Blockchain and Smart Contract for Trusted Decentralized Digital Genomics

Adnan Imeri, Nazim Agoulmine, Djamel Khadraoui

Blockchain and Smart Contract for Trusted Decentralized Digital Genomics

Adnan Imeri*
adnan.imeri@list.lu
Luxembourg Institute of Science and Technology & Université Paris-Saclay (UEVE)
Esch-sur-Alzette, Luxembourg

Nazim Agoulmine
Université Paris-Saclay (UEVE)
Évry-Courcouronnes, France
nazim.agoulmine@univ-evry.fr

Djamel Khadraoui
Luxembourg Institute of Science and Technology
Esch-sur-Alzette, Luxembourg
djamel.khadraoui@list.lu

ABSTRACT

Genomics data enables benefits in understanding human health and potential diseases. Besides being very useful in various medical fields, they encounter considerable issues regarding the usability of genomic data by unauthorized third parties. The unauthorized use or misuse of genomics data inflicts privacy and ethical problems. Several regulatory frameworks aim to regulate the use of genomic data, and parties involved and working with such data should comply with the regulatory framework. Ensuring compliance aspects when such data are shared imposes a clear challenge. The technologies used for processing, sharing, and maintaining genomic information do not have the capability to formally and empirically ensure that these data are used in compliance with the regulatory framework and consented to as determined by the data owner (based on the regulatory framework). We propose new technologies, such as blockchain and smart contracts, for improving information sharing in the genomic ecosystem.

KEYWORDS

blockchain, smart contract, information sharing, trust, automation, genomic data

ACM Reference Format:
Adnan Imeri, Nazim Agoulmine, and Djamel Khadraoui. 2018. Blockchain and Smart Contract for Trusted Decentralized Digital Genomics. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX)*. ACM, New York, NY, USA, 5 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

The technological components of the blockchain (BC) have disturbed various data-related domains, particularly the ones related to secure information sharing, and privacy concerns are a significant issue. Among such domains, genomics data (DG) imposes immense information sharing and privacy concerns.

GD is foreseen to be revolutionary in improving the health and well-being of humans. It has been applied in different use cases, such as biomedical, clinical, pharmaceutic, and general-purpose research [18]. Among the main characteristics that GD might be helpful to are the ability to use them in diagnostic testing, which enables testifying if an individual is affected and might be exposed in the future to specific diseases [4]; determining to parent among two individuals [16] and developing efficient drugs and therapies by using pharmacogenetics tests [8]. Therefore, it is essential to ensure good treatment of the GD to exploit it and engage donors in the processes by improving trust and transparency in GD management (storage, access, processing, and transformation). We aim to present a conceptual approach that enables trusted GD management.

This paper is structured as follows. Section 2 shows related works regarding GD management. In Section 3, we present BC technology and its characteristics. Section 4 presents our scientific approach. In Section 5, we show our conclusion and future works.

2 RELATED WORK STUDIES: PRIVACY, SECURITY, AND GOVERNANCE-RELATED ISSUES ON MANAGING GD

Besides the significant progress in using GD, it raises privacy-related issues when such data are treated. Privacy issues are increased when GD is collected and used for adversary purposes. Privacy attacks such as the identity attack [6, 18, 22] performed by collecting donors' demographic data and performing specific data matchmaking, statistics, and probabilistic prediction enable the triangularity of donor identity. Disclosing the donor's sensitive attributes (possible deserts association) presents an attribute disclosure attack [7, 11, 18], and being able to reconstruct or predict partially or "entirely" the doners genotype gives a completion attack [5, 10, 17]. Moreover, privacy-related issues are identity tracing [20], attribute disclosure, and completion attacks [18, 24]. Following well-known privacy related is the identification of participants associated with genome projects [22], attribute disclosure and identifying the presence of an individual [3, 11], re-identify an individual by performing long-range family relatives [9], re-identify individuals and their relatives within a beacon [18, 20], using GD for prediction of disease for an individual [12] and many other related privacy issues. Another related issue is the trust of donors related to the storage and processing of GD. The trust-related concerns for GD come from the fact that the data stored are not tamper-resistant. The GD attacker might access, use, and predict information about individuals [3, 11, 18]. Researchers have applied several cryptographic techniques for GD security, such as "Homographic encryption", "Differential privacy", "Secure cryptographic hardware" and "Secure multi-party computation" [18]. The cryptographic primitives

