ROBHOOT
Open Research Network
Whitepaper v.2.0

February 1, 2020

1 Summary

Global sustainability is a major goal of humanity. Many studies have shown global sustainability could be achieved by strengthening feedbacks between social, ecological and governance systems. Sustainability goals, however, strongly depend on filling evidence- and research-based knowledge gaps. Yet, the science ecosystem lack expertise in developing technologies to narrow down the knowledge gaps in a reproducible and open science ecosystem. In this whitepaper we introduce Robhoot, an open decentralized research network generating reproducible science fully accounting for the research cycle. Robhoot targets the reduction of knowledge gaps by generating open-access reproducible global science reports. Robhoot is to develop a hybrid-neutral-technology to lay out the foundation of an open-science ecosytem strengthening the robustness, decentralization 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

Science and technology contain multiple steps of information transfer among trusted/untrusted peers. As a consequence, science generates knowledge with specific features. Which are the desirable features of human-generated knowledge to gain informed decisions in complex social, governance, environmental and technological problems? How much knowledge generation can be automatically achieved to enrich human-expertise generated knowledge? Currently public funded science is highly centralized [1, 2], prone to errors [3], difficult to reproduce [4], and contains many biases [5]. This makes science an ecosystem with increasing problems to decrease the evidence- and research-based knowledge gaps in humanity. Here we introduce an automated open research network technology fully accounting for the research cycle. The goal of such a technology is to reduce the global knowledge gaps while accounting for centralization, bias, error-prone, non-reproducibility and lack of incentives in the existing science and technology ecosystem (Figure 1 and Table 1).

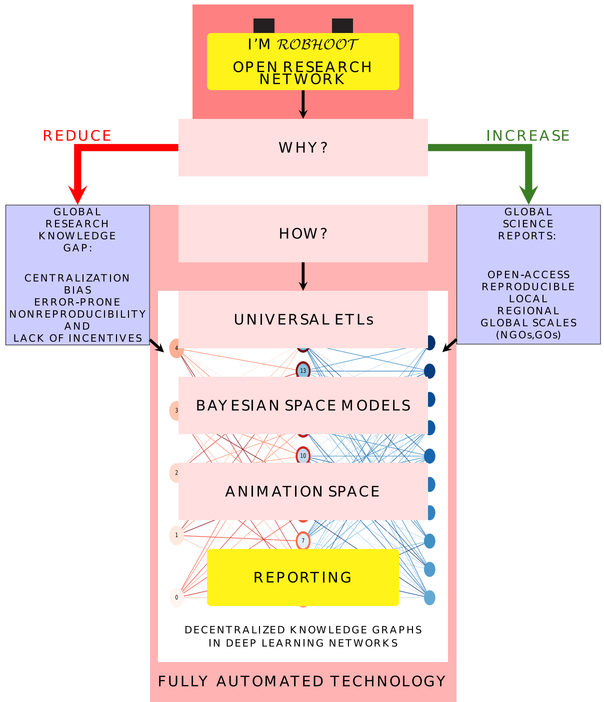

Figure 1: The architecture of an open automated research network technology. Robhoot targets a decrease of global knowledge gaps (red path) and an increase in open and reproducible global science reports (green path). Open automated knowledge generation enrich human-expertise generated science by simultaneously accounting for centralization, bias, error-prone, non-reproducibility, and lack of incentives (red path). Robhoot will explore open-access automation in digital ecosystems merging/developing the following technologies: First, **Universal ETLs** (i.e., Extraction, Transformation and Load algorithms) to reach complete database integration and complexity reduction. Second, **Bayesian Space Models** to explore deep-process-based learning networks and data-models optimization metrics. Third, **Animation Space** algorithms visualizing fitting procedures between empirical and model patterns. Four, **Reporting** in natural processing languages to generate reproducible reports based on the population of knowledge graphs, and fifth, **Distributed algorithms** to decentralize knowledge graph generation making them immutable, open and globally accessible.

The core features of Robhoot are decentralization and automation. Currently, many studies focusing on decentralized ecosystems are producing an immense gain of

Features	Science Ecosystem	Robhoot
Decentralization	No	Yes
Full automation	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: Decentralization, Automation, Open-access, Immutability, Robustness, Reproducibility, and Owner-controlled assets. These features will be added during the different stages of the development of Robhoot (see section "Robhoot Design Goals").*

knowledge about scalability, security and decentralization trade-offs [6, 7, 8, 9, 10], but its impact in the science ecosystem is currently not well understood. Automation and AI technologies is the other angle from which many advances are rapidly occurring [11, 12, 13]. Yet, while the existing technological paradigm in many sectors is rapidly shifting towards science-based decentralization and automation technologies, the science ecosystem lack open-source technologies accounting for decentralized, neutral and automated knowledge-inspired generation (Figure 1 and Table 1) [2]. Rapid advances of automated research platforms facilitating data integration accounting for parts of the research cycle are currently under development.[1] Yet, open-source decentralized and automated technologies accounting fully for the research cycle are at a very incipient stage of development. While conceptual frameworks conceptualizing the required steps in many research fields are well established (Figure 2a), there is currently a lack of integration, development and automated tools connecting knowledge graphs (Figure 2b) to deep process-based learning networks (Figure 2c) in fully decentralized ecosystems (Figure 2d).

