A new model for intra- and inter-institutional soil data sharing

José Padarian and Alex B. McBratney

Sydney Institute of Agriculture & School of Life and Environmental Sciences, The University of Sydney, New South Wales, Australia

Correspondence: José Padarian (jose.padarian@sydney.edu.au)

Abstract. Data sharing and collaboration are critical to solving large scale problems. The prevalent soil data-sharing model is based on different groups sending their data to a lead party. This model is of a centralised nature and, consequently, results in the participants ceding their control and governance over their data to the lead party. Here we explore the use of a distributed ledger (blockchain) to solve the aforementioned issues. We explain what a blockchain is and some of its characteristics to then
5 describe some features of a blockchain that makes it an interesting candidate for an inter-institutional database. Finally, we describe the potential use case of developing a global soil spectral library with multiple, independent international institutions constituting the network.

1 Introduction

10 Soil is a key component of ecosystems and the need for soil information to monitor its condition is increasing. A large amount of soil data has been collected in the last century, with a special increment during the 70s-80s, and many organisations are performing the exhaustive task of "rescuing" and organising that data in more accessible formats (Arrouays et al., 2017). Additionally, in many countries, a large amount of new soil data is being generated partially aided by the advancements in methods such as soil spectroscopy, which makes the acquisition of soil data faster and cheaper compared with traditional wet
15 chemistry methods (Padarian et al., 2019b).

Most collected soil data is useful to solve problems locally but it is too fragmented to tackle more general issues. This applies at various levels of granularity including different teams within an institution, a single institution in different regional locations, and multiple institutions either within a country or internationally. In these cases, collaboration and data sharing becomes paramount. The soil community recognises this collaboration need and has responded by creating different data-
20 sharing initiatives. For instance, Rossel et al. (2016) compiled a global soil spectral library for soil mapping, modelling and monitoring with datasets from 92 countries (mainly data from United States, Australia and Europe). Another global spectral library has been promoted by FAO's Global Soil Partnership via the Global Soil Laboratory Network (GLOSOLAN). FAO also promoted different initiatives to establish collaboration networks to share soil profile information, including between Latin American countries (SISLAC) or a Global Soil Information System (GLOSIS). All these initiatives are designed as

centralised systems were, in order to collaborate, different parties (either individuals or organisations) should send their data to the lead organisation (Fig. 2a). A centralised information system has a series of disadvantages that we explore in this work. Two of the most important disadvantages of using a centralised network, especially in the context of a collaborative network of multiple independent parties, are that the control and data governance (norms, principles and rules) are completely ceded to the initiating party. Usually, these aspects could be defined in a data-sharing agreement prior to the establishment of the collaboration but, in practice, there are no controls to avoid unilateral decisions.

A potential solution for the data control and governance issues derived from the implementation of a centralised data-sharing system in the use of distributed a ledger or blockchain. The aim of this paper is to delineate the requirements for a functional, decentralised, inter-institutional database (IIDB) to share soil information on a distributed ledger or blockchain. First, we introduce some terms that will be used throughout this paper to then explain what a blockchain is and some of its characteristics. Second, we describe some features of a blockchain that makes it an interesting candidate for an IIDB. Finally, we present a use case of collaborative effort that could be a good fit for using the proposed model.

2 Definitions

Public-private key-pair: In asymmetric cryptography, two keys are used — private and public keys. The private key is used to sign a transaction and the public key is used to verify the signature (Kumar et al., 2011). As the names suggest, the private key is only known by the signing party, and the public key is available to everyone to verify the signature.

Hashing: Mapping between data of an arbitrary size onto data a fixed size (Dworkin, 2015). When the original data is unknown, it is very difficult to reconstruct it from the hash value, which makes it a good candidate to ensure the integrity of a transaction.

Point of failure: A potential risk caused by a poor system design were a single fault at that point can affect the correct functioning of the system.

