A Blockchain-Based Approach to Provenance and Reproducibility in Research Workflows

Kevin Wittek
Institute for Internet Security
Westphalian University of Applied Sciences
RWTH Aachen University
Gelsenkirchen, Germany
wittek@internet-sicherheit.de

Neslihan Wittek
Department of Biopsychology
Ruhr University Bochum
Bochum, Germany
neslihan.wittek@rub.de

James Lawton
Max Planck Digital Library (MPG)
Max Planck Society
Munich, Germany
lawton@mpdl.mpg.de

Iryna Dohndorf
Karakun GmbH
Dortmund, Germany
iryna.dohndorf@karakun.com

Alexander Weinert
Institute for Software Technology
German Aerospace Center (DLR)
Cologne, Germany
alexander.weinert@dlr.de

Andrei Ionita
Fraunhofer FIT
Fraunhofer Society
Sankt Augustin, Germany
andrei.ionita@fit.fraunhofer.de

***Abstract*—The traditional Proof of Existence blockchain service on the Bitcoin network can be used to verify the existence of any research data at a specific point of time, and to validate the data integrity, without revealing its content. Several variants of the blockchain service exist to certify the existence of data relying on cryptographic fingerprinting, thus enabling an efficient verification of the authenticity of such certifications. However, nowadays research data is continuously changing and being modified through different processing steps in most scientific research workflows such that certifications of individual data objects seem to be constantly outdated in this setting. This paper describes how the blockchain and distributed ledger technology can be used to form a new certification model, that captures the research process as a whole in a more meaningful way, including the description of the used data through its different stages and the associated computational pipeline, code for analysis and the experimental design. The scientific blockchain infrastructure bloxberg, together with a deep learning based analysis from the behavioral science field are used to show the applicability of the approach.**

***Index Terms*—Blockchain, ethereum, bloxberg, open science, integrity, reproducibility**

I. Introduction

Nowadays technology organizations, researchers, publishers and individuals are faced with huge amounts of research data and other artifacts originating from scientific studies. Examples range from original data, code and tests, scientific workflow descriptions, experimental design variables to execution environments (software as well as hardware), and manuscript files, to name just a few. Researchers from different fields such as social sciences, natural and life sciences encounter a reproducibility crisis in severe ways, leading to many scientific studies that are difficult and sometimes even unfeasible to replicate or reproduce [1]–[3]. This reproducibility and replicability complication arise from inappropriate scientific practices, which are mainly originating from a lack of raw data [4]–[6] and p-value hacking by selective data analysis and reporting [7]–[11]. This irreproducibility is responsible not only for time lost in scientific ends but, also inevitable economic loss. In the United States, the irreproducibility rate in preclinical research is around 50% which points out 28 billion USD loss per year and a big proportion of the errors are induced by study design, data analysis and reporting [12].

Therefore, reproducibility approaches enclosing verification methods for scientific workflow data properties, e.g. its provenance and integrity, strengthen not only the quality and credibility of research findings, but also the efficiency of economics of scientific research and innovation. As such, reproducible science has become increasingly important in a diverse set of research landscape fields. In fact, even modern scientific data management systems aim to design optimized policies to prevent loss of scientific data set quality, to increase its reuse and to preserve trustworthy digital data [13]–[17]. An increasingly popular approach to deal with verification of data set authenticity and provenance comes by leveraging blockchain and *distributed ledger technology (DLT)* [18]–[20]. While most public blokchain networks operate according to a *Proof of Work* consensus mechanisms, other distributed trust architectures can be established using a *Proof of Authority* consensus mechanism based on a consortium of international research organizations in order to provide a *Proof of Existence* blockchain service which can be used to verify the existence of any research data and to validate its integrity, while still preserving data privacy mechanisms [21]–[24].

Several variants of trusted timestamping services and methods for providing signed certifications for research data and their properties have already been established. Verification of research data (or any kind of data) for authenticity is

978-0-7381-1420-0/21/$31.00

often based on the application of cryptographic hash functions for fingerprinting, and the appropriate trust-worthy distributed infrastructures providing digital signatures in order to validate the data ownership, e.g. the web-based but Ethereum backed *bloxberg Certify* DApp, as well as data models for emerging identifier standards [21], [23], [25], thus, supporting the certification of any type of data that might occur as part of a scientific research workflow. Certification of a generic research object, such as data or documents, during the scientific research workflow can improve the transparency, reproducibility, data integrity, and foster open science [26]–[29].

