ROBHOOT
Open Research Network
Whitepaper v.1.0

January 21, 2020

1 Summary

Robhoot aims to fully automate the research cycle in a open decentralized network. Research automation with reporting generation might help to contrast informed decisions when solving complex social, environmental and technological problems. Current technologies for scientific inquiry and decision-making are highly fragmented and thus only increase robustness, reproducibility, open-access and the interactions with the public marginally. The goal of Robhoot is to propose a hybrid-neutral-technology to lay out the foundation for an open-science research ecosystem aiming to strengthen the robustness and reproducibility of science. Robhoot is not set out to deliver a finished research open network in the science ecosystem, but to provide a science-enabled technology in establishing a prototype proof-of-principle to connect automated, decentralized and neutral-knowledge generation with knowledge-inspired societies.

2 The Science Ecosystem

The process of science and technology requires multiple steps of information transfer among trusted/untrusted peers. Science and technology produce knowledge and features like reproducibility, decentralization, and immutability of knowledge are key to reach neutral open-access reports when taking informed decisions in complex social, environmental and technological problems. However, currently public funded science is highly centralized [1, 2], prone to errors [3], difficult to reproduce [4], and contains many biases [5]. In this regard, automated-based knowledge following a secure peer-to-peer architecture storing the open-source knowledge graphs derived from the paths of the research cycle is far from reality. Taken together, these elements make the connection between the scientific process, open-access and reproducible research reporting for decision-making highly improbable. Despite many projects are aiming at making the science ecosystem less centralized and biased while increasing openness and reproducibility a science-enabled technological paradigm connecting open-science to knowledge-inspired societies is not currently in place [2].

Many studies in decentralized ecosystems are producing an immense gain in detailed knowledge about scalability, security and decentralization trade-offs [6, 7, 8, 9, 10]. Automation and AI technologies is the other angle from which many advances are rapidly occurring [11, 12, 13]. Yet, while the existing technological paradigm is rapidly shifting towards science-based decentralization and automation technologies, end-to-end open-source research accounting for decentralized, neutral and automated knowledge-inspired technologies are missing (Table 1). Rapid advances of automated research platforms facilitating data integration accounting for part of the research cycle are currently under development[1] but open-source decentralized and automated networks accounting fully for the research cycle are still at a very incipient stage of development. While conceptual frameworks conceptualizing the required layers in many research fields are well established (Figure 1a), there is currently a lack of integration, development and automated tools connecting knowledge graphs (Figure 1b) to deep process-based learning networks to explore their robustness (Figure 1c) in fully decentralized ecosystems (Figure 1d).

Features	**Science Ecosystem**	**Robhoot 1.0**
Decentralization	No	Yes
Open-access	Mostly No	Yes
Immutability	No	Yes
Robustness	Mostly No	Yes
Reproducibility	Mostly No	Yes
Owner-Controlled assets	No	Yes

Table 1: *Robhoot aims to be designed to resolve desirable properties of science: Robustness, Reproducibility, Decentralization, Open and Direct access to reporting by peers and not-peers.*

3 Robhoot Design Goals

Robhoot will be developed in four different stages following standard version protocols. The most advanced version is to provide real-time open-access reporting by

[1] This is by no means an exhaustive list but it gives an indication of the many projects currently in place: NakamotoT,BigQuery,Automated statistician,Modulos,Google AI,Iris,easeml

a decentralized neutral-knowledge network to gain informed decisions when solving complex social, environmental and technological problems. Automating the research cycle in a open research network ultimately aims to contrast human-produced science with machine-produced science to enrich the human-knowledge graphs in a neutral-knowledge-inspired society. Figures 1 to 3 show Robhoot stages, Robhoot in a digital ecosystem and the timeline for each of the stages, respectively.