3 Blockchain

In simple terms, a blockchain is a linked sequence of records of the transactions of digital assets (Fig. 1). These transactions can be of different types, including data creation (adding new data to the blockchain) or transfer (transferring the ownership of the data to another party, or to the same owner to edit data). The best known assets are crypto-currencies (e.g.: bitcoin), but in practice can be anything that can be represented by data. Each new transaction is cryptographically signed using the party's private key which is verified against the public key (included in the asset). The transaction also includes a hash that is generated using its public-private key-pair and the hash of the previous block. Any attempts to modify a block that has already been incorporated into the blockchain would change the signature and the hash of that transaction, which can be detected.
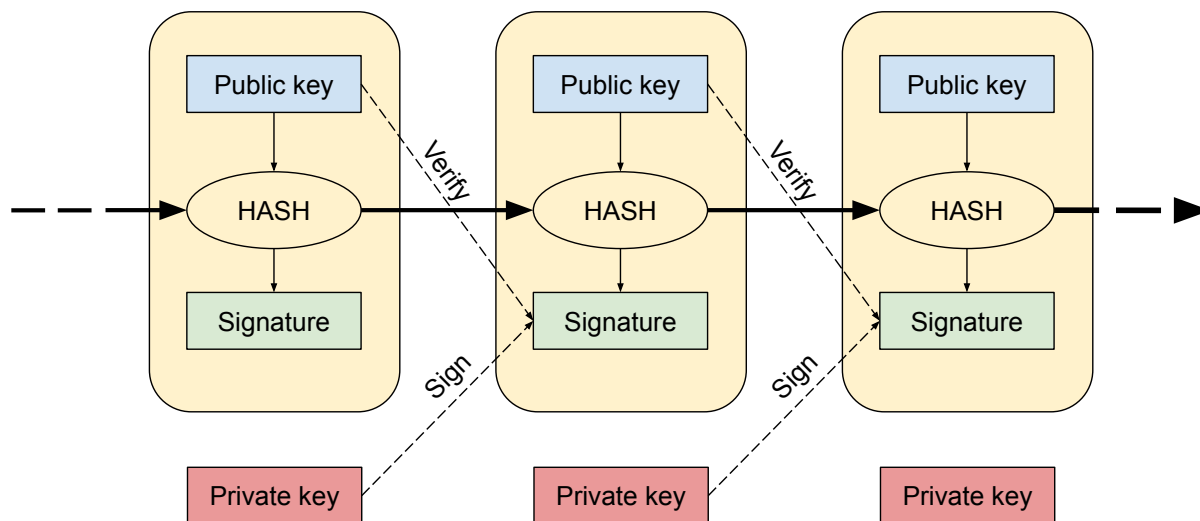
Figure 1. Diagram of three consecutive blocks (two transactions) within a blockchain.
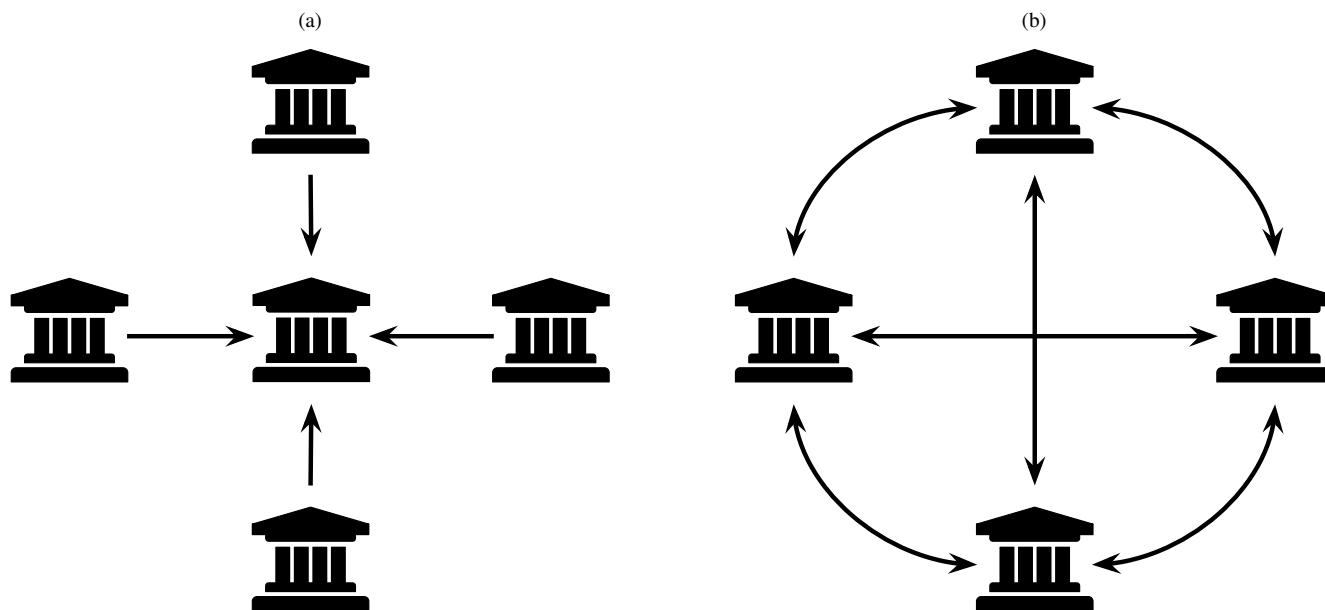
Figure 2. Data flows in two different soil information systems infrastructures (a) centralised (b) decentralised.

