., Schmitz, E. A., Sigette, E., & Cruz, P. V. (2018). ***Data Provenance in Agriculture***. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): Vol. 11017 LNCS* (Issue Sept, pp. 257–261). https://doi.org/10.1007/978-3-319-98379-0_31

Cruz, S. M. S, Ceddia, M. B., Costa, R. J. M., Schmitz, E. A., **Marinho, E. C**., Mendes, J., Correia, A. C., Cruz, S. O., Cruz, P. V., Klinger, F., Miranda, R. C. T., & Rizzo, G. S. C. (2018). ***Putting Soil Security on the Sustainable Development Agenda: Need for an open e-infrastructure***. *Proceedings of the Second International Conference on Agro Big Data and Decision Support Systems in Agriculture*, 59–62.

Cruz, S. M. S, Ceddia, M. B., Schmitz, E. A., Rizzo, G. S., Miranda, R. C. T., Cruz, S. O., Correa, A. C., Klinger, F., **Marinho, E. C**., & Cruz, P. V. (2018). ***Towards an e-infrastructure for Open Science in Soils Security***. *Anais Do Brazilian E-Science Workshop (BreSci), 2*(Figure 1). https://doi.org/10.5753/bresci.2018.3273

Cruz, S. M. S. da Ceddia, M. B., Cruz, P. V., Rizzo, G. S., Miranda, R. C. T., Cruz, S. O., Correa, A. C., Klinger, F., **Marinho, É. C.,** & Schmitz, E. A. (2018). ***OpenSoils: e-Science em Segurança de Solos***. Tical, September.

Resumo. A tecnologia Blockchain combinada com a proveniência de dados torna os dados do solo mais confiáveis e rastreáveis, fornecendo informações invioláveis sobre suas origens e transformações. Apresentamos FAIRCHAIN, uma infraestrutura computacional para gerenciar contratos inteligentes que utilizam dados curados de solos capaz de contribuir para mitigar os desafios em aberto na cadeia de abastecimento agroalimentar, mais especificamente na dificuldade de rastreabilidade de dados de solos. Neste trabalho, discutimos o mecanismo de estruturação de contratos inteligentes usando dados enriquecidos com metadados de proveniência retrospectiva. A infraestrutura é flexível e pode incorporar implementações de fluxo de trabalho.

Abstract. Blockchain technology combined with Data provenance is one way to make soil data more trustworthy and traceable by providing tamper-proof information about the origin, transformations, and history of pieces of data. We present Hyperledger Fabric of FAIRCHAIN, a computational infrastructure that manages smart contracts that uses soil data. We aim to mitigate the open challenges of the agricultural food supply chain, specifically in the difficulty of traceability of soil data. In this work, we present the mechanism to structure a smart contract using soil data enriched with retrospective provenance metadata. The infrastructure can hold workflow implementations.

Introduction

The problems facing soil security in Brazil today have enlarged considerably, as well as various factors and phenomena that augment the complexity of the issues, such as climate change, land grabbing, misuse of fertilizers, erosion, loss of agrobiodiversity, and other agricultural sustainability issues. However, from the computer science point of view, the scarcity of curated and trustful soil data is burdensome. For instance, Brazil has a small amount of soil data. Only 5% of the national territory is mapped on a scale of 1:100,000 or more (Bolfe et al., 2018).

Agronomic investigations require high-quality data. Today, there are thousands of soil data silos widespread on the Web. They have multiple semantics and formats and little or no provenance metadata. These silos are isolated, undocumented, and stored in laboratories or institutions with limited findability, accessibility, and security guarantees (Cruz et al., 2019; Marinho, Cruz et al., 2020).

Data provenance, also known as lineage or pedigree, is metadata that can be understood as the digital history of a piece of data or dataset (Freire et al., 2008). It is used in several areas, from computer science to digital agriculture (Allemang & Bobbin, 2016). For example, evidence of provenance is essential to ensure food quality and consumer protection. They ensure data originality, identify data generation processes, and increase confidence in the food supply chain.

Due to their flexibility, blockchains and data provenance are used for various applications, such as digital agriculture, food supply chains, and soil security (Marinho, Pinheiro, et al., 2020). Blockchain is quite a generic technology and has the potential to serve an array of industries due to its practical advantages. Every update on the Blockchain is incremental, making it impossible to modify or recreate any piece of data on the chain of data blocks.