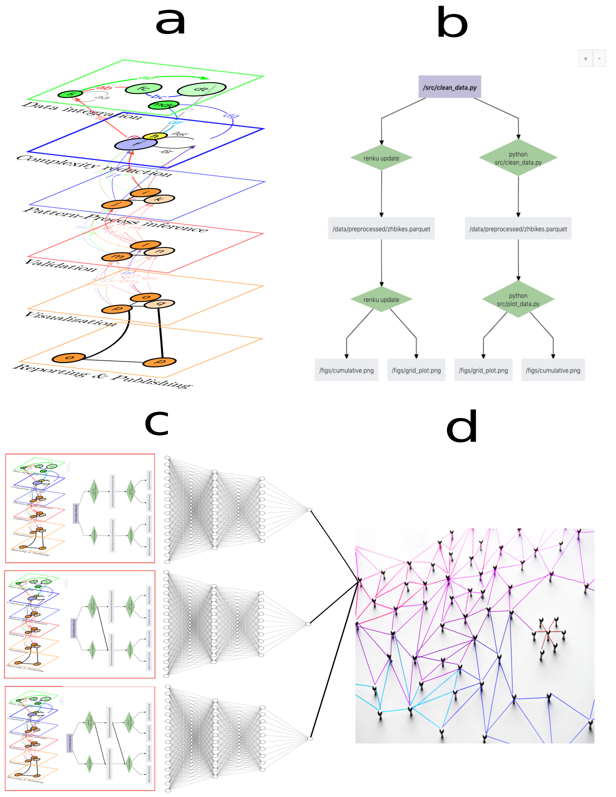

Figure 1: Automated knowledge-based network technology. **a) Robhoot 1.0** will account fully for the research cycle from data integration (top) to reporting generation (bottom). **b) Robhoot 2.0** will encode each path of the research cycle in **a** as a knowledge graph (KG). **c) Robhoot 3.0** will add deep knowledge-based networks to automatically explore populations of KGs to gain robustness of the process-based patterns contained in the data. **d) Robhoot 4.0** will deploy all KGs in a distributed network of mutually trusting/untrusting peers with every peer maintaining the population of the KGs.

The overall objectives and a brief numeration of the tools and methods to be used/developed in each stage for each of the four major Robhoot versions are the following:

3.1 Robhoot 1.0: Automated Research Cycle

- Develop, deploy and integrate open-source algorithms to fully automate the research cycle (Figure 1a).
- Exploration of robustness within and between layers from data integration, complexity reduction, inference, and validation to visualization and automated reporting (Figure 1a).
- Robhoot 1.0 testnet to explore a Open Research Network in Biodiversity and Global Change Research to connect open science (i.e., citizen and other data-driven models) to real-time open-access knowledge generation to gain informed decisions when solving local and global environmental problems.

- **Tools and Methods**: Multilayer networks, Bayesian Networks, Network metrics, Julia computing language, Open-source software protocols, Gitchain, ETLs algorithms, Kafka, Clickhouse.

3.2 Robhoot 2.0: Knowledge Graphs

- Implementation of algorithms to reproduce paths of the research cycle with Knowledge Graphs (KGs) (Figure 1b).
- Robustness and stability exploring a suite of open-source lineage client-tracker algorithms.

- **Tools and Methods**: Knowlegde graph algorithms and packages (i.e., Renku and others).

3.3 Robhoot 3.0: Deep learning networks

- Deploy automated deep learning algorithms to sample paths of the research cycle to produce populations of Knowledge Graphs (KGs) (Figures 1a-c).
- Exploration of the robustness of automated research cycle combining optimization algorithms and the population of Knowledge Graphs (Figure 1c).

- **Tools and Methods**: Multilayer networks, Neural Biological Networks, Bayesian Networks, Deep learning networks. Optimization algorithms.