By design, a blockchain usually operates within a network of interconnected nodes (Fig. 2b). Each node keeps a copy of the chain (public ledger) and acts as validator, assuring the validity of new transactions. After enough nodes have reached consensus about the validity of the transaction, the new data-block is appended to the chain.

4   A soil data-sharing platform based on blockchain

Besides providing a solution to the aforementioned problems, namely centralised data control and governance, a blockchain has other characteristics that makes it an interesting candidate for a IIDB. Some of these solutions and characteristics are described in this section.

4.1   Decentralisation

As mentioned before, the main characteristic of a blockchain is the decentralised nature of the system. Each node of the network keeps a copy of the blockchain, which is synchronised after every new transaction (creation or transfer). Assuming that each node of the network is controlled by a different party, there is no centralised data storage, hence no single point of failure or control. Normally, in a well-designed, diverse network, a significant number of the nodes can be compromised without affecting its integrity.

Because all the nodes have a copy of the blockchain and act as validators, malicious modifications to the data are very difficult (see immutability section). The only possible way of tampering the data is if most of the nodes are colluded, which can be avoided by ensuring a diverse network.

For intra-institutional data sharing, a blockchain system can also be implemented to replace a traditional, permissioned database. The advantages are similar to inter-institutional case, including each team leader having "ownership" of their data, data traceability, data access logging and potentially preventing unauthorised access, and preventing malicious modifications or deletions. Data is one of the most valuable assets of any company (ref) and adding this extra layer of security to ensure its integrity should be a priority, and even mandatory for publicly funded institutions.

4.2   Data ownership

When a new asset is created it is cryptographically signed and assigned to one or more users. Only those users have the capability of transferring that asset even if the whole blockchain data is available at every node. The transactions are signed with the party's private key, and the validation nodes ensure that the signature match with the owner(s) public-key(s) before proceeding.

4.3   Immutability

Since the blocks of the chain are linked (Fig. 1), in practice, it is not possible to remove or edit a transaction. When a party needs to change the content of some data that they own, the new edited version is appended to the blockchain. This design feature produces data redundancy but also makes possible to keep the history of every asset, which is key for auditability. Thanks to this immutability, the parties within the network can always trust that the data is reliable and that it has not been tampered.

4.4 Data Governance

Data governance defines the norms, principles and rules under which the activities of a consortium should be conducted. It might include important details such as data release and rights to publish with consortium data first, research output rules (e.g. authorship sequence in consortium publications), if the data should be shared with non-consortium members (Singh and Daar, 2009), and the addition of new members.

In a centralised network, data governance is stipulated on an agreement and any modifications can be agreed between the members. In practice, control and governance over the data are ceded to the central node and the system has no way to prevent that unilateral changes are made. Using blockchain, the network is collectively governed, allowing the participants to cast votes to make important changes such as adding new nodes.