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

are applied to the data layer and cover only some of the spectrum of GD security aspects. Therefore, they remain limited in the scope of operation towards privacy issues. GD is generally stored in public and private clouds [1, 2]. Besides SLA, which describes a legal guarantee that the information is treated according to specified rules, there needs to be formal proof (neither empirical) that guarantees the data are treated accordingly. Furthermore, controlling these data from the donor [1] By having a certain level of sovereignty, i.e., deciding with whom this information is currently shared, is impossible. In the GD sharing ecosystem, the data are shared among stakeholders (donor, private or public entities, and third parties involved, i.e., government). The data integrity (usability according to regulatory framework and ethics) and traceability of sharing this information are beyond the donor's perspective. The pieces of evidence on data sharing, even those that are possible with existing exploited technologies for the GD, are still subject to the possibility of being tampered with by different actors involved (or unauthorized parties) and do not guarantee sufficient transparency. Therefore, to improve such issues, we propose a BC-based conceptual solution to improve transparency in GD sharing and management. Following the literature review, we enlist the identified challenges related to GD:

(1) Privacy of GD remains one of the main challenges. Tracking and sharing identity-related data based on genomic information concerns many individuals [18].
(2) Lack of transparency on processing GD as the current technologies fail to ensure immutable end-to-end traceability of information.
(3) Lack of data governance from the user perspective. Users do not have any control over the information they intend to share.
(4) Allowing access/query GD from unauthorized and unidentified parties.
(5) The applied cryptography primitives do not solve the problems entirely as they are focused on specific levels, e.g., data, while other issues, such as governance aspects, authorized access, and process GD, where an immutable log of action is not considered.
(6) Trust and confidentiality issues where information immutability is not guaranteed, and pieces of evidence over information sharing with unauthorized parties are not available.
(7) Architectural issues
   (a) Centralization and cloud-based architectures with limited transparency
   (b) Lack of integration and upgrade into new technological advancements

2.1 Identified challenges. GD beyond the state of the art

Besides the previously mentioned significant research challenges, there are enormous concerns related to privacy, security, and management of GD. We present the context, considering relevant aspects of non-functional requirements and technology-related issues that are only partially considered in the literature.

- In the GD ecosystem, there are many actors (stakeholders) involved, e.g., donors, research centers, biomedicine centers, authorities, and many others, and which one with different access rights and operation level (create, add, remove, modify, transfer, store, etc). That makes this complex and almost impossible to manage with traditional technologies.
- Not considering entirely (including regulatory framework) the functional and non-functional requirements of the stakeholders, donors, and regulatory frameworks.
- Not considering compliance aspects related to data policies and regulatory framework in the solution design and not applying the policy aspect correctly and comprehensively at the technology level.
- Technological issues. Choosing technologies that do not have the ability to support security and privacy-preserving aspects entirely or partially.
- Centralization storage of GD. The current approach considers GD storage in a centralized cloud database, which limits traceability and transparency over the processing of GD.
- Inconsideration of the governance aspects. GD are entirely stored, processed, and shared by specific stakeholders. The donor does not have any governance impact once the data are provided to these stakeholders.
- Not allowing donors to have a certain sovereign level over the provided data. That is sourced from the fact that confidentiality is not guaranteed formally or empirically, as the chosen technologies do not have such ability to

We propose a research activity, i.e., direction/theme, to holistically address the issues mentioned above. The objective is to advance the knowledge and thinking process about enforcing trust, transparency, traceability, and availability of GD among different stakeholders. Initially, we introduce BC technology and related technologies that we intend to propose in our scientific approach.

We consider that applying BC technology in a complex ecosystem composed of multiple stakeholders, such as GD spaces, advocates for better process management. We believe that BC technology has much better features compared to classical technology when it comes to information security, trust, transparency in information sharing, advanced control over data sharing from authorities' perspectives, enabling decentralized governance, and improving interoperability.

