STI 2018 Leiden

23rd International Conference on Science and Technology Indicators
"Science, Technology and Innovation Indicators in Transition"

STI 2018 Conference Proceedings

Proceedings of the 23rd International Conference on Science and Technology Indicators

Chair of the Conference

Paul Wouters

Scientific Editors

Rodrigo Costas
Thomas Franssen
Alfredo Yegros-Yegros

Layout

Andrea Reyes Elizondo
Suze van der Luijt-Jansen

The articles of this collection can be accessed at https://hdl.handle.net/1887/64521

ISBN: 978-90-9031204-0

23rd International Conference on Science and Technology Indicators (STI 2018)
"Science, Technology and Innovation indicators in transition"
12 - 14 September 2018 | Leiden, The Netherlands
#STI18LDN

Applying Blockchain Solutions to Address Research Reproducibility and Enable Scientometric Analysis

David Kochalko*, Courtney Morris**, Jason Rollins***

*dkochalko@artifacts.ai, **cmorris@artifacts.ai, ***jrollins@artifacts.ai
ARTiFACTS, 955 Massachusetts Ave. #184, Cambridge, MA, 02139, USA

Blockchain hash: 0xa1e251a666adb45d292f6b0c46339929d5098ae5f81f934574e2ab9a4560f8ed

Introduction

Awareness of the unreliability of many published research findings has been percolating within the public and academic consciousness for decades; many have called the current situation a "reproducibility crisis" (Baker, 2016). There are high-profile examples from myriad scientific disciplines. Even casual readers of science-related news are certainly aware of the cold fusion debate raging since the 1980's (Ioannidis, 2005) and the scandal over fabricated human stem cell data underlying the seminal papers by Hwang, et al. published in (and subsequently retracted from) *Science* (Kennedy, 2006). The societal impacts of this crisis are far-reaching and include undermining the public trust, drug and food safety, and misuse of scarce grant funding resources. Empirical research on the extent of reproducibility-related challenges and their contributing factors (Goodman, Fanelli, Ioannidis, 2016; Munafo et. al., 2017) point to behavioral, economic, governance, and technological causes (Nosek, Spies, Motyl, 2012). More specifically within the bibliometrics community, recent scrutiny has focused on details of methodology, interpretation of data, and broader implications of study results; Van den Besselaar's (2017) close analysis of Butler's (2003) claims about funding and Australian research is one recent and representative example. As well, there are burgeoning efforts to define taxonomies for Direct and Conceptual reproducibility for bibliometrics (Waltman, et. al., 2018; Velden, et. al., 2018). With quantitative computation at the heart of bibliometric analysis, looking to emerging technologies to help mitigate reproducibility challenges seems a natural choice and will likely be essential for affecting lasting change. Specifically, Web 3.0 solutions such as artificial intelligence and distributed computing enable 3 key improvements for science and scientometrics: 1) greater visibility of findings and researchers, in near real-time; 2) provenance, with clearer digital trails, stronger and more confident signal intelligence; and 3) a community-curated citation index that will augment the Open Citation Initiative (http://opencitations.net/) and provide a rich resource for the bibliometrics community. ARTiFACTS.ai, a platform developed by the authors of this paper, is currently the only system purpose-built to achieve these aims.

Provenance and Attribution of All Research Outputs

Establishing and sustaining the complete provenance of research outputs has always been a laudable goal but is only becoming practical with current technical solutions. Scientists and scholars use many products and tools that create digital outputs. These artifacts are revised, new ones created as the research process continues, and outputs build. It is well established that researchers withhold information for fear of losing control of their intellectual property (Campbell, Clarridge, Gokhale, 2002). Therefore, most research outputs reside in disconnected silos, available only to researchers in the project or shared on a limited basis. Version controls that chain together revisions are missing from (or inconsistently used in) many of these tools. Common file storage systems neither connect and associate different file types, nor do they augment these files with relevant metadata valuable for attribution and scientometric analysis (Garfield, 1979). An emerging solution to this long-standing challenge is the ability to establish digital proof of existence (PoE) and provenance over research outputs that is registered into a secure data structure for enabling verification of research. When augmented with methods of data capture from device-to-blockchain, confirming the identity of creators, and linking creators with their research outputs, it becomes possible both to confirm proof of authorship and the unaltered authenticity of source data. This ability to clearly record research data, methods, results, and interpretations in a tamper-evident, and traceable way is the specific aspect of blockchain technology that some feel most applicable to improving reproducibility. Bartling, et. al., (2017) explicitly state "…by opening the research cycle to scientific self-control beyond the final publication … might therefore be a fix to the current reproducibility crisis in science." Furlanello (2017) echoes this with "...blockchain technology can be key to address the issues of replicability, accountability and trust in scientific studies, providing an immutable ledger for all the steps, from protocols to all outcomes…"