In many applications, however, it appears to be of interest to incorporate and accurately record the whole research workflow and related artifacts including raw data, software dependencies, computational pipelines, code files, and experiment settings into the certification model, while also capturing a chronological sequence of data creation and change events. One can record how data has changed through the research process, for instance between initial acquisition and subsequent preprocessing. Then, considering a sequence of multiple generated certificates and the corresponding timestamps, the workflow of a research process can be certified and validated at a later point in time and by interested third parties. Tamper proof recording of data and research processes allows to develop verification processes, either completely automated or by providing means to support human reviewers, with more confidence and therefore ensure better guarantees and degrees of authenticity and reusability.

II. Related Work

To address the aforementioned issues, different solutions and technologies have been born from varied disciplines over the last decade. The simplest approach is sharing the data and the used code during the data acquisition and analysis [30], [31] by using a traditional version control system such as Git [32]–[34]. Besides using traditional version control methods in the scientific domain, open access collaborative research platforms have been developed [35], [36]. This is in line with the original interest of reproducibility and advancing computational research and science processes, where users can run the code and test for reusability, publish, archive, share and collaborate.

Another approach, specifically tailored to evolutionary genomics, helps researchers to create reusable and reproducible bioinformatics pipelines that can be deployed on a desktop workstation or in the cloud [37]. Various tools and standard software distributions used in evolutionary biology based on *Debian GNU/Linux*, *GNU Guix*, and *Bioconda* are provided along with containerized execution environments and distribution mechanisms such as *Docker*, *GNU Guix*, and *Singularity*. Bundling these open software distributions and defining reproducible and shareable pipelines using workflow engines such as the *Common Workflow Language (CWL)*, *Guix Workflow Language (GWL)*, *Snakemake*, and *Nextflow* provide extra time and productivity to the researchers, while also lending assistance to reproducible science.

Similarly, the creation of reusable workflows has also taken hold in the field of multidisciplinary design, analysis, and optimization. Here, numerous engineers from a variety of disciplines collaborate to design complex systems such as ships [38], aircraft [39], helicopters [40], or spacecraft [41]. Such analyses often comprise tools that are under rapid development during the course of an analysis. Thus, the dependencies between the tools employed, the tool versions, and the tools itself may change between subsequent runs of the same analysis. Current tool support captures both the inter dependencies of the tools and the data distributed among the tools in such an analysis, thus aiding transparency of the analysis itself [42]. Identifying software versions [43], storing such information on software and its computing environment in a trustworthy way [44] and making such analyses reproducible is a topic of ongoing research.

Recently, many technologies and container-based virtualization methods have been proposed to deal with efficiency, quality of service, reusability and portability of scientific workflows. On the one hand, a large number of technical advances and developments of infrastructure models with appropriate workflow specifications have been made [45], [46]. The infrastructures covered range from distributed environments to cloud solutions [47]–[50]. On the other hand, solutions that integrate blockchain technology are gaining particular interest and development. Examples include an architecture of blockchain-enabled service workflows with a focus on quality of service preserving mechanisms [51], as well as blockchain-based platform for storing and managing electronic medical records in the cloud [24], where concepts for data and properties were proposed but without consideration of the whole scientific workflow.

More recent research suggests using blockchain technology for the development of research workflow solutions for provenance, reliability, and collaboration [52]–[57], and in different steps of the research activities to ensure transparency in data acquisition, data processing, research fund distribution, researchers' contribution and the publication review process [58]–[61].

III. Use Case Description & Design Considerations

In order to explore the solution space, we use a machine learning supported animal study as a reference point to infer potential requirements and specifications as depicted in Fig. 1. Considering the ML supported research process, as shown in Fig. 1, our research objective is to develop a blockchain-based method for tracking and certifying research artifacts in each step of the research process. First, we analyze at which points of the research process and the experimental design manipulation (willingly or unwillingly) might occur.

In a behavioral experimental design that aims to apply deep learning analysis for tracking and analyzing animal movement and behavior, researchers record videos from their model animals under different conditions. These conditions and the order under which the different parts of the experiment are to be conducted, are *a priori* encoded in an experiment plan,

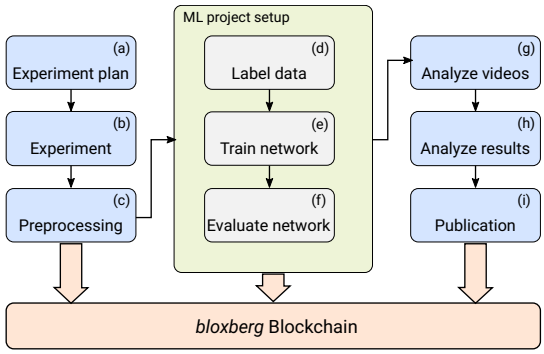

Fig. 1. ML supported animal research process. Potential and most likely feedback cycle is visualized by box, however, steps back can in theory appear at every stage of the process. In each pipeline step, several research artifacts are produced, e.g., in the step (g) videos resulting from the preprocessing step (c) and model from the training step (e) are used to generate time series data. The existence of all resulting artifacts can be passively or actively documented using the bloxberg blockchain infrastructure.

which can generally be considered a free form document. According to the experiment plan, the experiment is subsequently conducted. Note that already at this point in time, the experiment plan and the reality of the experiment might start to diverge. This by itself is not necessary a result of malicious behavior or bad practice, however, we can see a value in documenting this fact at the point in time, when the divergence initially occurs. Valid reasons for the differences can be outlined as part of the final publication.