3.4 Robhoot 4.0: Distributed ledger network

- Deploy a permissioned-permissionless distributed ledger technology to guarantee decentralization, open-access, neutral-knowledge-based network and prior confidenciality/posterior reproducibility of the KGs populations (Figures 1c and 1d).
- Exploration of a suite of consensus algorithms and smart contracts among trusted-untrusted peer-to-peer interactions to infer macroscopic metrics of the open research network (Figure 1d).
- Quantification of metrics to study the scalability-security-decentralization trade-offs when storing KGs in the research network (Figure 1d).
- Testnet case study to explore the interaction between consensus protocols and the scalability-security-decentralization trade-offs when committing the KGs to the distributed ledger.
- Mainnet to cryptographically link each population of KGs to previous KGs-ledger to create an historical KGs-ledger chain that goes back to the genesis ledger in the open research network. The mainnet aims to connect multiple database with real-time open-access citizen data science to knowledge-inspired societies.

- **Tools and Methods**: Distributed computing algorithms, Blockchain and consensus algorithms, BighainDB, Gitchain. Telegram open network, Golem.

4 Robhoot in Digital Ecosystems

The science ecosystem currently lack technologies fully automating the research cycle into digital ecosystems. Despite public institutions are demanding more reproducibility and openness of the data and the scientific process, and overall a shifting towards open and reproducible scientific and engineering landscapes, there are not currently open and integrated technologies aiming to compactly facilitate and distribute the scientific and engineering knowledge in open, reproducible and immutable knowledge networks.

Automating knowledge-generation requires the integration of many distinct traits or features. Usually, knowledge-generation comes from interactions within- and between-layers of the scientific process (Figure 1a). The feedbacks occurring among layers in the science and technology ecosystem also provide unexpected behaviors that are difficult to anticipate. Therefore many feedbacks and interactions within- and between-layers are not easy to reproduce if not properly accounted for. Robhoot will take advantage of the open-source software community to explore how knowledge graphs, optimization, automation, and decentralization algorithms can be connected to the robustness and reproducibility of the scientific process.

Therefore, Robhoot aims to be a hybrid-technology accounting for many traits (Table 1: decentralized, open-access, immutable, robust, reproducible, and secure with trusted/untrusted peer-to-peer interactions). Producing such a multitrait technology means multidisciplinarity teams compactly making functional interactions within a rapidly evolving digital ecosystem. In this regard, Robhoot aims to put together scientists and engineers from data science, computer science (i.e., distributed computing, software development), the physics of complex systems (i.e., multilayer networks), artificial intelligence (i.e., deep learning and automation) and the biology, ecology and evolution of social, natural and technological ecosystems.

One way of visualizing the multitrait dimensionality of Robhoot in the digital ecosystem is to connect each layer of the scientific process (Figure 1a) to the open-source software required to gain functionality of the automated research cycle (Figure 2). For example, Node 0 (left column, Figure 2) can be the Data Integration layer (Figure 1a). This node is connected to seven nodes that might represent open-source ETLs (i.e., extract, transform, load data open-source software, central column, Figure 2). Connections between Node 0 and nodes 5, 6, 8, 9, 10, 12 and 13 might be rapidly evolving (i.e., indicated by the different red tones of the connections). Indeed, open-source ETLs are rapidly evolving towards accounting for many heterogeneous aspects of data integration (i.e., formats, historical-real time, storage, dimensions, size, bias and spatiotemporal resolution). ETLs might also be connected to a gradient of reporting generation (i.e., right column, Figure 2) noting reports containing only a subset of the interactions of the digital ecosystem network. The network of the fully automated research cycle can be one where Nodes 0, 1, 2, 3, and 4 represent the different layers of the research cycle (left column, Figure 2 and Figure 1a) connected to the open-source software of the digital ecosystem (central column, Figure 2) to produce the full population of reports (right column, Figure 2).