It is important to highlight the unintended effects of a scholarly communication system that is dominated, nearly exclusively, by edited and peer-reviewed, published works. This is not to assert those processes should be abandoned. Rather, we reference them to highlight the lack of attention, investment, and novel solutions for revealing the precursors of the published work, the meaningful outputs generated during research, most of which become dark knowledge and remain unavailable for subsequent use, verification, or reproducibility studies. Web 3.0 technologies leveraged into flexible, workflow-oriented systems that connect all related artifacts are now becoming achievable and a practical option for addressing reproducibility challenges. At the same time, the Open Access and Open Science movements have helped to voice general concerns of accessibility of data and publications—challenges that might also reasonably be addressed with similar technical solutions.

While conventional citations provide general pointers to prior science, these guides to earlier works appear long after the current research itself has been completed and typically years from when the cited work occurred. Due to the protracted publication process, that same 'current' research will then itself typically go years before it becomes cited further impeding the natural iterative process of research. Compounding this delay, these citations typically only occur between published articles and thus fail to point to and index vast amounts of additional and valuable research outputs. The ability to cite a wider range of output has been emerging for a decade but these efforts have myopically focused on specific types such as datasets (Borgman, 2015) or pre-prints which, while a step in the right direction, are still manuscript focused. Overall, citations and their information quotient have been slow to evolve. Source records lack information about the specific contributions of the authors or creators, and citations given often provide no reason or rationale. With creative approaches and contemporary computer technology, attributions to be given, received, and recorded for all types of pre-published research output is now becoming practical. Enabling attributions to be given to all types of creative works, together with additional context that are securely registered into a distributed database in real-time and without publishing delay, will improve verification analyses and provide new insight into citation activity, while also allowing research reputation to grow in real-time as research is being conducted.

Leveraging Web 3.0 Technologies

The infrastructure, building blocks, and applications necessary for realizing the benefits of the Web 3.0 era in science and scholarship are apparent and expanding. And while no single technology can be expected to address all needs, the collective and creative use of many Web 3.0 innovations together with conventional solutions will bring measurable improvements to reproducibility in science and scientometrics.

In just the past couple of years there has been a virtual explosion of interest and development in artificial intelligence, big data, machine learning, and blockchain technology for scientific research and academic publishing. There is general enthusiasm (tempered with justified skepticism over the hype) about "…the potential...to transform scholarly communication and research in general...blockchain can touch many critical aspects…including transparency, trust, reproducibility and credit…" (Van Rossum, 2017). Industry-academic partnerships and start-up companies, including Blockchain for Peer Review, Frankl, Iris.ai, Knowbella, Orvium, Protocols.io, Scienceroot, and others are beginning to build Web 3.0 applications for peer review, open access, data sharing, and other core scholarly communication uses cases. Some of these efforts are still conceptual or under development and seem generally intent on replacing existing processes with a direct, distributed web alternative, so are not necessarily focused on improving reproducibility. ARTiFACTS.ai, a live platform developed by the authors of this paper, is arguably the most mature of these new systems and is purpose-built to address many of the challenges noted above. ARTiFACTS.ai comprises four key components:

- A Hyperledger-based, permissioned blockchain engine powering existence (PoEs) and attribution (citation) transactions to be permanently recorded, verified, versioned, stored, and linked.
- A web-based project management and collaboration platform based on the Open Science Framework (Center for Open Science, 2018), designed to empower distributed research teams to work together and integrate with their existing productivity and workflow software tools.
- A set of plug-ins and APIs for direct integration with existing research-focused workflow, analysis, and storage software allowing streamlined connections to the ARTiFACTS blockchain.
- A comprehensive, and ever-growing, metadata archive of scholarly artifacts (articles, datasets, research methods, in-progress reports, etc.) comprising a prospective index created in real-time and joined with a retrospective index that is curated, enhanced, and governed

by the research community—who are also incentivized and directly rewarded for their contributions to this work, including ultimate majority ownership in the historical index.

Through its prospective indexing, ARTiFACTS empowers researchers at any point from inception to completion to organize and manage their diverse research products, transacting PoEs and citations systematically into a publicly accessible distributed ledger that creates an immutable, persistent, and reliable mechanism for accessing version changes and all related artifact types, essential for confirming reproducibility of their science. Importantly, this information is accessible in real-time while a research effort is underway, during peer review, or thereafter for stakeholders concerned with the reliability of scientific findings. Scientometricians will have immediate access to the underlying citation data that enables control groups to be established for purposes of multiple analyses and reproducibility studies.