Before starting the analysis, data cleanup and preprocessing, e.g., cutting the videos (start and the endpoint of the experiment), renaming the video files according to a predefined schema in order to encode additional information (animal id, test condition, session number, etc.) or normalization of physical factors, i.e., lighting or camera rotation, are performed. Since these steps are often highly individualized for the experiment at hand and might be executed either manually or by hand-rolled software solutions, errors might be easily introduced, that would propagate further through the research process. It is therefore advisable, to record the different states of the video files to allow for detailed reviews and fault analysis capabilities.

For the further analysis of the videos, one option to check the differences is comparing animals' pose alteration occurring in different conditions and the *DeepLabCut* deep learning software package has been widely used among researchers for markerless animal pose estimation [62]. The software is configured using a JSON configuration file, containing essential parameters such as which body parts need to be labelled, how many frames will be labelled and the number of targeted training iterations. Since the initial configuration will influence the further training of the model significantly, the integrity of the file is of high importance. The labelling of training data is a software aided manual process, that will generate 2 sets of image files: A train and a test data set. To make the performance and quality of the resulting model traceable, the used train data set needs to be as closely and auditable connected to the model as possible. Once the training has been conducted, the model is saved in a file, which also contains the number of used iterations for training encoded in the filename. Since the performance of the model is subsequently evaluated in a software aided manual process, an inadequate performance might lead to restarting from previous steps and making changes to parameters such as preprocessing, number of training iterations or adding addition training data. This could lead to the previous model and model parameters never being actively used in the further process, however, documenting its existence and performance at a specific point in time is still a valuable step towards better transparency.

A sufficiently trained network can be used for analyzing the rest of the experiment videos, creating *Hierarchical Data Format (HDF5)* files for each video, which consist of time series coordinate data for the labelled body points. In order to generate scientific results, this time series data can be evaluated by using different statistical methods depending on the research questions, either by using specialized statics software such as *SPSS*, or by writing custom code and using special software packages in languages such as *Python*, *R* or *Matlab*. To decrease the possibility of human error and improve reproducibility, approaches that are describable in machine executable languages, i.e. code, are favorable, especially since the existence of content of such files could be recorded similarly to the actual data. However, since the analysis code usually involves further software dependencies, a machine readable and reproducible description of the runtime environment as a combination of language specific software dependency management tools such as *Maven*, *pipenv* or *NPM* should be used in conjunction with container engines such as *Docker*.

IV. System Design

A. bloxberg

bloxberg is a scientific consortium blockchain that is governed and operated by a set of international research organizations [23]. This consortium operated blockchain network is primarily targeted towards applications and services that enhance research with the benefits of distributed ledger technology and is implemented as a public-permissioned PoA Ethereum network, with research organizations acting as validators. By ensuring that a diverse group of research organizations is governing the chain, collective standards for the scientific domain can be readily established. The name bloxberg has been inspired by the mythological name *Blocksberg*, which nowadays acts as a substitute for the geographical name of the German *Brocken* mountain. In the middle ages, the Blocksberg was considered a meeting place for witch covens and although historically different geographical locations have been attributed as being a Blocksberg (since it originally and etymologically describes the property of a meeting place for witches), over time the legends and stories merged into a single mythological location, which finally got attributed to the Brocken mountain in the 17th century [63].