This work presents the Hyperledger Fabric (HL) of FAIRCHAIN, a computational infrastructure that manages soil data, ensuring provenance, trustworthiness, and traceability. This proposal is a contribution that can be used in the agricultural food supply chain. FAIRCHAIN consists of structuring smart contracts using soil data enriched with provenance metadata in the chaincode of the Hyperledger Fabric (HL). The structured provenance enriches soil data with rastreability, while blockchain technology ensures the data's integrity, auditability, and confidence.

Related works

Azzi *et al.* (2019) described how the Blockchain could be integrated into the supply chain to create reliable, transparent, authentic, and secure systems. The authors analyzed the benefits and challenges encountered in blockchain-based supply chain management.

Yadav et al. (2020) reviewed the agricultural food supply chain considering food safety and soil security. Even traditional retailers like Walmart stress that Blockchain and provenance are critical for running their business (Kamath, 2018).Lautert *et al.* (2021) pointed out that the transparency of blockchain is highlighted by data provenance. It also described the adherence of blockchain technology with data provenance and data immutability.

Cruz *et al.* (2018) present OpenSoils, a provenance-oriented multilayered e-infrastructure tailored to store curated soil profiles in a cloud-based computational framework. It is a lightweight computational approach that allows researchers to collect, store, describe, curate, and harmonize various soil resources, increasing the compliance and trustability of food supply chains. The e-infrastructure can be extended to support the Blockchain and store data in a tamper-proof way.

Our proposal aims to be more embracing than related works. We propose a decentralized platform with member governance and decentralized data storage respecting data ownership. In addition, new types of soil-related data can be integrated into the platform, and we have the possibility of scientific workflow implementation.

Data Provenance

Data provenance identifies the derivation history of data. It is well-known in e-Science, Databases, Data Science, and Digital Libraries (Tan, 2007). However, in digital agriculture, few works consider the critical importance of data provenance to achieve reproducibility and trust in agronomic results (Adams et al., 2017).

Our proposal considers the use of retrospective provenance in next-generation digital agronomic systems. Provenance is metadata derived from the computational steps (da Cruz & do Nascimento, 2016). It represents information such as "which process was executed," "who performed it," and "how long it took" without having prior knowledge of the sequence of computational steps involved in it. (Freire et al., 2008).

Metadata can be collected and recorded as retrospective provenance in FAIRCHAIN, such as who is in charge of collecting the soil samples in the field, when it started and ended, what is the geolocation of samples; what devices and software were used. This metadata must also be saved in case of further change. In the case of images, another set of metadata must be collected, such as timestamp, executing agent, geolocation, equipment, image resolution, and others.

Data provenance can be more tangible in the food supply chain due to Blockchain's immutable and decentralized nature; it can enhance, and the benefits are diverse. For instance, data and provenance about soil samples can be used to create fertility maps of a given property to get carbon credits in the market. Insurance companies can rely on the maps because Blockchain ensures that once a specific dataset is committed to the ledger, it can never be deleted. Thus, allowing them to audit or follow the variations of carbon rates in the property through the years, they will never be deceived because they can always bring up the records with trust, transparency, and confidence.

Blockchain and Hyperledger Fabric

Blockchain technology emerged to deal with the intermediation problem by certifying entities (e.g., banks and insurance agencies), allowing untrusted entities to interact in a decentralized and distributed manner without brokers or trusted third parties. Blockchain uses a distributed database that records immutable transactions on a network.

The type of access is the core characteristic of a blockchain. In a public blockchain, any user can enter without any prior identification. A private blockchain (a.k.a permissioned) requires that the user has permission from the company, institution, or group responsible for the Blockchain. The last kind comprises parts that know each other and share a previously established relationship (Androulaki et al., 2018).

As soon as a permissioned blockchain is created, each user profile would be allowed to write a set of data and metadata controlled by the consensus defined among the organizations. The permission is essential to respect the organizations' roles (groups, companies, and wet labs, among others).