3 BLOCKCHAIN TECHNOLOGY AND ITS COMPONENTS

"BC is a distributed decentralized database that allows the storage of append-only transaction data. The BC network comprises several decentralized nodes that communicate with each other in a peer-to-peer mode. All the nodes in the BC network contain the same copy of the ledger, and they rely on communication in distributed nodes, thus avoiding any central authority [13, 15, 23, 25]. "The BC nodes gather transactions into "blocks". The transactions are initially validated by performing cryptographic checks (public-private key cryptography). The block of transactions stored in the BC is immutable, and cryptography tools ensure data integrity"[15] [13, 23, 25]. "Among the main BCs, the fundamental characteristic is that the block of data is linked together, so block N contains

[1] A donor/user presents any patient who accepts to give (donate) GD for research purposes. We refer to users as donors in the jargon of GD.

the hash address of the previous block N-1" [15] [13, 23, 25]. "The tendency to change the information stored in the BC is denied by consensus protocol, which verifies the state of data"[15] [13, 23, 25].

3.1 Smart Contract

"The smart contract (SC) is an autonomous computer code encoded to be self-executed and linked to a specific task [13]. It is deployed on the BC and executed based on its specifications to perform a particular task. SC implements a certain level of business logic and, in combination with BC technological capabilities, constitutes a powerful tool to solve information-related issues such as transparency, traceability, immutability, availability, and interoperability" [13, 14, 23]. "SC logic is mainly based on domain-specific. It encodes any set of rules emerging from the source of the SC into the programming language. The SC source can be a natural language law, the scope of any agreement between parties, and other possible sources depending on the business process requirements. For the transaction that is accepted on the BC, and if it contains the SC address as a message received, the miners will execute the SC code and react according to the SC's specific tasks. A SC is a self-executed program; moreover, it can invoke another SC, call an external service (oracles), fulfill given tasks, and implement and automate a wide range of domain-specific applications." [13]

3.2 Self-Sovereign Identity (SSI)

Self-sovereign identity (SSI)[2] enables an individual, object, or device to win and control their identity with third-party involvement [19, 21]. The SSI is based on two standards provided by the World Wide Web Consortium (W3C):

Decentralized Identifiers (DIDs)[3]: A new way for individuals to generate unique identifiers that allows interaction with the digital world.

Verifiable Credentials (VC)[4]: These digital credentials contain attributes that are a self-issued or authorized party (government, regulatory organization). Our aim in using SSI and BC is to be able to issue VC for GD sets, donors, stakeholders, and other parties involved. That enables decentralized authentication and control of GD space to improve privacy, storage, and data sharing.

4 SCIENTIFIC APPROACH FOR SECURE SHARING AND MANAGEMENT OF GENOMIC DATA

(A) **Trusted data sharing and management.** From the GD sharing perspective, having technological features that allow tracking and logging transactions on data sharing and versioning (possible modification) improves transparency and trust of users towards GD stakeholders. For such technological features, we propose using the immutability of information features offered by distributed ledger technologies, i.e., BCs. The GD user, i.e., donner, might verify if the information is shared with relevant parties as indicated in the information-sharing policy. The **information-sharing** policy will be available for definition in cooperation between GD donors, stakeholders, and authorities. It will allow the GD used to indicate which type of information can be shared for a specific group of users and to what extent. This approach is based on a BC and smart contract that will use a specific API service that further identifies and writes data changes (sharing the files (folders), versioning, Decentralized Identifiers (DID), and Verifiable Credentials (CV)) on the ledger based on an information-sharing policy.

(B) **Regulatory Framework procedures semi-automatically adjusted in the GD ecosystem.** Data migration, sharing, or processing in the GD ecosystem from one geographic area to another (cross-border) might be subject to different rules and guidelines (regulatory framework, standards, and policies). Such rules, sourced from the regulatory framework, might deny information sharing with the specific user (organization or entity) or processing such information in a specific context. Similarly, a particular rule sourced from the regulatory framework might require denying access to GD and services to a specific genomic stakeholder (or private user) located in another country (different jurisdiction). We propose using BC and smart contracts to automatically resolve such a situation, automatically triggering different procedures on GD sharing according to the rules and procedures. Using smart contracts to cover automatic detection of the regulatory framework will be possible by reading the API GD sharing ecosystems. Based on the information provided by these APIs, the smart contract triggers specific action (based on rules related to that action) and denies sharing of information or disabling any service related to GD.