B. Research Object Certification

The bloxberg network is operated and developed according to the *bloxberg Improvement Proposals (BLIP)* process, which itself is modeled after other community development and decision processes, such as Ethereum's EIP, Bitcoin's BIP or Python's PEP [64]. One particular topic is *Research Object Certification (ROC)* as described in BLIP-2, which proposes a standard for object certification which is generalizable such that it could be utilized for a variety of scientific fields. Each individual research object, representing a set of files, is formulated as a *JSON-LD* [65] file that complies with the *Verifiable Credentials Data Model 1.0* W3C standard [66]. A merkle proof [67] is calculated from each batch of files that are to be certified, which is then transacted on the *bloxberg* blockchain as an *ERC-721* token, extended by the *ERC721Metadata* interface [68]. Finally, the resulting transaction and merkle proof are encoded in the JSON-LD files. In addition, the ERC-721 token can be updated with a *RFC 3986 URI* specifying the resource under which the JSON-LD file can be accessed, to improve the self-describing properties of the token itself. Additional metadata can be provided, which describes the object in more detail or might contain URIs resolving to the underlying files. Apart from the machine-readable JSON-LD format, more user-friendly formats such as HTML or PDF are used in practice to present the certificate to the user. These add mark-up to the information stored in JSON-LD and are by no means the complete and authoritative version of the certificate, but merely accompany the JSON-LD file as a human readable presentation. The issued certificates can be checked for validity using a verification service developed and deployed as a DApp. Using the generated proof value for the batch of research objects a hexadecimal value can be decoded that needs to match against a transaction id on the *bloxberg* chain. Subsequently the ERC-721 token hash stored for the respective transaction id is verified as well, by matching it against the *crid* value contained in the local certificate. Note that since the *Research Object Certification* process is oblivious to file format, file sets can consist of data as well as of code and execution environment descriptions, while also allows for a heterogeneous combination of those classes.

C. Research Object Chain Links

While ROC provides a way to certify the existence of a certain set of files at a specific point in time using a scientific blockchain infrastructure, by itself they provide no way of semantic or logical clustering of certificates. We therefore propose an additional layer, that allows to create a chronological sequence of certificates, to better capture the realities of scientific processes as described in III, *Research Object Chain Links (ROCL)*. On this layer, each ROCL contains the Ethereum transaction of the previous ROCL, thereby strongly chaining them together in way akin to the blockchain data structure itself (see Fig. 2). To further support computational reproducibility and discoverability of data and accompanying code, each ROCL contains two distinct ROC certificates (one for data, one for code and runtime environment description)

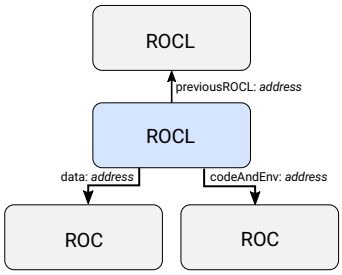

Fig. 2. Research Object Chain Links (ROCL) data structure

in the form of Ethereum transaction addresses. Researchers are advised to regularly append ROCL entries to the chain representing their current research project, to capture progress with high fidelity. A regular schedule, similar to the daily usage of a lab notebook, would be recommended, however, it is not enforced in this technical specification.

V. Conclusion

An envisioned upgrade to the certification process consists of using *Decentralized Identifiers (DIDs)* [69] for the involved actors, to further strengthen the system as a trust architecture. The DIDs will essentially identify both the issuing instance and the owner of the certificate. Issued certificates can subsequently be added to its owner's DID profile that, in turn, can be discovered via a service endpoint linked in the corresponding DID document. Hence the list of obtained certificates can be publicly advertised similarly to the list of authored papers. This upgraded version of the issued certificates complies with the *Verifiable Credentials Data Model 1.0* [66], for which multiple use cases have been designed [70], including but not limited to educational purposes. Although the current design does not contain any specification with regards to the actual data storage, future iterations might include recommendations and specification, to further improve transparency and collaboration. It would be advisable to closely analyze if decentralized storage solution such as *IPFS* [71] or *Swarm* [72] are suitable with regards to non-functional requirements such as performance and scalability. In addition, future work will aim at performance and cost evaluation of the proposed blockchain-based approach. Obviously, certifying the existence of a certain set of files and the generation of chronological sequences of such certificates is only one step in a set of frameworks containing solutions for reproducibility and authenticity leading to automated verification of the whole research process. Thus, there is still a need for further research on how to exploit the proposed blockchain-based solution in actual scientific workflows, how to use the infrastructure for workflow execution and verification in order to guarantee the consistency with scientific, institutional and performance requirements, and how e.g., formal verification techniques can be applied for achieving these goals. Although our approach does not consider semantics or any quality aspects, the proposed system still contains the potential to improve the transparency, integrity and reproducibility of research.