Hyperledger Fabric (HL), or simply Fabric, is a permissioned blockchain architecture aiming for resilience, flexibility, scalability, and confidentiality. Fabric is designed as a modular and extensible general-purpose blockchain. It supports running distributed applications written in standard programming languages, allowing them to run consistently across many nodes, giving the sense of running on a single global distributed blockchain network.

An HL network supports multiple connected blockchains. Each blockchain can be referred to as a channel and has different peers as members. The introduction of the architecture, which follows the execute-order-validate blockchain paradigm, consists of an endorsement and wise contract policy, executing untrusted code in a distributed way in an untrusted environment (Androulaki et al., 2018)

Practical Use Case: HL and soil data repository in FAIRCHAIN

In this section, we describe the implementation of an HL Fabric network in FAIRCHAIN and its relationship to digital agriculture, considering the repository of OpenSoils e-infrastructure. Figure 1 summarizes the canonical model used to ensure traceability and immutability of soil data in the repository.

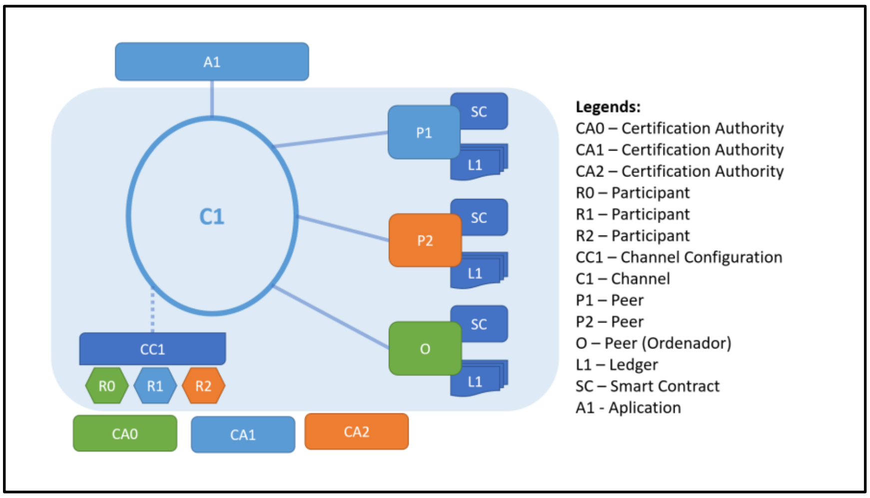

Figure 1 - Network model in HL Fabric. Adapted from: (Hyperledger, 2020)

One or more participants can start a channel (C1). They define how the operation of the future network channel will be configured. The participant's information is stored in the channel's configuration file. This setting defines restrictions to the network resources, i.e., who can read or write the channel or use a specific chaincode API via an access control list. The setup occurs before the channel's creation or is specified when instantiating the chaincode on the channel.

Organizations must agree on the channel configuration (CC1) in the "Configuration Block" block. In our scenario, there are three organizations (R0, R1, and R2), the organizations may be composed of technicians and pedologists, and each organization belongs to a different institution. As mentioned before, one of the organizations might unilaterally create the channel and invite the others. In this model, the organizations will collaborate on the channel since its initialization.

Users of organizations need to be authorized to act within this channel. Each organization must define which certification authority serves they need. Hyperledger Fabric provides a standard Certificate Authority component (CA), which issues PKI-based certificates to network member organizations and users.

In our model, a root certificate is issued for each organization. An enrollment certificate is issued for each user authorized by the respective certification authority (CA0, CA1, and CA2). A user (pedologist, student, researcher, among others) must be associated with an organization (institution, organization, or group) authorized to send requests to the Blockchain. In this use case, the requests would be the datasets of soils.

Endorsement Policy

These certificates can identify components belonging to an organization, sign transactions, and indicate if the organization endorses the transaction's outcome; this is a precondition for the transaction to be accepted in the ledger.

The endorsement policy is a static library for validating transactions, which a chaincode can parameterize. It allows the chaincode to specify the endorsers for a transaction as a set of nodes (Ruan et al., 2020).

The transaction flow has three steps: i) execution, verification of correctness, and endorsement of transactions; ii) order and consensus protocol; iii) validation against application-specific confidence assumptions, avoiding concurrency problems (Manevich et al., 2021; Ruan et al., 2020).