3 Robhoot Design Goals

Robhoot will be developed in four stages following standard version protocols. The most advanced version is to provide real-time open-access reporting in a decentralized network to gain informed decisions when solving complex social, governance, environmental and technological problems. Automating the research cycle in a open research network ultimately aims to contrast human-generated science with machine-generated science to target neutral-knowledge-generation in knowledge-inspired societies. Figures 1 to 4 show Robhoot goals and architecture, the four main milestones of Robhoot, Robhoot in a digital ecosystem and the timeline of working packages for each of the stages, respectively.

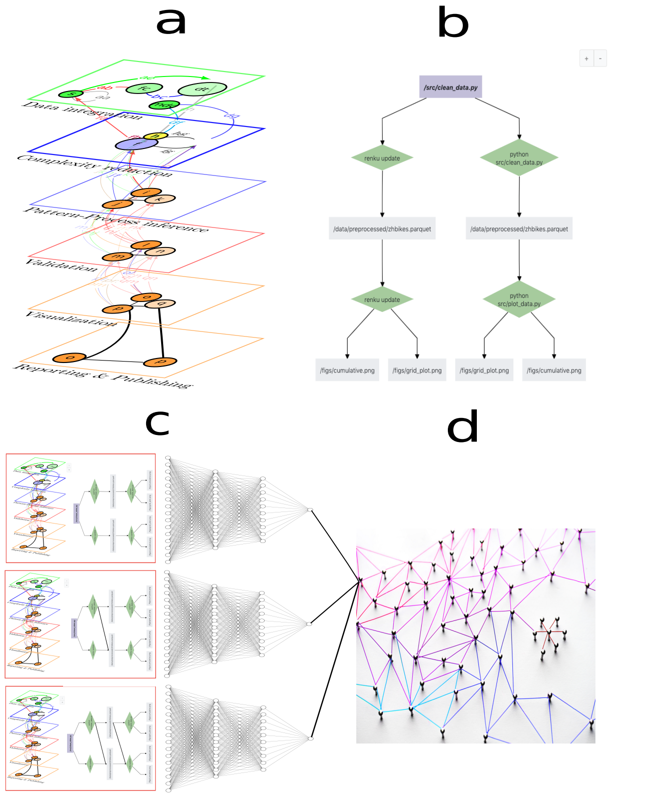

Figure 2: Milestones of an automated knowledge-based network technology. a) **Robhoot 1.0** will feature a fully research cycle from data integration (top) to reporting generation (bottom). **b)** **Robhoot 2.0** will automatically track the research cycle as shown in **a** using knowledge graphs (KGs). **c)** **Robhoot 3.0** will feature deep leaning networks to automatically explore populations of KGs to gain undestanding of the process-based patterns contained in the data, and **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 with the tools, methods, and potential bottlenecks in each stage for each of the four major Robhoot versions are the following:

3.1 Robhoot 1.0: Automated Research Cycle

- **Universal ETLs** will connect generalized algorithms to open-source software to extract, transform and load data with different properties (i.e., formats, historical-real time, storage, dimensions, size, sampling bias and spatiotemporal resolution) (Figures 1 and 2a, two top layers).
- **Bayesian space models** will explore generalized open-ended language of models combining Bayesian networks and optimization methods. The Bayesian space models module will search, evaluate models, trading-off complexity, fit to data and quantify resource usage (Figures 1 and 2a, inference and validation layers).
- **Animation Space** will connect open-source visualization software to the exploration of open-ended models to make the whole search transparent, highly visual and reproducible (Figures 1 and 2a, visualiza-

[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

tion layer).

- **Reporting** will develop a procedure to automatically explain the structure of the Bayesian space modeling module. It will also communicate the module using visualizations of the procedures followed by the Universal ETLs and Bayesian space models modules (Figures 1 and 2a, reporting layer).
- Robhoot 1.0 testnet will use "Biodiversity and Global Change Research databases" to explore the robustness of the automated research cycle, from the **Universal ETLs** and **Bayesian space models** to the **Animation space** and **Reporting**.

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

3.2 Robhoot 2.0: Knowledge Graphs

- Implementation of algorithms to reproduce paths of the research cycle with Knowledge Graphs (KGs) (Figure 2b).
- 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 2a-c).
- Exploration of the robustness of automated research cycle combining optimization algorithms and the population of Knowledge Graphs (Figure 2c).

- **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 2c and 2d).
- 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 2d).
- Quantification of metrics to study the scalability-security-decentralization trade-offs when storing KGs in the research network (Figure 2d).
- 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 the open-source digital ecosystem. 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 features. Usually, knowledge-generation comes from interactions within- and between-layers of the scientific process (Figure 2a). The feedbacks occurring within and 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 (Figure 1).