References

[1] O. S. Collaboration, "Estimating the reproducibility of psychological science," *Science*, vol. 349, no. 6251, 2015. [Online]. Available: https://science.sciencemag.org/content/349/6251/aac4716

[2] W. Świątkowski and B. Dompnier, "Replicability Crisis in Social Psychology: Looking at the Past to Find New Pathways for the Future," *International Review of Social Psychology*, vol. 30, no. 1, pp. 111–124, May 2017. [Online]. Available: http://www.rips-irsp.com/articles/10.5334/irsp.66/

[3] W. M. Hensel, "Double trouble? The communication dimension of the reproducibility crisis in experimental psychology and neuroscience," *European Journal for Philosophy of Science*, vol. 10, no. 3, p. 44, Oct. 2020. [Online]. Available: https://doi.org/10.1007/s13194-020-00317-6

[4] M. Pawlik, T. Hütter, D. Kocher, W. Mann, and N. Augsten, "A Link is not Enough – Reproducibility of Data," *Datenbank-Spektrum*, vol. 19, no. 2, pp. 107–115, Jul. 2019. [Online]. Available: https://doi.org/10.1007/s13222-019-00317-8

[5] A. Zuiderwijk and H. Spiers, "Sharing and re-using open data: A case study of motivations in astrophysics," *International Journal of Information Management*, vol. 49, pp. 228–241, Dec. 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0268401218311836

[6] T. Miyakawa, "No raw data, no science: another possible source of the reproducibility crisis," *Molecular Brain*, vol. 13, no. 1, p. 24, Feb. 2020. [Online]. Available: https://doi.org/10.1186/s13041-020-0552-2

[7] G. D. Smith and S. Ebrahim, "Data dredging, bias, or confounding," *BMJ : British Medical Journal*, vol. 325, no. 7378, pp. 1437–1438, Dec. 2002. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1124898/

[8] K. Dwan, D. G. Altman, J. A. Arnaiz, J. Bloom, A.-W. Chan, E. Cronin, E. Decullier, P. J. Easterbrook, E. V. Elm, C. Gamble, D. Ghersi, J. P. A. Ioannidis, J. Simes, and P. R. Williamson, "Systematic Review of the Empirical Evidence of Study Publication Bias and Outcome Reporting Bias," *PLOS ONE*, vol. 3, no. 8, p. e3081, Aug. 2008, publisher: Public Library of Science. [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0003081

[9] G. L. Gadbury and D. B. Allison, "Inappropriate Fiddling with Statistical Analyses to Obtain a Desirable P-value: Tests to Detect its Presence in Published Literature," *PLOS ONE*, vol. 7, no. 10, p. e46363, Oct. 2012, publisher: Public Library of Science. [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0046363

[10] M. J. Page, J. E. McKenzie, J. Kirkham, K. Dwan, S. Kramer, S. Green, and A. Forbes, "Bias due to selective inclusion and reporting of outcomes and analyses in systematic reviews of randomised trials of healthcare interventions," *The Cochrane Database of Systematic Reviews*, no. 10, p. MR000035, Oct. 2014.

[11] M. L. Head, L. Holman, R. Lanfear, A. T. Kahn, and M. D. Jennions, "The Extent and Consequences of P-Hacking in Science," *PLOS Biology*, vol. 13, no. 3, p. e1002106, Mar. 2015, publisher: Public Library of Science. [Online]. Available: https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1002106

[12] L. P. Freedman, I. M. Cockburn, and T. S. Simcoe, "The Economics of Reproducibility in Preclinical Research," *PLOS Biology*, vol. 13, no. 6, p. e1002165, Jun. 2015, publisher: Public Library of Science. [Online]. Available: https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1002165

[13] S.-A. Santone, P. McQuilton, P. Rocca-Serra, A. Gonzalez-Beltran, M. Izzo, A. L. Lister, M. Thurston, and the FAIRsharing Community, "Fairsharing as a community approach to standards, repositories and policies," *Nature Biotechnology*, vol. 37, pp. 358–367, 2019.

[14] A. A. Siyal, A. Z. Junejo, M. Zawish, K. Ahmed, A. Khalil, and G. Soursou, "Applications of blockchain technology in medicine and healthcare: Challenges and future perspectives," *Cryptography*, vol. 3, no. 1, 2019. [Online]. Available: https://www.mdpi.com/2410-387X/3/1/3

[15] A. K. Shrestha and J. Vassileva, "User data sharing frameworks: A blockchain-based incentive solution," in *2019 IEEE 10th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, 2019, pp. 0360–0366.

[16] A. Anderberg, E. Andonova, M. Bellia, L. Calès, A. Inamorato Dos Santos, I. Kounelis, I. Nai Fovino, M. Petracco Giudici, E. Papanagiotou, M. Sobolewski, F. Rossetti, and L. Spirito, "Blockchain now and tomorrow," European Commission, Scientific and Technical Research Reports EUR 29813 EN, 09 2019. [Online]. Available: http://dx.doi.org/10.2760/901029

[17] V. L. Lemieux, "Trusting records: Is blockchain technology the answer?" *Records Management Journal*, vol. 2, pp. 110–139, 2016.