In the execution phase, the transaction proposals are signed by users and sent for endorsement. These proposals contain the sender's identity, the operation to be executed, the parameters, the chaincode identifier, a nonce[1], and a transaction identifier derived from the users' ID and the nonce. Every chaincode implicitly specifies a set of endorsers through the channel's endorsement policy (Androulaki et al., 2018).

The results of the execution process are input to the sorting service. The nodes perform an atomic transmission protocol resulting in an ordered sequence of output strings grouped into blocks. (Manevich et al., 2021). The transaction flow is anchored in consensus; each originator uses the same block formation protocol to group transactions into blocks and forward them to the nodes. Then, a block is formed as the number of pending transactions reaches the limit or when a timeout occurs (Ruan et al., 2020).

[1] In cryptography, a nonce is an arbitrary value that can only be used once.

Blocks can be delivered to nodes by the ordering service or gossip. A new block enters the validation phase. This phase consists of three sequential steps: the evaluation of the endorsement policy, the read-write conflict check-up (a.k.a verification) of all transactions in the block, and, finally, the update of the local ledger with blockchain state update. Block transactions are validated sequentially based on the corresponding endorsement policy and serialization.

These organizations can physically host the ledgers (L1) and chaincodes (SC) within the channel. These hosting points can be called nodes (P1 and P2). They are one of the physical points where organizations transact on a channel connected to it (the other being an application). The number of channels in a node is limited only to physical and country legislation.

After the creation of the channel, transactions are forwarded to endorsing nodes. Next, transactions are passed to the Order Service for ordering and structuring the block. The last step is the validation stage, where the block is validated, recorded in the ledger, and distributed to other nodes.

Smart Contracts and Chaincodes

The next step is introducing smart contracts, or chaincodes, to the channel nodes. Smart contracts are codes that connect a user to the blockchain network. Smart contracts with similar characteristics can be grouped in HL Fabric called chaincode (SC). A chaincode is installed on the nodes and may group more than one smart contract.

Chaincodes are installed on the nodes, approved by the organizations, and persisted in the channel. Install, approve, and confirm a chaincode is known as the "lifecycle" of the chaincode. Organizations are not required to install all chaincodes. Note, in Figure 2, that the ordering service does not have the chaincode installed, as these nodes usually do not propose transactions.

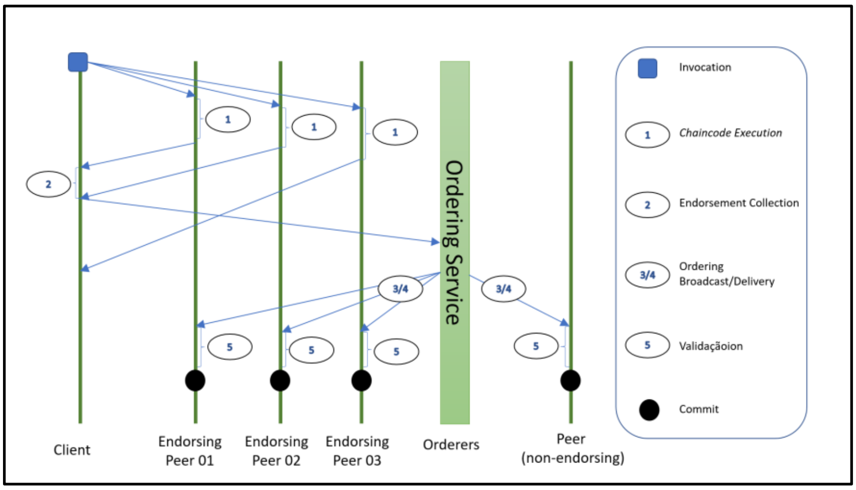

Figure 2 – Fabric's transaction flow, adapted from (Androulaki et al., 2018)**.**

In the practical use case presented (HL and soil data repository in FAIRCHAIN), a chaincode will be generated with only one smart contract to manage the data that will be persisted in the Blockchain. This contract will be responsible for validating the user's access to the Blockchain prior to the information persistence. This contract allows the management of soil profiles and horizons.