5 Potential use case: global soil spectral library

Although a blockchain data-sharing model has applications at many levels of granularity (inter- and intra-institutional, and international), we would like to focus on the use case of creating a multi-party (e.g., multi-institutional, multi-national, global) soil spectral library. Spectral soil data can be compared to the digital fingerprint of a particular soil sample which encodes information about its physical, chemical and biological properties (Grunwald, 2016). In pedometrics, a discipline that applies quantitative methods to study the variation of soils, the use of spectral data in conjunction with statistical or machine learning models to predict soil properties is already broadly implemented (Padarian et al., 2019b). Nevertheless, the development and application of models derived from spectral data still presents a series of challenges. For instance, models derived from local data, despite showing a good performance, have a limited applicability to other areas since they might lose their validity (Grinand et al., 2008). A potential solution is to develop models trained on data obtained from larger extents, which can be then "localised", taking advantage of the global knowledge to make predictions at the local scale (Padarian et al., 2019a). This approach has shown significant improvements of local predictions for multiple soil properties. To take full advantage of these advanced models, and since they are considered as "data-hungry" methods, it is recommended to train them on a large soil spectral library. Of course, collating a large spectral library that spans a large extent is not a trivial task. This is when collaboration and data sharing becomes important. Multiple, independent organisations can join efforts to reach a solution to a problem that it is very difficult to solve independently, which yields institutional (local) benefits greater than what it is possible working in isolation.

After all the efforts from different institutions to collaborate in a common initiative, it is only fair that the data-sharing infrastructure is carefully designed to ensure a democratic access, and control and governance over the data. We believe that, in general, a decentralised system can guard those interest for all parties involved. Particularly in the case of a global soil spectral library, the use of a decentralised database is of critical importance since the resulting database could be used by national reference centres for soil analysis. The level of transparency and security that a distributed ledger offers ensures that the reference data has not been tampered and also, given its decentralised model, will maximise accessibility. In the following

Snippet 1. Example *Sample* asset. The soil property codes are just for illustration purposes.

```
{
    "date": "2019−05−27",
    "user": "University1",
    "top": 0,
    "bottom": 10,
    "oc": 5.2,
    "sand": 10,
    "silt": 20,
    "clay": 70,
    "bd": 1.1,
    "spectra": "EOhM2lSd6j/o19ZP/9nqP+..."  # This spectra is encoded as base64
}
```

sections we explore certain implementation aspects of a decentralised data-sharing system in the context of a global soil spectral library.

5.1 Consortium initiation

The potential members of the consortium would have enough analytical capacity to measure the spectral response of soil
5 samples and also to perform laboratory analyses to measure the corresponding physical, chemical and biological soil properties. This includes universities and commercial soil laboratories from different countries.

Each member should have available the computational infrastructure to become a node of the network. The requirements are not prohibitive and include enough capacity to store all the data and internet connection. Each node should generate their public-private key-pair, securely store a copy of the private key and distribute the public key to the rest of the members. To
10 start the network, all the public keys should be known by all the members. Once the network is functional, more nodes can be added with the approval of most of the current members via an election process.

In terms of the network users, it is possible to have multiple users per node (e.g. different researchers from a single University). Ideally, all the users should have their own public-private key-pair to sign their transactions, and their public keys should be known to all the users. This information can also be stored in the blockchain as a public ledger of who can access the data.

15 ## 5.2 Providing data

After the network is functional, any member can create new transactions to add data that will be synchronised between all the nodes, ensuring immediate accessibility to the data to all the members. The structure of what constitutes an "asset" should be defined during the initiation period. For instance, the asset could be a single soil sample with its corresponding analytical data (Snippet 1). The system should support the use of numerical and text data to store all the necessary soil properties and
20 metadata. Complex data structures such as soil spectral data can be stored as comma separated numbers or compressed.

The new transaction should be signed with the user's private key and the asset ownership set to the user's public key. This provides a way of authenticating the origin of the data and allows the node, and only that node, to create updated versions of

that asset if needed (e.g. when new properties are measured or to correct potential errors). Before a new transaction is appended to the blockchain, a "super-majority" of the nodes (usually a large proportion such as at least 2/3) must agree on the validity of that transaction. The most basic validation is to assure that the owner(s) are signing the transaction, but in practice it is possible to set any logical rules. This provides the opportunity to give certain groups of users the control over an asset, define minimum
5   number of owners, perform basic data integrity checks, etc.