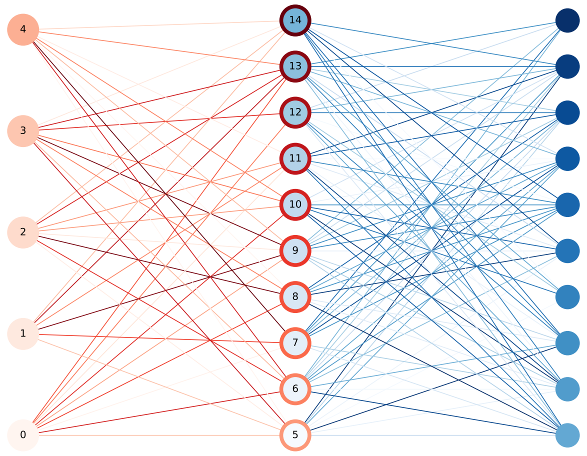

Figure 2: Robhoot in Digital Ecosystems: **Left column**: **Robhoot 1.0** containing the research cycle represented as nodes (i.e., from node 0 to 4: Data integration (0), Complexity Reduction (1), Inference (2), Validation (3), and Visualization(4)). **Central column**: The research cycle layers connected to the open-source software in the digital ecosystem. Nodes can for example represent the ETLs open-source software required to produce a general data integration accounting for many data heterogeneities. **Right column**: Reporting gradient connected to the open-source software where each report (i.e., represented as a node) is generated only using a subset of the research layers and ETLs.

5 Conclusion

Science and technology ecosystems are in need of accounting for the uncertainties, reproducibility and immutability related to the complexity of the research process. This need is not just for a specific stage of the research cycle, but from data acquisition and integration to automated reporting generation because knowledge-inspired societies and decentralized governance will demand full research cycle transparency to solve complex social, environmental and technological problems. This need brings many challenges to our research proposal because obtaining robust knowledge from integrating many layers of the research cycle, each containing its own set of methods and uncertainties, can generate divergent, fragile and contradictory outcomes.

We will develop a flexible research method focusing step by step in different stages with varying levels of complexity (i.e., from Robhoot 1.0 to 4.0, Figure 3). Our motivation will be to provide a first open-access proof of concept of how the technology works: we will automate reproducible research paths along a multilayer network (Robhoot 1.0) to sample the KGs (Robhoot 2.0) using different deep learning algorithms to estimate the uncertainty of the ruled-based inference obtained by fitting predictions to simulated data (Robhoot 3.0). Accounting for the uncertainties of each of the research stages when sampling