4.1 Conceptual approach to fulfill proposed scientific topics from A and B

Our approach is based on **BC** and **smart contracts and** consists of Trusted Decentralized Digital Genomic (TDDGen). For the TDDGen approach, we propose two scientific concepts: the **genomic multiparty smart contract (GMPSC)** for GD sharing and management and the **digital genomic certificate** that enables traceability and transparency of using GD. To fulfill the GMPSC, our methodology consists of the following steps:

(a) Analysis of ontological information related to GD sharing and management.
  (i) Use specific methodology for information extraction as well as domain expert advice.
  (ii) Consideration of industry (business) requirements to involve them in the model.
(b) Designing data sharing model (information sharing policy) based on official policy-related
  (i) Ontology and reasoning model
  (ii) Logical-based approach
(c) Model Upload: Engage mode into GMPSC.
  (i) Requires design and development of a method to translate the ontological model into GMPSC code (smart contract). This method is based on the model driven engineering method, which will enable the creation of an engineering method for model translation/applicability.

[2]https://ssi-ambassador.medium.com/an-introduction-to-self-sovereign-identity-ssi-916eb42f0490
[3]https://www.w3.org/TR/did-core/
[4]https://www.w3.org/TR/vc-data-model/

(ii) Designing and developing an engine to decode workflow for TDDGen.

(d) API-based plug-in service that enables engaging different Smart Contracts into GMPSC

(i) Enables applicability in different use cases with different conditions aligned with b and c.

The points mentioned in b), c), and d) intend to create standards for GD management. This standard proposes a broad set of smart contract functions, enabling applicability in different GD sets and contexts. Changing context and applying it to different GD sets means using another model, e.g., switching it on b) and applying it in c). That indicates the core approach remains the same; only the model changes.

Figure1 presents the conceptual approach of TDDGen, including GMPSC. The approach considers regulatory framework documents (Corpus of legal documents) to ensure a transparent and trusted digital GD sharing and management lifecycle. For the involved stakeholders (SGT), which might be national or international, we propose using specific technological components (API-based layer) that rely on smart contracts to ensure the traceability, compliance, and usefulness of GD. SGTs can interact with BC[5] using GMPSC. GMPSC will play the policy "data guard" role to enable particular automation of procedures by applying different ontological models sourced from regulatory and business rules (defined in a and b). The digital details, i.e., the instance of meta-data of GD, an example of digital components of STGs, and other related instances, will be used to enable interaction with different GD sets.

In general, for the stakeholder onboarding, authentication, and authorization, we propose an SSI approach. We enable decentralized authentication of stakeholders, and any operation can be carried out only by authorized stakeholders (according to the data sharing policy). As mentioned in the previous section, DID enables a unique identification of stakeholders, authorities, and donors. We aim for the self-sovereignty of multi-parties over GD according to GD sharing policy.

Complementary in the TDDGen, we propose the concept of a Digital Genomic Certificate (DGC). The DGC plays the role of an information assistant used to enlist stakeholders, maintain additional rules and regulations expressed digitally, collect and store DIDs and VC, and track and trace specific actions (access, processing, and transfer) of GD.

The proposed conceptual approach is quantified as a second layer of hybrid digital GD management. The "hybrid" concept introduces the treatment of meta-data for GD, then the management of interactions (information sharing, processing, and storage) via smart contract, which does not necessarily access the GD, and finally, the applicability of this approach in different contexts and datasets. The whole and unmodified datasets of GD remain in the first layer.

At a glance, our approach aims at the following:

(1) Researching for applying BC and smart contracts further to enhance information security and privacy of GD. Moreover, we intend to use several anonymization techniques to secure GD sharing. That relies on a combination of smart contra abilities and different encryption techniques.

(2) Ensuring a certain level of sovereignty for donors by enabling them to decide which data to share and with whom. This technique is based on SSI and BC. Self-covering identification of stakeholders, donors, and other involved parties. Via the SSI approach, we propose a new way of managing. This technique allows for the sharing of genomic information only with the authorized party.

(3) Proposing a new decentralized architectural solution for GD sharing and management.

(4) Governance of GD with the help of the smart contract (combination of on-chain and off-chain techniques, based on regulatory framework studies in B). The core idea is to have a certain level of control over access to GD sets.

