ROBHOOT
Open Discovery Network
FET v.1.0

February 28, 2020

1 Summary

Global sustainability is a major goal of humanity. Many studies have shown global sustainability could be achieved by strengthening transparency and feedbacks between social, ecological, technological and governance systems. Sustainability goals, however, strongly depend on global access to evidence- and research-based knowledge gaps. Yet, the science ecosystem lacks open-source technologies narrow down knowledge gaps. We introduce an open discovery network targeting knowledge gaps throughout open-access generation of fully reproducible science reports. Open discovery network encompasses a hybrid-automated-technology to lay out the foundation of an open-science ecosytem strengthening the robustness, decentralization, reproducibility and social accessibility of science. The project summarized here is not set out to deliver a finished open discovery network, but to provide the architecture of a science-enabled technology as a proof-of-principle to connect decentralized and neutral-knowledge generation with knowledge-inspired societies.

2 Excellence

2.1 Radical vision of a science-enabled technology

We are in the midst of the fourth industrial revolution, a transformation revolving around data driven intelligent machines. Yet, despite the rapid evolution of the digital ecosystem around data driven intelligent machines, open discovery-based technologies facilitating global access to informed decisions when solving complex social, environmental and technological problems are particularly lacking. How can data driven intelligent machines help to reach global sustainability goals by reducing knowledge gaps? The *ROBHOOT* project introduces new concepts to global knowledge gaps. Current technologies for automated scientific inquiry are highly fragmented, partly solve reproducibility and are mostly developed in close-source software. Thus, despite the importance of global access to knowledge gaps for reaching sustainability goals, open-source technologies fully accounting for the research cycle to generate reproducible science reports are lacking. The goal of *ROBHOOT* is to propose a new hybrid-technology concept combining question- and reproducibility-based decentralized knowledge-graphs to lay the foundation for a novel, global access to scientific discovery technology.

ontain multiple steps of information transfer among trusted and untrusted peers. As a consequence, science generates knowledge with specific features. Which are the desirable features of human-generated knowledge? Should such features be aligned with taking informed decisions in