The smart contract will be responsible for validating the entry data and then persist on the blockchain. All results from blockchain-based queries will be made available in JSON[2] format, allowing interoperability with other machines that know this result format. With the smart contract confirmed by the peers, the client application (A1) can be used to invoke transactions in this chaincode (Hyperledger, 2019). The use case was implemented on a local machine to validate concepts and punctual adjustments in the smart contract with satisfactory results.

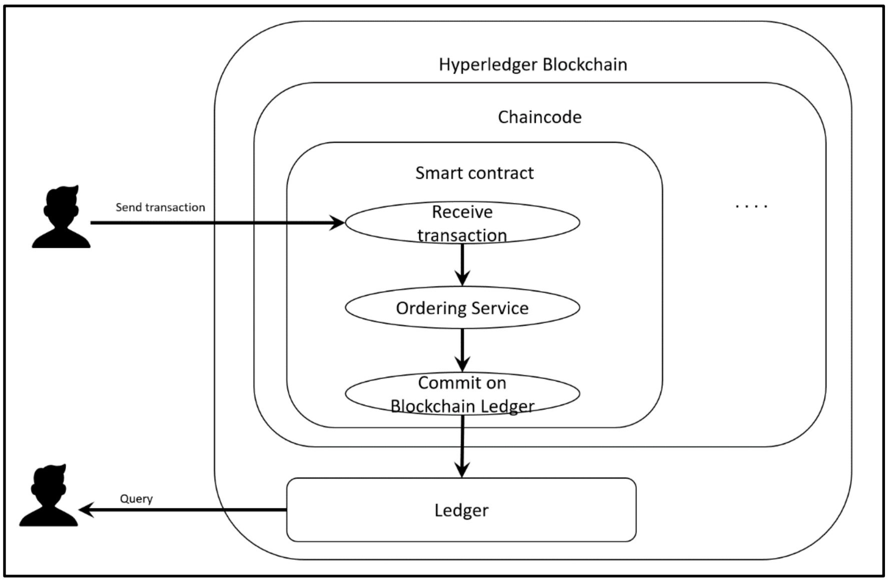

Figure 1 - Smart contract workflow (Source: the authors)

Conclusions

The FAIRCHAIN proposal increases trustworthiness and traceability in the food supply chain. For instance, if agronomic services achieve better knowledge about the soils, they can face the soil security issues more appropriately, getting a deeper understanding of complex issues like land grabbing or soil loss by erosion.

Today, there are few works in the literature about provenance and blockchain in soil data. The way data can be stored on the FAIRCHAIN blockchain offers tamper-proof information about the origin, transformations, and history of pieces of data. Another essential point to note is that soil datasets receive unique identifiers, which increases the possibility of reusing this data and increases their confidence in further studies.

We stress that high-quality soil data can be transformed into high-value agronomic products like geographic maps or soil properties maps, which governments or private companies, like insurance companies, banks, and agronomic services, can trust to make more accurate decisions. For instance, as the data on the blockchain cannot be deleted, variations of the rates of carbon in the soil samples can be followed through time.

The FAIRCHAIN blockchain is being developed as part of a broader platform called Opensoils (Cruz et al., 2018). The platform is constantly evolving to meet the demands of soil science professionals.

[2] JavaScript Online Notation

Acknowledgments

This work has been supported in part by the following Brazilian research agencies: CNPq, CAPES and MEC/PET. The last author in funded by grant # 306115/2021-2.

References

Adams, R., Parry, G., Godsiff, P., & Ward, P. (2017). The future of money and further applications of the blockchain. *Strategic Change*, *26*(5), 417–422. https://doi.org/10.1002/jsc.2141

Allemang, D., & Bobbin, T. (2016). *A Global Data Ecosystem for Agriculture and Food*. *24*(2), 23. https://www.godan.info/sites/default/files/documents/Godan_Global_Data_Ecosystem_Publication_lowres.pdf

Androulaki, E., Barger, A., Bortnikov, V., Cachin, C., Christidis, K., de Caro, A., Enyeart, D., Ferris, C., Laventman, G., Manevich, Y., Muralidharan, S., Murthy, C., Nguyen, B., Sethi, M., Singh, G., Smith, K., Sorniotti, A., Stathakopoulou, C., Vukolić, M., … Yellick, J. (2018). Hyperledger fabric. *Proceedings of the Thirteenth EuroSys Conference*, 1–15. https://doi.org/10.1145/3190508.3190538

Azzi, R., Chamoun, R. K., & Sokhn, M. (2019). The power of a blockchain-based supply chain. *Computers & Industrial Engineering*, *135*, 582–592. https://doi.org/10.1016/j.cie.2019.06.042

Bolfe, É. L., Campos, S. K., Júnior, M. A. G. P., Contini, E., Rodrigues, R. de A. R., & Santana, C. A. M. (2018). Futuro da Agricultura Brasileira. *Embrapa*, 212.