(5) Automation aspects, to secure information sharing with the help of smart contract based on knowledge and reasoning achieved by regulatory (policy) framework.

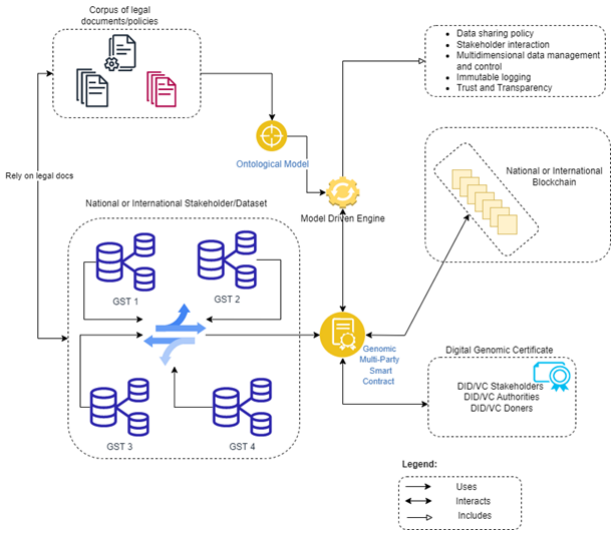

Figure 1: The conceptual presentation of the TDDGen for GD sharing and management.

5 CONCLUSION AND FUTURE WORKS

BC technology, Smart Contact, and SSI and its related components are new research areas combined to solve the privacy issues related to GD. In this research paper, we presented a research theme towards GD. Initially, we presented privacy-related problems related to GD. The technological characteristics of BC, smart contracts, and SSI were presented. Further, we proposed a scientific approach that aims to reduce privacy-related concerns significantly.

For future works, we aim to extend this approach by conducting systematic and rigorous research to foster innovation in secure information sharing by considering intensely privacy-related attributes. We aim to develop the mentioned approach into proof of concepts to measure technological capabilities for supporting genetic data sharing and management.

[5]The used blockchain can be governed by a national organization or any other international BC such as EBSI. This needs to be clarified with national objectives and a regulatory framework governing GD.

ACKNOWLEDGMENTS

This research was supported by Genopole[6] under the ATIGE grant.