[18] S. Nakamoto, "Bitcoin: A peer-to-peer electronic cash system," Web document., 2008. [Online]. Available: https://bitcoin.org/bitcoin.pdf

[19] V. Buterin, "Ethereum: A next-generation smart contract and decentralized application platform." white paper, 2014. [Online]. Available: http://ethereum.org/ethereum.html

[20] M. Sharples and J. Domingue, "The blockchain and kudos: A distributed system for educational record, reputation and reward," *Adaptive and Adaptable Learning, EC-TEL 2016*, pp. 490–496, 2016.

[21] K. Wittek, D. Krakau, N. Wittek, J. Lawton, and N. Pohlmann, "Integrating bloxberg's proof of existence service with matlab," *Frontiers in Blockchain*, vol. 3, p. 49, 2020. [Online]. Available: https://www.frontiersin.org/article/10.3389/fbloc.2020.546264

[22] E. Duffield and D. Diaz, "Dash: A privacy centric cryptocurrency," 2015.

[23] S. Vengadasalam, F. Kleinfercher, and J. Lawton, "First international blockchain for science: bloxberg," white paper, 2019. [Online]. Available: https://www.mpg.de/13413856/erste-internationale-blockchain-fuer-die-wissenschaft-bloxberg

[24] H. Kaur, M. Alam, R. Jameel, A. Mourya, and V. Chang, "A proposed solution and future direction for blockchain-based heterogeneous medicare data in cloud environment," *Journal of Medical Systems*, vol. 42, Jul. 2018.

[25] I. Barclay, S. Radha, A. Preece, I. Taylor, and J. Nabrzyski, "Certifying provenance of scientific datasets with self-sovereign identity and verifiable credentials," 2020.

[26] P. Wortner, M. Schubotz, C. Breitinger, S. Leible, and B. Gipp, "Securing the integrity of time series data in open science projects using blockchain-based trusted timestamping," in *Proceedings of the Workshop on Web Archiving and Digital Libraries (WADL'19)*, 2019, pp. 1–3.

[27] ——, "A decentralized method for making sensor measurements tamper-proof to support open science applications," *arXiv preprint arXiv:1904.00237*, 2019.

[28] J. Erbguth and J.-H. Morin, "Towards distributed trustworthy traceability and accountability," *Earticle. net*, pp. 223–225, 2016.

[29] M. Aturban, S. Alam, M. Nelson, and M. Weigle, "Archive assisted archival fixity verification framework," in *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*. IEEE, 2019, pp. 162–171.

[30] T. E. Hardwicke, M. B. Mathur, K. MacDonald, G. Nilsonne, G. C. Banks, M. C. Kidwell, A. Hofelich Mohr, E. Clayton, E. J. Yoon, M. Henry Tessler, R. L. Lenne, S. Altman, B. Long, and M. C. Frank, "Data availability, reusability, and analytic reproducibility: evaluating the impact of a mandatory open data policy at the journal Cognition," *Royal Society Open Science*, vol. 5, no. 8, p. 180448, 2018. [Online]. Available: https://royalsocietypublishing.org/doi/10.1098/rsos.180448

[31] I. Hrynaszkiewicz, *Publishers' Responsibilities in Promoting Data Quality and Reproducibility*, ser. Handbook of Experimental Pharmacology. Cham: Springer International Publishing, 2020, pp. 319–348. [Online]. Available: https://doi.org/10.1007/164{_}2019{_}290

[32] K. Ram, "Git can facilitate greater reproducibility and increased transparency in science," *Source Code for Biology and Medicine*, vol. 8, no. 1, p. 7, Feb. 2013. [Online]. Available: https://doi.org/10.1186/1751-0473-8-7

[33] N. Strupler and T. C. Wilkinson, "Reproducibility in the Field: Transparency, Version Control and Collaboration on the Project Panormos Survey," *Open Archaeology*, vol. 3, no. 1, pp. 279–304, Nov. 2017. [Online]. Available: https://www.degruyter.com/view/journals/opar/3/1/article-p279.xml

[34] J. D. Blischak, P. Carbonetto, and M. Stephens, "Creating and sharing reproducible research code the workflowr way," *F1000Research*, vol. 8, p. 1749, Oct. 2019. [Online]. Available: https://f1000research.com/articles/8-1749/v1

[35] "Code Ocean," http://codeocean.com, accessed: 2020-12-14.

[36] "protocols.io," http://protocols.io, accessed: 2020-12-14.