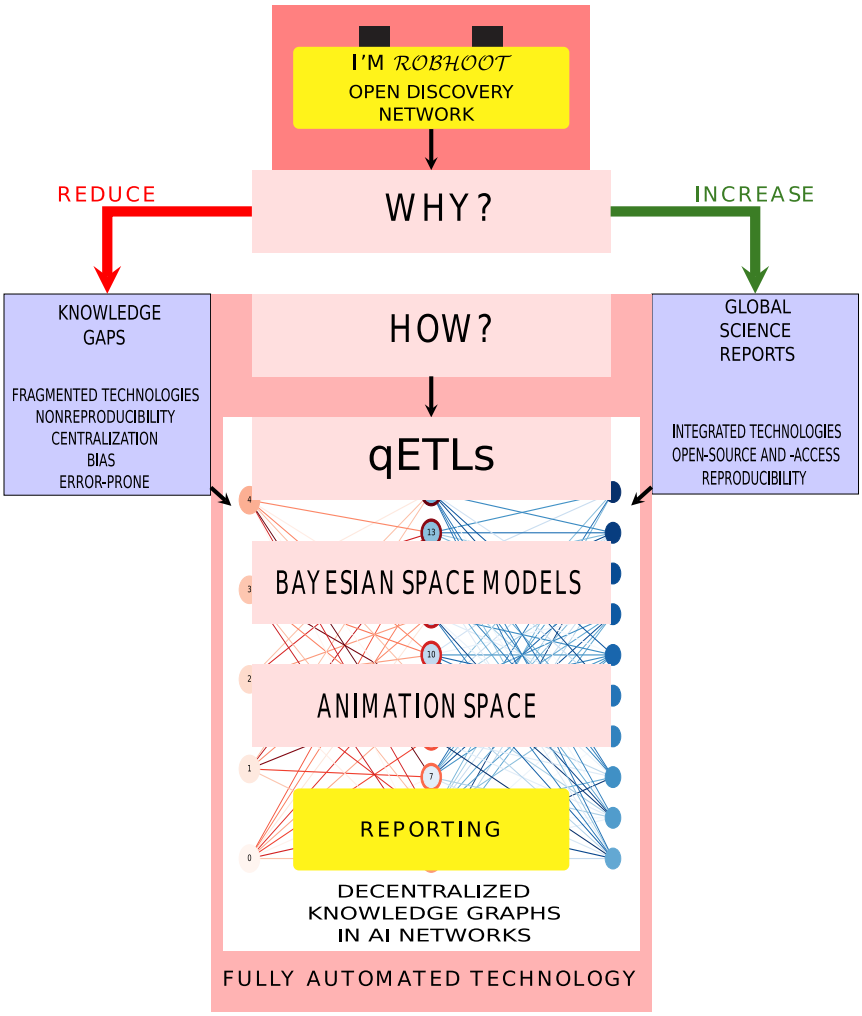


Figure 1: Open discovery network technology. *ROBHOOT targets global knowledge gaps (red path) and open-access fully reproducible discovery reports (green path). ROBHOOT is open-source science-based automated technology decentralizing reproducible knowledge graphs. ROBHOOT integrates 1) question-based knowledge gaps with data extraction, transformation and loading algorithms for database integration and complexity reduction (***qETLs***), 2)* **Bayesian Space Models** *accounting for open-ended model optimization techniques, 3)* **Animation Space** *visualizing fitting procedures and pattern generation connecting empirical data and open-ended models, and 4) open and globally accessible* **Reporting** *formalized in natural processing language.*

complex social, governance, environmental and technological problems? Can global transparency in knowledge generation be achieved to facilitate sustainability goals of humanity? 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 building up problems to decrease the evidence- and research-based knowledge gaps in humanity [6]. Here we propose an open research network fully accounting for the research cycle. The goal of such a technology is to reduce 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).

The core feature of an open research network is to embed knowledge-generation into reproducible automation and decentralization. Currently, many studies focusing on decentralized ecosystems are producing an immense gain of knowledge in a variety of sectors about scalability, security and decentralization trade-offs [7, 8, 9, 10, 11]. In the open science ecosystem, only a few implementations of decentralized technologies exist [2]. Automation and AI technologies represent the other angle from which many advances are rapidly occurring in digital ecosystems [12, 13, 14].

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 is designed to resolve desirable properties of science: Open-access, immutability, robustness, reproducibility, and owner-controlled assets. These features will be added during the different stages of development of the project (section "Design Goals").*

While the existing technological paradigm in many sectors is rapidly shifting towards science-based decentralization and automation technologies, the science ecosystem currently lack decentralized, neutral and open-source knowledge-inspired technologies strongly impacting knowledge-inspired societies (Figure 1 and Tables 1 to 3). Rapid advances of research platforms automating parts of the research cycle are currently under development [15, 16].[1] However, most of the existing research projects aiming to automate part of the research cycle are being built around close-source software. Therefore, open-source, reproducible, decentralized and automated technologies accounting for the research cycle are at a very incipient stage of development. To move forward open-source technologies accounting for the research cycle we need to compactly integrate knowledge-generation (Figure 2a) to automated tools connecting knowledge graphs (Figure 2b) and deep learning networks (Figure 2c) in fully decentralized ecosystems (Figure 2d).

3 The objectives of the proposal

The ROBHOOT consortium will build on the achievements of many open-source software projects, the open-source digital ecosystem, exploring novel features to reach a fully automated technology targeting a global reduction of knowledge gaps. ROBHOOT proposes to go far beyond the existing partial solutions to reproducibility and automation to track, understand and predict how knowledge is made during the discovery process. ROBHOOT will build science-process understanding by making the full research cycle reproducible.

It is a vision that can steer discovery into the future global access of knowledge-inspired societies.

Many technologies at the center of LifeTime are key European research strengths that the Flagship could boost. These include single-cell technologies combined with advanced imaging, artificial intelligence and patient-matched organoids, or organ-on-a-chip disease models to study the progression of an illness and develop novel therapeutics.

is set to be developed in four stages each with one main goal (Figure 2). The most advanced version is to provide open-access automated and fully reproducible reporting in a decentralized network. ROBHOOT v.1.0 features an automated research cycle connecting question generation to data integration and reporting. **b**) ROBHOOT **v.2.0** traces research paths as shown in **a** using reproducible knowledge graphs (KGs). **c**) ROBHOOT **v.3.0** contrasts deep leaning networks to explore populations of KGs for gaining undestanding of the process-based patterns contained in

[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,datarobot,aito,eureqa

the data, and **d**) *ROBHOOT* **v.4.0** deploys KGs in a decentralized network of trusting/untrusting peers with every peer maintaining the population of the KGs.