While ARTiFACTS.ai invokes a mix of technologies, the measured skepticism around blockchain in some technical and academic writing compels additional explanation to warrant its relevance. We incorporate blockchain into our solution based on the essential characteristics of the ARTiFACTS vision which make blockchain a necessary (though not singularly sufficient) component of our technology :

- We are creating an open and shared resource, one that will receive inputs from a community of researchers.
- The community brings multiple creators, contributors, editors, all of whom are widely distributed physically and not known or inherently trusted uniformly.
- The transaction activity will entail some degree of challenges, corrections, and confirmations, where collaborative interactions among multiple contributors are managed.
- Incentives in the form of tokens provide personal connection to the quality of the retrospective index and establish governance influence over policies and development innovations.

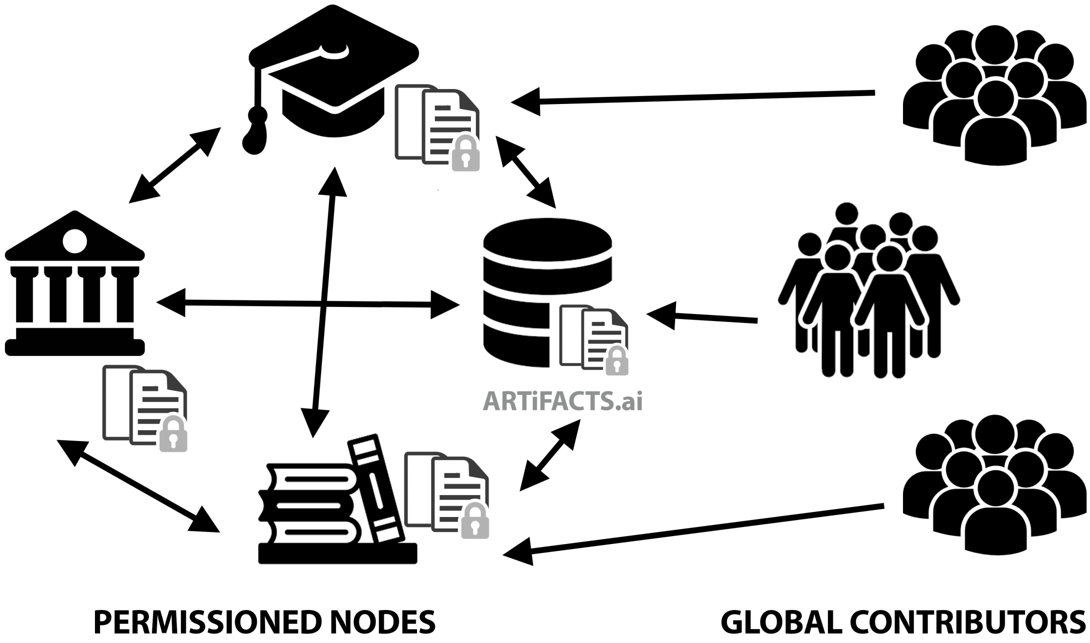
Distributing trust beyond a single, dominant organization by employing permissioned nodes brings together representative stakeholders, all of whom share core values for upholding scientific inquiry, and make the system economically viable by significantly reducing operating costs.

Enabling Next-Generation Scientometric Analysis

Scopus and Web of Science are commercial citation datasets that have dominated much of scientometric analysis for the past several decades. Despite their dominance, they offer a limited subset of traditionally published, and deeply retrospective, research outputs that are indexed and linked based on the limitations of the companies's economics and less so on the purported interest of selectivity and curatorial control. Yet research has not supported the long-term value of these constraining editorial policies, artificially created metrics, nor their reliance on a narrow view of traditional citations within their walled gardens (Wang, Song, Barabási, 2013). In contrast, a more expansive view—powered by Web 3.0 technologies and broad community engagement—will readily enable conventional citations to be augmented with new, and useful information. Source records may contain context about the creators; attributions will contain information explaining the reasons for acknowledging prior works. New forms of attribution and scientometric analysis will emerge (Moed, 2009) including for pre-published, and traditionally non-published artifacts, and relating the quality of supporting evidence to citation patterns and measures of scholarly influence. Research validation and scientometric studies may be conducted on more comprehensive and connected sets of evidence, in real-time rather than retrospectively.