Robhoot aims to be a hybrid-technology accounting for many featurers (Table 1). Producing such a multi-feature technology requires multidisciplinarity teams making functional contributions for each of the Robhoot features while integrating all these features in 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 2a) to the open-source software community required to gain functionality of the automated research cycle (Figure 3). For example, Node 0 (left column, Figure 3) can be the Data Integration layer in Figure 2a. This node is connected to seven nodes representing open-source ETLs open-source software (i.e., extract, transform, load data, central column, Figure 3). Connections between Node 0 and nodes 5, 6, 8, 9, 10, 12 and 13 can 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 can also be connected to a gradient of reporting generation (i.e., right column, Figure 3) 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 3 and Figure 2a) connected to the open-source software of the digital ecosystem (central column, Figure 3) to generate full populations of reports (right column, Figure 3).

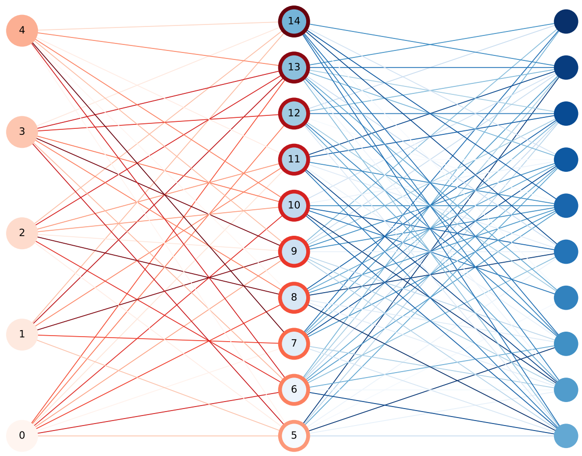

Figure 3: Robhoot in Digital Ecosystems: **Left column**: **Robhoot 1.0** representing the research cycle as nodes from number 0 to 4: Data integration (0), Complexity Reduction (1), Inference (2), Validation (3), and Visualization(4)). **Central column**: Nodes representing the research cycle in the left column are connected to open-source software in the digital ecosystem. Connections with node number 0 in the left column can, for example, represent the ETLs open-source software interactions required to generate the **Universal ETLs** module. The same meaning applies to the different nodes of the left column. **Right column**: Each node represents a report meaning there is a reporting gradient generated by the connections to the open-source software from where each report is generated only using a subset of the research layers and open-source software.

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 4). 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 the KGs comes from the many distinct paths within and across the layers in the research cycle (Figure 2a). 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 3).

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`.

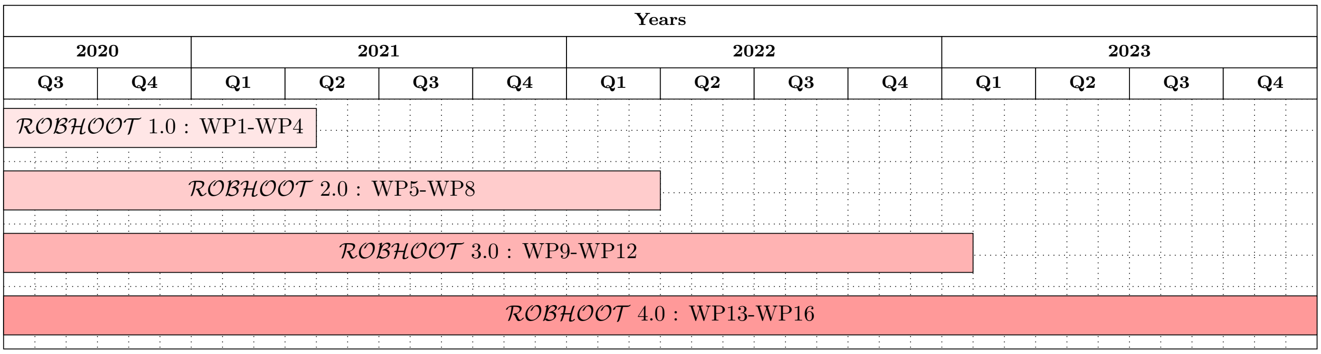

Figure 4: Robhoot roadmap: **Robhoot 1.0** working packages **WP1** to **WP4** will take care of the integration of **Universal ETLs**, **Bayesian Space Models**, **Animation Space**, and **Reporting** to generate a fully automated research cycle (Figure 2a). **Robhoot 2.0 WP5** to **WP8** will deploy knowledge graphs (KGs) into the research cycle (Figure 2b). **Robhoot 3.0 WP9** to **WP12** will explore deep learning networks scenarios to sample KGs populations to gain understanding of the robustness of the patterns in the data under distinct research routes (Figure 2c). **Robhoot 4.0 WP13** to **WP16** will deploy KGs populations into a decentralized network of mutually trusting/untrusting peers with every peer maintaining the population of the KGs.

[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`.