4 How they will be achieved

Mapping partly the discovery process is highly informative and challenging by itself, but a diverse group of X scientists across Europe decided that merely taking for reproducibility parts of the scientific process is not enough. Science is a highy dynamic process and there are many paths from where it can be achieved. To understand how discovery is generated, these scientists want to track the steps from questions to reporting and dissemination. To this end, they formed the *ROBHOOT* consortium with the goal of establishing an integrated toolbox containing several novel methods. *ROBHOOT* scientists will develop analytical and computational strategies such as machine-learning and artificial intelligence methods that help to understand discovery mechanisms and predict the future of global disemination of knowledge to science towards knowledge-inspired society. This strategy is expected to improve early access to discovery, predict the course of how science is evaluated and indentify new emerging targets where automation and global reports can play a key role in knowledge-inspired societies. *ROBHOOT*'s goals will be developed in four different stages.

4.1 *ROBHOOT* v.1.0: Automated Research Cycle

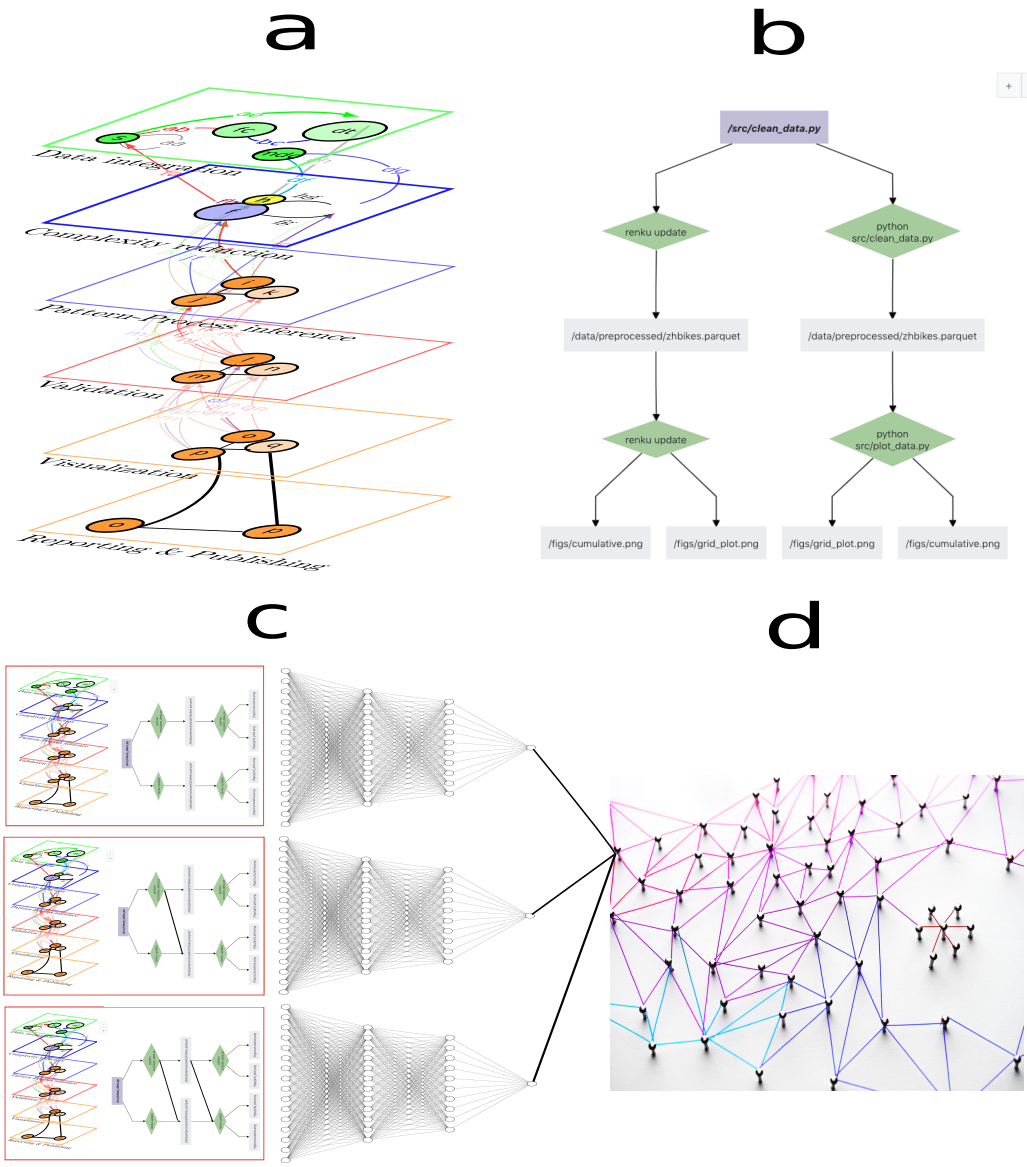
Discovery usually starts with an idea, and the research process is all about successfully bringing such an idea to the science ecosystem. and to the Algorithms detecting questions poorly explored but important for multidisciplinarity in the discovery process.