Cruz, S. M. S. da, Ceddia, M. B., Schmitz, E. A., Rizzo, G. S., Miranda, R. C. T., Cruz, S. O., Correa, A. C., Klinger, F., Marinho, E. C., & Cruz, P. V. (2018). Towards an e-infrastructure for Open Science in Soils Security. *Anais Do Brazilian E-Science Workshop (BreSci)*, *2*(Figure 1). https://doi.org/10.5753/bresci.2018.3273

Cruz, S. M. S. da, Klinger, F., Duarte, A., Mendes, J., Cruz, P. V., Marinho, E. C., & Schmitz, E. A. (2019). OpenSoils : Uma Plataforma de Apoio à Ciência do Solo. *SBIAgro*, *November*. https://www.researchgate.net/publication/337305016_OpenSoils_Uma_Plataforma_de_Apoio_a_Ciencia_do_Solo

da Cruz, S. M. S., & do Nascimento, J. A. P. (2016). SisGExp: Rethinking Long-Tail Agronomic Experiments. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 9672, Issue June, pp. 214–217). https://doi.org/10.1007/978-3-319-40593-3_24

Freire, J., Koop, D., Santos, E., & Silva, C. T. (2008). Provenance for computational tasks: A survey. *Computing in Science and Engineering*, *10*(3), 11–21. https://doi.org/10.1109/MCSE.2008.79

Hyperledger. (2019). *Hyperledger – Open Source Blockchain Technologies*. Hyperledger. https://www.hyperledger.org/

Hyperledger. (2020). *How Fabric networks are structured*. Hyperledger 2020. https://hyperledger-fabric.readthedocs.io/fr/latest/network/network.html

Kamath, R. (2018). Food Traceability on Blockchain: Walmart's Pork and Mango Pilots with IBM. *The Journal of the British Blockchain Association*, *1*(1), 1–12. https://doi.org/10.31585/jbba-1-1-(10)2018

Lautert, F., Pigatto, D. F. G., & Gomes-JR, L. C. (2021). Blockchain-based Data Provenance. *Anais Do Workshop Em Blockchain: Teoria, Tecnologias e Aplicações (WBlockchain 2020)*, 120–125. https://doi.org/10.5753/wblockchain.2020.12975

Manevich, Y., Barger, A., & Assa, G. (2021). Redacting transactions from execute-order-validate blockchains. *IEEE International Conference on Blockchain and Cryptocurrency, ICBC 2021*, 1–9. https://doi.org/10.1109/ICBC51069.2021.9461093

Marinho, E. C., Cruz, S. M. S. da, & Schmitz, E. A. (2020, October). An approach for sustainable and efficient soil data management. *Anais Do BigDssAgro*.

Marinho, E. C., Pinheiro, A. G., Bessa, A. C. F., Cruz, S. M. S. da, & Schmitz, E. A. (2020). Proteção de Dados : Proposta de gerenciamento de dados de solos usando os princípios FAIR e a tecnologia blockchain. *Proceeedings of TICAL*.

Ruan, P., Loghin, D., Ta, Q. T., Zhang, M., Chen, G., & Ooi, B. C. (2020). A Transactional Perspective on Execute-order-validate Blockchains. *Proceedings of the ACM SIGMOD International Conference on Management of Data*, 543–557. https://doi.org/10.1145/3318464.3389693

Tan, W.-C. (2007). Provenance in Databases: Past, Current, and Future. *IEEE Data Engineering Bulletin*, *30*(4), 3–12.

Yadav, S., & Singh, S. P. (2020). Blockchain critical success factors for sustainable supply chain. *Resources, Conservation and Recycling*, *152*, 104505. https://doi.org/10.1016/j.resconrec.2019.104505