REFERENCES

[1] [n. d.]. http://www.internationalgenome.org/
[2] 2023. http://www.ukbiobank.ac.uk/
[3] Ruichu Cai, Zhifeng Hao, Marianne Winslett, Xiaokui Xiao, Yin Yang, Zhenjie Zhang, and Shuigeng Zhou. 2015. Deterministic identification of specific individuals from GWAS results. *Bioinformatics* 31, 11 (2015), 1701–1707.
[4] Xi Chen and Hemant Ishwaran. 2012. Random forests for genomic data analysis. *Genomics* 99, 6 (2012), 323–329.
[5] Iman Deznabi, Mohammad Mobayen, Nazanin Jafari, Oznur Tastan, and Erman Ayday. 2017. An inference attack on genomic data using kinship, complex correlations, and phenotype information. *IEEE/ACM transactions on computational biology and bioinformatics* 15, 4 (2017), 1333–1343.
[6] Yaniv Erlich, Tal Shor, Itsik Pe'er, and Shai Carmi. 2018. Identity inference of genomic data using long-range familial searches. *Science* 362, 6415 (2018), 690–694.
[7] Matthew Fredrikson, Eric Lantz, Somesh Jha, Simon Lin, David Page, and Thomas Ristenpart. 2014. Privacy in pharmacogenetics: An {End-to-End} case study of personalized warfarin dosing. In *23rd USENIX security symposium (USENIX Security 14)*. 17–32.
[8] David B Goldstein, Sarah K Tate, and Sanjay M Sisodiya. 2003. Pharmacogenetics goes genomic. *Nature Reviews Genetics* 4, 12 (2003), 937–947.
[9] Melissa Gymrek, Amy L McGuire, David Golan, Eran Halperin, and Yaniv Erlich. 2013. Identifying personal genomes by surname inference. *Science* 339, 6117 (2013), 321–324.
[10] Zaobo He, Jiguo Yu, Ji Li, Qilong Han, Guangchun Luo, and Yingshu Li. 2018. Inference attacks and controls on genotypes and phenotypes for individual genomic data. *IEEE/ACM transactions on computational biology and bioinformatics* 17, 3 (2018), 930–937.
[11] Nils Homer, Szabolcs Szelinger, Margot Redman, David Duggan, Waibhav Tembe, Jill Muehling, John V Pearson, Dietrich A Stephan, Stanley F Nelson, and David W Craig. 2008. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS genetics* 4, 8 (2008), e1000167.
[12] Mathias Humbert, Kévin Huguenin, Joachim Hugonot, Erman Ayday, and Jean-Pierre Hubaux. 2015. De-anonymizing genomic databases using phenotypic traits. In *15th Privacy Enhancing Technologies Symposium (PETS)*, Vol. 2015. 99–114.
[13] Adnan Imeri. 2021. *Using the blockchain technology for trust improvement of processes in Logistics and Transportation.* Ph. D. Dissertation. University of Luxembourg, Esch-sur-Alzette, Luxembourg.
[14] Adnan Imeri, Nazim Agoulmine, Christophe Feltus, and Djamel Khadraoui. 2019. Blockchain: Analysis of the new technological components as opportunity to solve the trust issues in supply chain management. In *Intelligent Computing: Proceedings of the 2019 Computing Conference, Volume 2*. Springer, 474–493.
[15] Adnan Imeri, Christophe Feltus, Nazim Agoulmine, and Djamel Khadraoui. 2022. The Blockchain-Based Digital Certificate for the Transport of Dangerous Goods. In *Blockchain Driven Supply Chains and Enterprise Information Systems*. Springer, 43–61.
[16] Mark A Jobling, ARPITA Pandya, and CHRIS Tyler-Smith. 1997. The Y chromosome in forensic analysis and paternity testing. *International journal of legal medicine* 110 (1997), 118–124.
[17] Christoph Lippert, Riccardo Sabatini, M Cyrus Maher, Eun Yong Kang, Seunghak Lee, Okan Arikan, Alena Harley, Axel Bernal, Peter Garst, Victor Lavrenko, et al. 2017. Identification of individuals by trait prediction using whole-genome sequencing data. *Proceedings of the National Academy of Sciences* 114, 38 (2017), 10166–10171.
[18] Abukari Mohammed Yakubu and Yi-Ping Phoebe Chen. 2020. Ensuring privacy and security of genomic data and functionalities. *Briefings in bioinformatics* 21, 2 (2020), 511–526.
[19] Alex Preukschat and Drummond Reed. 2021. *Self-sovereign identity.* Manning Publications.
[20] Jean Louis Raisaro, Florian Tramer, Zhanglong Ji, Diyue Bu, Yongan Zhao, Knox Carey, David Lloyd, Heidi Sofia, Dixie Baker, Paul Flicek, et al. 2017. Addressing Beacon re-identification attacks: quantification and mitigation of privacy risks. *Journal of the American Medical Informatics Association* 24, 4 (2017), 799–805.
[21] Sovrin. 2018. Sovrin. https://sovrin.org/faq/what-is-self-sovereign-identity/ Accessed: 4 November 2023..
[22] Latanya Sweeney, Akua Abu, and Julia Winn. 2013. Identifying participants in the personal genome project by name (a re-identification experiment). *arXiv preprint arXiv:1304.7605* (2013).
[23] Xiwei Xu, Ingo Weber, Mark Staples, Liming Zhu, Jan Bosch, Len Bass, Cesare Pautasso, and Paul Rimba. 2017. A taxonomy of blockchain-based systems for architecture design. In *2017 IEEE international conference on software architecture (ICSA)*. IEEE, 243–252.
[24] Yanjun Zhang, Guangdong Bai, Xue Li, Surya Nepal, Marthie Grobler, Chen Chen, and Ryan KL Ko. 2022. Preserving Privacy for Distributed Genome-Wide Analysis Against Identity Tracing Attacks. *IEEE Transactions on Dependable and Secure Computing* (2022).
[25] Zibin Zheng, Shaoan Xie, Hong-Ning Dai, Xiangping Chen, and Huaimin Wang. 2018. Blockchain challenges and opportunities: A survey. *International journal of web and grid services* 14, 4 (2018), 352–375.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009

[6] https://www.genopole.fr/