- **Question generator** finds the weaknesses of question knowledge graphs to discover relevant topics across discipines.
- **Universal ETLs** connects open-source generalized fault-tolerance algorithms connecting relevant questions to the extraction, transformation and load of data with unique features (i.e., formats, historical-real time, storage, dimensions, size, sampling bias and spatiotemporal resolution, etc) (Figures 1 and 2a, two top layers).
- **Bayesian space models** explore 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, visualization 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 is an automated reporting generation on "Biodiversity, Global Change and Sustainability Research" to explore the robustness of the automated research cycle accounting for **Universal ETLs**, **Bayesian space models**, **Animation space** algorithms and **Reporting** in natural processing languages.

- **Tools and Methods**: Multilayer networks metrics, Bayesian Networks, Julia, Python, Open-source software protocols, Gitchain, ETLs open-source software, Kafka, Clickhouse, Fluentd,

Hadoop.

- **Novel territory**: Develop universal open-source ETLs algorithms and Bayesian space models and connect them to reporting automation in a “Biodiversity, Global Change and Sustainability Research” case study.



4.2 *ROBHOOT* v.2.0: Reproducible Knowledge Graphs

- Implementation of algorithms tracking paths of the research cycle with Reproducible Knowledge Graphs (KGs) (Figure 2b).
- Robustness and stability of searching and fitting procedures following a suite of open-source lineage client-tracker algorithms.

- **Tools and Methods**: Reproducible knowlegde graph algorithms and open-source packages (i.e., Renku and others).

- **Novel Territory**: Contrasting a set of Reproducible Knowledge Graphs algorithms to quantify the reproducibility, reusability, and recovery properties of the full research cycle.

4.3 *ROBHOOT* v.3.0: Deep learning networks

- Deployment of 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 the automated research cycle combining optimization algorithms and the population of Knowledge Graphs (Figure 2c).

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

- **Novel Territory**: Join Bayesian networks models to biology inspired deep-learning networks to efficiently explore constrained model space and the robustness properties of the populations of KGs along ensembles of the research cycle.

4.4 *ROBHOOT* v.4.0: Distributed ledger network

- Deployment of a permissioned-permissionless distributed ledger technology to guarantee decentralization, open-access, neutral-knowledge-based network generation 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 of the open research network. Launching of the mainnet to connect multiple database integration with real-time open-access citizen data science and knowledge-inspired societies.

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

- **Novel Territory**: Deployment of contrasting functional consensus algorithms to explore decentralization and robustness properties of the KGs populations along ensembles of the research cycle space.

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 (Tables 1 and 2).

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. We will take advantage of the open-source software community to explore knowledge graphs, optimization, automation, and decentralization algorithms together to study the robustness and reproducibility properties of the scientific process (Figures 1 and 2).

One way of visualizing the dimensionality of *ROBHOOT* in the digital ecosystem is to connect each layer of the scientific process (Figure 2a) to open-source software to gain functionality of the open research network (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., 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).