[37] F. Strozzi, R. Janssen, R. Wurmus, M. R. Crusoe, G. Githinji, P. Di Tommaso, D. Belhachemi, S. Möller, G. Smant, J. de Ligt, and P. Prins, "Scalable Workflows and Reproducible Data Analysis for Genomics," in *Evolutionary Genomics: Statistical and Computational Methods*, ser. Methods in Molecular Biology, M. Anisimova, Ed. New York, NY: Springer, 2019, pp. 723–745. [Online]. Available: https://doi.org/10.1007/978-1-4939-9074-0{_}24

[38] A. Papanikolaou, S. Harries, P. Hooijmans, J. Marzi, R. Le Néna, S. Torben, A. Yrjänäinen, and B. Boden, "A holistic approach to ship design: Tools and applications," *Journal of Ship Research*, pp. 1–29, 2020.

[39] T. Wunderlich, M. Abu-Zurayk, Časlav. Ilić, J. Jepsen, M. Schulze, M. Leitner, A. Schuster, S. Dähne, M. Petsch, R.-G. Becker, S.-A. Zur, and S. Gottfried, "Overview of collaborative high performance computing-based MDO of transport aircraft in the DLR project VicToria," *Deutscher Luft-und Raumfahrtkongress, Friedrichshafen, Germany*, 2018.

[40] P. Weiand, M. Buchwald, and D. Schwinn, "Process development for integrated and distributed rotorcraft design," *Aerospace*, vol. 6, no. 2, p. 23, 2019.

[41] P. M. Fischer, M. Deshmukh, A. Koch, R. Mischke, A. Martelo Gomez, A. Schreiber, and A. Gerndt, "Enabling a conceptual data model and workflow integration environment for concurrent launch vehicle analysis," in *Proceedings of the International Astronautical Congress, IAC*, 2018.

[42] B. Boden, J. Flink, R. Mischke, K. Schaffert, A. Weinert, A. Wohlan, and A. Schreiber, "RCE: an integration environment for engineering and science," *CoRR*, vol. abs/1908.03461, 2019. [Online]. Available: http://arxiv.org/abs/1908.03461

[43] S. Druskat, "Software and Dependencies in Research Citation Graphs," *Computing in Science & Engineering*, vol. 22, no. 2, pp. 8–21, 2019.

[44] M. Stoffers, "Trustworthy provenance recording using a blockchain-like database," Master's thesis, Leipzig University, 2017.

[45] R. Qasha, Z. Wen, J. Cala, and P. Watson, "Sharing and performance optimization of reproducible workflows in the cloud," *Future Gener. Comput. Syst.*, vol. 98, pp. 487–502, 2019. [Online]. Available: https://doi.org/10.1016/j.future.2019.03.045

[46] M. H. Hilman, M. A. Rodriguez, and R. Buyya, "Multiple workflows scheduling in multi-tenant distributed systems: A taxonomy and future directions," *ACM Comput. Surv.*, vol. 53, no. 1, Feb. 2020. [Online]. Available: https://doi.org/10.1145/3368036

[47] J. M. Erbel, S. Wittek, J. Grabowski, and A. Rausch, "Dynamic Management of Multi-Level-Simulation Workflows in the Cloud," in *Proceedings of the 2nd International Workshop on Simulation Science (SimScience 2019)*, ser. Communications in Computer and Information Science (CCIS). Springer, Cham, 2019.

[48] Y. D. Dessalk, N. Nikolov, M. Matskin, A. Soylu, and D. Roman, "Scalable execution of big data workflows using software containers," in *Proceedings of the 12th International Conference on Management of Digital EcoSystems*, ser. MEDES '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 76–83. [Online]. Available: https://doi.org/10.1145/3415958.3433082

[49] M. Orzechowski, B. Baliś, R. G. Słota, and J. Kitowski, "Reproducibility of computational experiments on kubernetes-managed container clouds with hyperflow," in *Computational Science – ICCS 2020*, V. V. Krzhizhanovskaya, G. Závodszky, M. H. Lees, J. J. Dongarra, P. M. A. Sloot, S. Brissos, and J. Teixeira, Eds. Cham: Springer International Publishing, 2020, p. 220–233.

[50] K. Burkat, M. Pawlik, B. Balis, M. Malawski, K. Vahi, M. Rynge, R. Ferreira da Silva, and E. Deelman, "Serverless Containers – rising viable approach to Scientific Workflows," *arXiv e-prints*, p. arXiv:2010.11320, Oct. 2020.

[51] W. Viriyasitavat, L. D. Xu, and Z. Bi, "Specification patterns of service-based applications using blockchain technology," *IEEE Transactions on Computational Social Systems*, vol. 7, no. 4, pp. 886–896, 2020.

[52] R. Coelho, R. Braga, J. M. David, M. Dantas, V. Stroele, and F. Campos, "Integrating blockchain for data sharing and collaboration support in scientific ecosystem platform," in *Proceedings of the 54th Hawaii International Conference on System Sciences*, 2021.