5.3   Retrieving data

Since every node keeps a copy of the blockchain locally, it is possible to retrieve data from any node from the network, providing extra redundancy and hence assuring accessibility in case of malfunction of some of the nodes. Advanced users can query their local copy of the database directly. A friendlier way of providing access to read the data is via an Application Programming
10   Interface (API) that connects any user with a node. That API can perform tasks such as querying the blockchain to retrieve specific data, provide the history of any asset, and potentially process data using pipelines approved by the consortium.

5.4   User interaction

Most of the specific blockchain operations (i.e. signing and verifying transaction) are performed in the background. There is no extra overhead for the users besides keeping their respective private keys safe. A user interface can be build on top of an API
15   so users can access the system as it were a traditional data management system (DMS), with capabilities to query and retrieve the data from the network.

In terms of the type of users with access to the system, any person with access to a node have complete reading access to the blockchain. If public access is required to allow non-consortium members to connect to the database, multiple solutions are available including single or multiple nodes acting as a web server. Using multiple nodes as web servers might reduce
20   latency, specially when the consortium spans different countries (i.e. an external user can connect to the closest node). Again, a platform can be build to ensure the public experience is identical to a normal DMS.

6   Summary

The prevalent soil data-sharing model is centralised, with users ceding their control and governance over their data to a lead party. We propose the use of a public ledger (blockchain) to create a decentralised soil data-sharing network. This network
25   provides a series of advantages to the participant institutions, including:

- allowing institutions to preserve the ownership and control over their data,

- instant access to the complete database,

- ensures that once the data is appended to the blockchain it cannot be tampered,

- actively participate in governance decisions such as adding new members through elections facilitated by the system.

Ultimately, any consortium data-sharing agreement is based on trust between the participants. By using a blockchain network, the need of trust is removed since rules can be programmatically enforced and the data becomes tamper-resistant. This protects the already existing trust-bond between the consortium members and, potentially, allows the consortium to expand its reach by working with new parties that are not fully-trusted.

5   For intra-institutional data sharing, a blockchain system can also be implemented to replace a traditional, permissioned database. The advantages include each team leader having "ownership" of their data, data traceability, data access logging and potentially preventing unauthorised access, and preventing malicious modifications or deletions.

Competing interests. The authors declare that they have no conflict of interest.

References

Arrouays, D., Leenaars, J. G., de Forges, A. C. R., Adhikari, K., Ballabio, C., Greve, M., Grundy, M., Guerrero, E., Hempel, J., Hengl, T., et al.: Soil legacy data rescue via GlobalSoilMap and other international and national initiatives, GeoResJ, 14, 1–19, 2017.

Dworkin, M. J.: SHA-3 standard: Permutation-based hash and extendable-output functions, Tech. rep., 2015.

5 Grinand, C., Arrouays, D., Laroche, B., and Martin, M. P.: Extrapolating regional soil landscapes from an existing soil map: sampling intensity, validation procedures, and integration of spatial context, Geoderma, 143, 180–190, 2008.

Grunwald, S.: Environmental soil-landscape modeling: Geographic information technologies and pedometrics, CRC Press, 2016.

Kumar, Y., Munjal, R., and Sharma, H.: Comparison of symmetric and asymmetric cryptography with existing vulnerabilities and counter-measures, International Journal of Computer Science and Management Studies, 11, 2011.

10 Padarian, J., Minasny, B., and McBratney, A.: Transfer learning to localise a continental soil vis-NIR calibration model, Geoderma, 340, 279–288, 2019a.

Padarian, J., Minasny, B., and McBratney, A. B.: Machine learning and soil sciences: A review aided by machine learning tools, Soil Discussion, 2019b.

Rossel, R. V., Behrens, T., Ben-Dor, E., Brown, D., Demattê, J., Shepherd, K. D., Shi, Z., Stenberg, B., Stevens, A., Adamchuk, V., et al.: A
15 global spectral library to characterize the world's soil, Earth-Science Reviews, 155, 198–230, 2016.

Singh, J. A. and Daar, A. S.: Intra-consortium data sharing in multi-national, multi-institutional genomic studies: gaps and guidance, Hugo J, 3, 11–4, 2009.