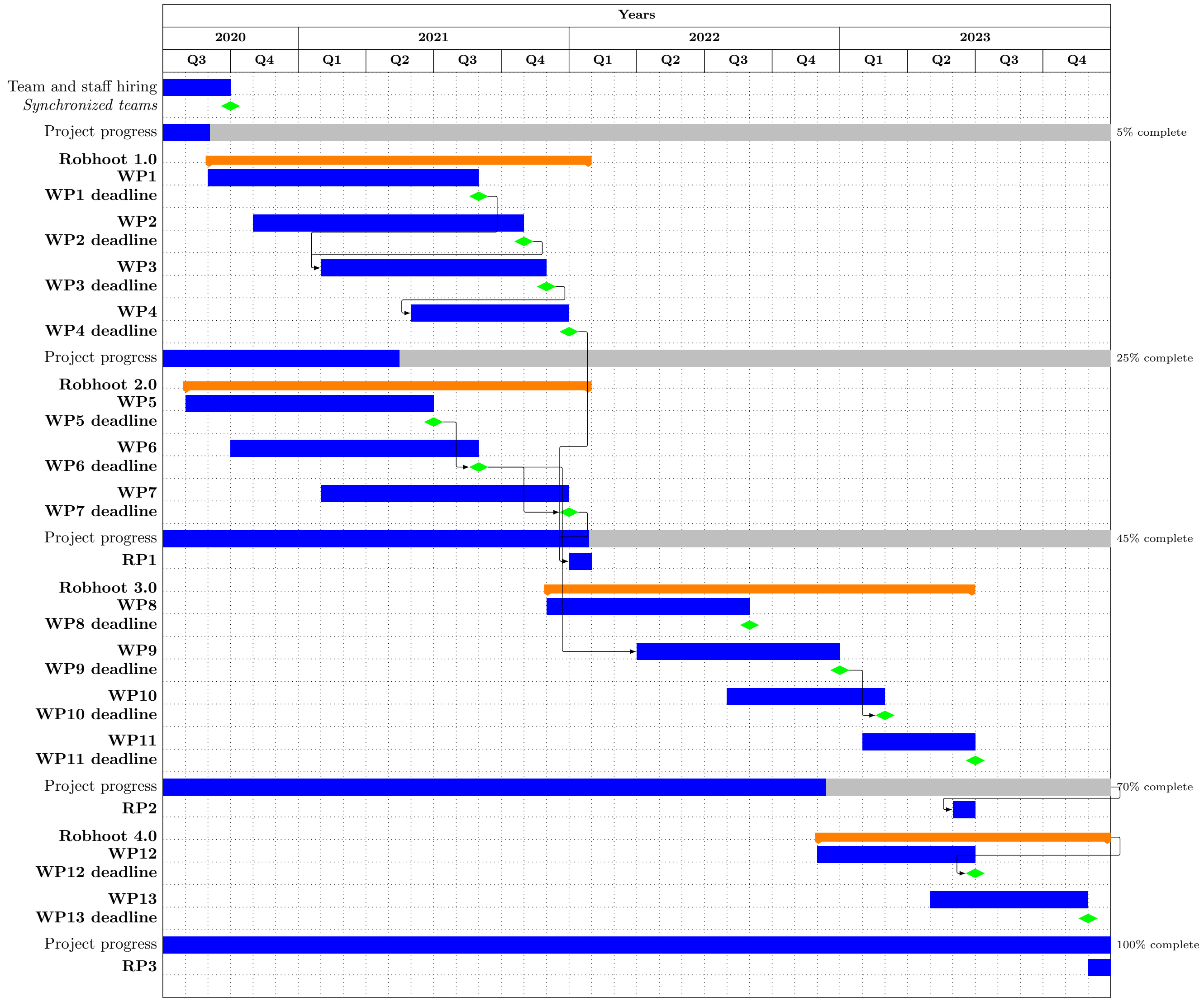

Figure 3: The Robhoot roadmap: **Robhoot 1.0** working packages **WP1** to **WP4**: Develop, integrate, deploy and test the functionality of interacting open-source algorithms to fully automate the research cycle (Figure 1a). **Robhoot 2.0** working packages **WP5** to **WP7**: Fully automate end-to-end research cycle exploring, implementing an testing knowledge graphs. **Robhoot 3.0** working packages **WP8** to **WP11**: Developing, implementing and testing deep knowledge-based networks to automatically explore populations of KGs to gain robustness of the process-based patterns contained in the data. **Robhoot 4.0** working packages **12** to **13**: Deploy all KGs in a distributed network of mutually trusting/untrusting peers with every peer maintaining the population of the KGs.

the KGs comes from the many distinct paths within and across the layers in the research cycle (Figure 1a). Robhoot will test a variety of consensus algorithms to explore the degree of security, decentralization and scalability of the ledger knowledge network using the generated population of KGs (Robhoot 4.0).

Despite our focus will be bias towards the algorithmic robustness during the four stages of Robhoot development, we will develop a domain-specific case study, a Robhoot Open Network, to test the robustness of the rule-based inference obtained by fitting each of the generated KG to empirical patterns. The high risk associated to robustly automate the full research cycle for producing immutable open knowledge will be buffered to a great extend because the existing digital ecosystem of highly reliable open-source software tools (Figure 2).

References

[1] H. Inhaber. Changes in centralization of science. *Research Policy*, 6(2):178–193, apr 1977. ISSN 0048-7333. doi: 10.1016/0048-7333(77)90024-5. URL `https://www.sciencedirect.com/science/article/abs/pii/0048733377900245`.