Blockchain technology allows for instantly accessible, time efficient solutions for scientific collaboration. While no system is completely immune to errors, a more open infrastructure that enables access and reuse of research objects throughout the scientific process will very likely result in more transparency. The "trustless" and "anonymous" nature of some blockchain-based systems are often touted as core benefits, but there are downsides that could be mitigated with a system typology leveraging the best of the distributed aspect of blockchains combined with elements of established authority and control. A permissioned blockchain model does just this by allowing some well-known authorities (universities, charitable foundations, research consortia, scholarly societies, etc.) to operate "trusted nodes" to help balance the completely open and distributed blockchain data structure. Such a design offers clear advantages for data quality and, ultimately, reproducibility, while also minimizing costs.

Figure 1. Network diagram



We envision an organization like ORCiD (https://orcid.org/) operating a trusted node where they could validate identities and associate them with ORCiD ID's, collaborating with ARTiFACTS and the Sovrin Foundation (https://sovrin.org/) in leveraging a decentralized digital identity network together with ORCiD's user information and enriched metadata (based on user choice and permissions), which can be linked with the community-managed citation dataset. Their node would not need to be the ultimate authority through which all transactions pass, but instead could help to create a more complete and traceable chain of links, objects, and associations. We postulate that more frequent transacting of granular research artifacts on such a permissioned system will encourage more forthright and principled behavior as the progress trail will be easier to follow, validate, and identify any issues—accidental or fraudulent. Our long-term view is to enable wider sharing of methods, experimental notes, software code, and protocols that could all help with reproducibility challenges. Our vision also includes new individual and cumulative metrics and indicators built on the community-managed citation corpus, validated by the impartial partners within the trusted ecosystem, and available for all stakeholders to reproduce and examine at any time. Perhaps most importantly, ARTiFACTS provides a more holistic view of the contributions and impact scientists and scholars are making in their fields. The scientometrics community is ideally positioned to provide guidance on the most suitable metrics and indicators for assessing researcher impact.

Economic and Reputational Incentives in an Evolving Research Ecosystem

For scientific communication to meaningfully evolve, the underlying reputation and recognition systems must adapt. New incentive systems are also needed, but only where their application is appropriate. A researcher's reputation, in theory, reflects their full contribution to their field of study and, by extension, the attribution they receive for those contributions. As discussed, ARTiFACTS creates the ability to provide, receive, and, most significantly, record attribution across all research outputs, not just a limited index of published articles. With this approach, theory can increasingly move closer to real world application.

But what is the role for new incentive systems made possible by distributed computing and blockchain, that allow for incentives in the form of value tokens? The ARTiFACTS.ai platform incorporates tokens as a central component to compensate the community for curation of the retrospective index—an appropriate application that importantly does not invoke their use for influencing or determining academic reputation. ARTiFACTS.ai will use tokens to incent community contribution to creating the retrospective index, establishing a personal connection to the quality of the collective citation data, providing governance over policies and development of the corpus, and ultimately achieving community ownership of the retrospective index. This token structure will also enable the community to allocate how surplus proceeds are awarded toward research proposals, including bibliometric research. These tokens will be approved by, and compliant with, United States Security and Exchange Commission (SEC) regulations, in distinct contrast to so-called cryptocurrencies. Some of the organizations mentioned above are already extending this idea of a distributed 'web of value' to other core research activities, like peer reviews, to formally quantify contributions that have traditionally been expected but often seen as secondary to publishing. This approach is quite new but is supported by some emerging research (see Shrestha, & Vassileva, 2018 for one recent discussion). Furlanello (2017) notes "...the blockchain...can be used to build the reputation of researchers...and, overall, the reproducibility...embeds a direct, transparent, objective and unbiased...reputation score...to be a valid alternative to...other bibliometrics measures widely adopted for ranking…" Key challenges currently facing many bibliometric researchers (along with many in other disciplines) certainly include: access to data—especially large-scale, curated datasets, open and fair licensing to share data, clear and reusable methods and code, and interpretations and claims tied to actual experimental results. Based on our current work, and the burgeoning research of others, we are confident that novel new approaches leveraging Web 3.0 technologies, including the new systems noted above, could readily solve these inherent issues.

Conclusions and Future Study

Reproducibility challenges persist in science. Empirical studies expose many causal factors. Application of blockchain-based solutions, including the ARTiFACTS system we have deployed, will strengthen efforts to verify research findings and enable new forms of scientometric analysis. The key factors that lend themselves to improving reproducibility through contemporary technical solutions are: provenance of a broader range of research work products, enabling unconstrained attribution, unencumbered access to datasets, and empowerment of researchers to share in-progress work more confidently and freely. Scientometric analysis can be applied beyond the artificial boundaries of published, indexed literature in new and novel ways to provide insight into research paths and their influence on advancing knowledge and understanding.