[53] R. Coelho, R. Braga, J. M. N. David, M. Dantas, V. Ströele, and F. Campos, "Blockchain for reliability in collaborative scientific workflows on cloud platforms," in *2020 IEEE Symposium on Computers and Communications (ISCC)*, 2020, pp. 1–7.

[54] Q. Meng and R. Sun, "Towards secure and efficient scientific research project management using consortium blockchain," *Journal of Signal Processing Systems*, vol. 93, pp. 1–10, 2021.

[55] X.-F. Zhang, "Application of blockchain technology in data management of university scientific research," in *International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing*, 2020, pp. 606–613.

[56] W. Jeng, S.-H. Wang, H.-W. Chen, P.-W. Huang, Y.-J. Chen, and H.-C. Hsiao, "A decentralized framework for cultivating research lifecycle transparency." *PLOS ONE*, vol. 15, no. 11, pp. 1–17, 2020.

[57] D. Fernando, S. Kulshrestha, J. D. Herath, N. Mahadik, Y. Ma, C. Bai, P. Yang, G. Yan, and S. Lu, "Sciblock: A blockchain-based tamper-proof non-repudiable storage for scientific workflow provenance," in *2019 IEEE 5th International Conference on Collaboration and Internet Computing (CIC)*, 2019, pp. 81–90.

[58] T. Mackey, N. Shah, K. Miyachi, J. Short, and K. Clauson, "A framework proposal for blockchain-based scientific publishing using shared governance," *Frontiers in Blockchain*, vol. 2, p. 19, 2019.

[59] D. Science and J. V. Rossum, "Blockchain for Research," Digital Science, Tech. Rep., 2017. [Online]. Available: https://digitalscience.figshare.com/articles/report/Blockchain{_}for{_}Research/5607778

[60] C. Furlanello, M. De Domenico, G. Jurman, and N. Bussola, "Towards a scientific blockchain framework for reproducible data analysis," *arXiv:1707.06552 [cs, q-bio]*, Jul. 2017. [Online]. Available: http://arxiv.org/abs/1707.06552

[61] S. Bartling, "Blockchain for Science and Knowledge Creation," in *Gesundheit digital: Perspektiven zur Digitalisierung im Gesundheitswesen*, R. Haring, Ed. Berlin, Heidelberg: Springer, 2019, pp. 159–180. [Online]. Available: https://doi.org/10.1007/978-3-662-57611-3_10

[62] A. Mathis, P. Mamidanna, K. M. Cury, T. Abe, V. N. Murthy, M. W. Mathis, and M. Bethge, "DeepLabCut: markerless pose estimation of user-defined body parts with deep learning," *Nature Neuroscience*, vol. 21, no. 9, pp. 1281–1289, Sep. 2018, number: 9 Publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s41593-018-0209-y

[63] J. Praetorius, *Blockes-Berges Verrichtung*, 1st ed. Leipzig; Frankfurt (Main): Scheibe; Arnst, 1668.

[64] J. Lawton, A. Ionita, A. Tenorio-Fornés, K. Wittek, and K. Uzdogan, "bloxberg-org/blips: Zenodo doi release," Dec. 2020. [Online]. Available: https://doi.org/10.5281/zenodo.4327465

[65] M. Sporny, D. Longley, G. Kellogg, M. Lanthaler, P.-A. Champin, and N. Lindström, "Json-ld 1.1–a json-based serialization for linked data," Ph.D. dissertation, W3C, 2019.

[66] M. Sporny, D. Longley, and D. Chadwick, "Verifiable credentials data model 1.0," Nov 2019. [Online]. Available: https://www.w3.org/TR/2019/REC-vc-data-model-20191119/

[67] K. H. Duffy and D. Semenovskiy. [Online]. Available: https://w3c-ccg.github.io/lds-merkle-proof-2019/

[68] D. S. William Entriken, "Eip-721: Erc-721 non-fungible token standard," Jan 2018. [Online]. Available: https://eips.ethereum.org/EIPS/eip-721

[69] "Decentralized Identifiers (DIDs) v1.0." [Online]. Available: https://www.w3.org/TR/did-core/

[70] "Verifiable Credentials Use Cases." [Online]. Available: https://www.w3.org/TR/vc-use-cases/

[71] J. Benet, "Ipfs-content addressed, versioned, p2p file system," *arXiv preprint arXiv:1407.3561*, 2014. [Online]. Available: https://arxiv.org/abs/1407.3561

[72] J. H. Hartman, I. Murdock, and T. Spalink, "The swarm scalable storage system," in *Proceedings. 19th IEEE International Conference on Distributed Computing Systems (Cat. No.99CB37003)*, 1999, pp. 74–81.