[2] Vlad Günther and Alexandru Chirita. " Scienceroot " Whitepaper. 2018. URL `https://www.scienceroot.com/`.

[3] Ferric C Fang and Arturo Casadevall. Retracted Science and the Retraction Index. *Infection and Immunity*, 79(10):3855 LP – 3859, oct 2011. doi: 10.1128/IAI.05661-11. URL `http://iai.asm.org/content/79/10/3855.abstract`.

[4] Tom E. Hardwicke, Maya B. Mathur, Kyle MacDonald, Gustav Nilsonne, George C. Banks, Mallory C. Kidwell, Alicia Hofelich Mohr, Elizabeth Clayton, Erica J. Yoon, Michael Henry Tessler, Richie L. Lenne, Sara Altman, Bria Long, and Michael 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*, 5(8):180448, sep 2018. ISSN 20545703. doi: 10.1098/rsos.180448. URL `https://doi.org/10.1098/rsos.180448`.

[5] John P a Ioannidis. Why most published research findings are false. *PLoS medicine*, 2(8):e124, aug 2005. ISSN 1549-1676. doi: 10.1371/journal.pmed.0020124. URL `http://www.ncbi.nlm.nih.gov/pubmed/16060722`.

[6] Golem. The Golem Project Crowdfunding Whitepaper. *Golem.Network*, (November):1–28, 2016. URL `https://golem.network/crowdfunding/Golemwhitepaper.pdf`.

[7] Nikolai Durov. Telegram Open Network. pages 1–132, 2017.

[8] Elli Androulaki, Artem Barger, Vita Bortnikov, Srinivasan Muralidharan, Christian Cachin, Konstantinos Christidis, Angelo De Caro, David Enyeart, Chet Murthy, Christopher Ferris, Gennady Laventman, Yacov Manevich, Binh Nguyen, Manish Sethi, Gari Singh, Keith Smith, Alessandro Sorniotti, Chrysoula Stathakopoulou, Marko Vukolić, Sharon Weed Cocco, and Jason Yellick. Hyperledger Fabric: A Distributed Operating System for Permissioned Blockchains. *Proceedings of the 13th EuroSys Conference, EuroSys 2018*, 2018-Janua, 2018. doi: 10.1145/3190508.3190538.

[9] Ocean Protocol Foundation, BigchainDB GmbH, and DEX Pte. Ltd. Ocean Protocol: A Decentralized Substrate for AI Data & Services Technical Whitepaper. pages 1–51, 2018. URL `https://oceanprotocol.com/`.

[10] BigchainDB GmbH. BigchainDB: The blockchain database. *BigchainDB. The blockchain database.*, (May):1–14, 2018. doi: 10.1111/j.1365-2958.2006.05434.x. URL `https://www.bigchaindb.com/whitepaper/bigchaindb-whitepaper.pdf`.

[11] J Schmidhuber. Deep learning in neural networks: An overview. *Neural Networks*, 61:85–117, 2015.

[12] Markus Reichstein, Gustau Camps-Valls, Bjorn Stevens, Martin Jung, Joachim Denzler, Nuno Carvalhais, and & Prabhat. Deep learning and process understanding for data-driven Earth system science. *Nature*. ISSN 0028-0836. doi: 10.1038/s41586-019-0912-1. URL `www.nature.com/nature`.

[13] Yolanda Gil, Bart Selman, Marie Desjardins, Ken Forbus, Kathy Mckeown, Dan Weld, Tom Dietterich, Fei Fei Li, Liz Bradley, Daniel Lopresti, Nina Mishra, David Parkes, and Ann Schwartz Drobnis. A 20-Year Community Roadmap for Artificial Intelligence Research in the US Roadmap Co-chairs: Workshop Chairs: Steering Committee. Technical report, 2019. URL `https://bit.ly/2ZNVBVb`.