As some of our discussion in the paper has been speculative, further research is required to explore and understand the nuanced issues and causal factors associated with scientometric reproducibility where they may be distinguished from more general scientific reproducibility—the information content that citations should contain in the Web 3.0 era, the meaning and utility of citations to unpublished work products, and methods for interpreting relevance and influence of research based on citation data, among others. A broad community of collaborators that includes the creators of research, their research organizations, funding agencies, publishers, and service providers to the scientific and scholarly communications ecosystem is called for if we are to enjoy the benefits of these Web 3.0 era innovations.

Competing Interests Disclosure

The authors of this paper are employees of a for-profit entity, ARTiFACTS.ai (Artifacts of Research, Inc.), which is included in the analysis in this paper. Reasonable efforts have been made to be objective; regardless, this is noted in the interest of transparency.

References

Baker, M. (2016). 1,500 Scientists Lift The Lid On Reproducibility. Nature, 533, 452–454. https://doi:10.1038/533452a

Bartling, S., et. al. (2018). Blockchain for Open Science and Knowledge Creation (living document). Retrieved from https://10.5281/zenodo.401369

Borgman, C. L. (2015). Big Data, Little Data, No Data Scholarship in the Networked World. Cambridge: MIT Press.

Butler, L. (2003). Explaining Australia's increased share of ISI publications—The effects of a funding formula based on publication counts. Research Policy, 32, 143–155.

Campbell, E.G., Clarridge, B.R., Gokhale, M., et al. (2002). Data Withholding in Academic Genetics Evidence From a National Survey. JAMA. 287:4, 473–480. https://doi:10.1001/jama.287.4.473.

Center for Open Science. (2018). The Open Science Framework. Retrieved from https://osf.io/

Furlanello, C., De Domenico, M., Jurman, G., & Bussola, N. (2017). Towards A Scientific Blockchain Framework For Reproducible Data Analysis. arXiv preprint arXiv:1707.06552.

Garfield, E. (1979). Citation Indexing. Its Theory And Application In Science, Technology And Humanities. New York: Wiley.

Goodman, S. N., Fanelli, D., & Ioannidis, J. P. A. (2016). What Does Research Reproducibility Mean? Science Translational Medicine, 8:341.

Ioannidis, J.P.A. (2005). Why Most Published Research Findings Are False. PLoS Med 2(8): e124. https://doi.org/10.1371/journal.pmed.0020124.

Kennedy, D. (2006). Editorial Retraction of Hwang et al. Papers. Science 20: 311. 335. https://doi:10.1126/science.1124926

Moed, H.F. (2009). New Developments in The Use Of Citation Analysis in Research Evaluation. Archivum Immunologiae et Therapiae Experimentalis 57:13. https://doi.org/10.1007/s00005-009-0001-5.

Munafo, M.R., Nosek, B.A., Bishop, D.V.M., Button, K.S., Chambers, C.D., du Sert, N.P., Simonsohn, U., et. al., (2017). A Manifesto For Reproducible Science. Nature Human Behaviour 1. Article Number: 0021. https://doi:10.1038/s41562-016-0021.

Nosek, B.A., Spies, J.R., Motyl, M., (2012). Scientific Utopia: II. Restructuring Incentives and Practices to Promote Truth Over Publishability. Perspectives on Psychological Science, 7(6), 615-631. https://doi.org/10.1177/1745691612459058.

Shrestha, A. K., & Vassileva, J. (2018). Blockchain-Based Research Data Sharing Framework for Incentivizing the Data Owners. In International Conference on Blockchain (pp. 259-266). Springer.

Van den Besselaar, P., Heyman, U., & Sandström, U. (2017). Perverse Effects Of Output-based Research Funding? Butler's Australian Case Revisited. Journal of Informetrics, 11(3), 905–918.

Van Rossum, J. (2017). Digital Science Report Blockchain for Research Perspectives on a New Paradigm for Scholarly Communication. https://www.digital-science.com/resources/digital-research-reports/blockchain-for-research/

Velden, T., Hinze, S., Scharnhorst, A., Schneider, J.W., Waltman, L. (2018). Exploration of Reproducibility Issues in Scientometric Research Part 2: Conceptual Reproducibility https://arxiv.org/abs/1804.05026

Waltman, L., Hinze, S., Scharnhorst, A., Schneider, J.W., Velden, T. (2018). Exploration of reproducibility issues in Scientometric research Part 1: Direct Reproducibility. https://arxiv.org/abs/1804.05024

Wang, D., Song, C., Barabási, A. (2013). Quantifying Long-Term Scientific Impact. Science, 342:6154, 127-132. https://doi.org/10.1126/science.1237825