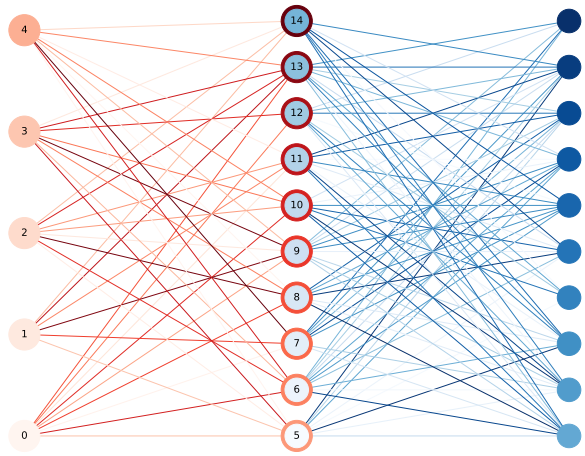


Figure 3: Robhoot in Digital Ecosystems: **Left column**: *ROBHOOT* **v.1.0** representing the research cycle as nodes from number 0 to 4: Data integration (0), Complexity Reduction (1), Inference (2), Validation (3), and Visualization(4)). **Central column**: Nodes representing the research cycle in 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 Their relevance in terms of Future and Emerging Technologies

Table 2

5.1 Scientific concept

Research cycle automation combining novel decentralization algorithms with data, inference, and knowledge graphs integration

5.2 Identified problem

Global sustainability is a major goal of humanity. Many studies have shown global sustainability could be achieved by strengthening transparency and feedbacks between social, ecological and governance systems. Sustainability goals, however, strongly depend on global society access to evidence- and research-based knowledge gaps. Yet, the science ecosystem lacks open-source technologies narrowing down the different aspects of knowledge gaps.

5.3 Potential solutions envisaged

- Global access to fully reproducible reports
- Testnet case study in Sustainability and Biodiversity research to facilitate open an global access

5.4 Describe the vision of a radically-new science-enabled

5.5 Technology that the project would contribute towards

Implementation 3.1 Research methodology and work plan – Work packages, deliverables We will use a hybrid modular approach (refs for selecting project management) containing four objectives, eleven work packages and X deliverables. The following is the research methodology and timing of the work packages and the connections among the packages within and across goals: O1: Multilayer WP1 (ADAI): Automated data acquisition and integration. Open-source ETLs are rapidly evolving towards accounting for many key aspects of data integration: Data manipulation across formats (CloverDX), merging of historical and real-time streaming data pipelines (i.e., Kafka) and data structures facilitating the storage and access of large amounts of data (i.e., clickhouse.) Our research methodology will be focused on developing an automated workflow using the geographically distributed cloud on computing and storage to test the robustness of data integration metrics across gradients of simulated data containing dimensions, biases, sizes, formats, temporal and spatial resolution (should we be more domain specific here? Or should we stay general and thinking broadly about simulated data with along complexity gradients and explore data integration metrics? How is the SDSC dealing with data integration for Renku? MapReduce Golem network Resource distribution in data storage and simulating data (Ease.ml constraints and novelty and how Fluence network can help to solve it) WP2 (PROPANCE): Process pattern automated inference. Automated classification scheme to explore many inference methods across the population of KGs. WP3 (VISUR): Intralayer visualization and reporting. We will integrate and develop new algorithms to merge distinct database into one large db using mySQL, clickhouse or similar db-open-source software. WP4 (XX) Multilayer and KG integration: Existing knowledge of neural networks (refs by Luis) and deep knowledge-based algorithms to obtain the KGs (integrating WP1-3.) automated ledger knowledge network technology

to compactly facilitate open-science, decentralization, reproducibility and security in the science ecosystem O2: Ledger WP5 (): DeepKlen consensus protocol (DCP). The ledger represents the DeepKlen universe at a given point in time. It contains the KGs list and all the orders in the distributed network. We will define and implement a protocol to which KG set to apply to the last ledger. WP6 (): Scalability, decentralization and security protocols. WP7 (): Ledger automation O3: DeepKlen WP8 (): Testnet WP9 (): Mainnet O4: Robhoot WP10 (): Testnet Showcase of the julia packages to be integrated within and between layers and the existing gaps. WP11 (): The Robhoot Open Network (RON) for Biodiversity research

5.6 Describe how this vision surpasses paradigms that currently exist

5.7 Overall and specific objectives for the project

High risk, plausibility and flexibility of the research approach
We are in need of accounting for the uncertainties, the reproducibility and immutability related to automation in science and engineering. This need is not just for a specific stage of the research cycle, but for the full research cycle, from data acquisition to reporting generation because knowledge-inspired societies and governance will demand full research cycle transparency in solving complex social, environmental and technological problems. This need brings many challenges to our research proposal because obtaining robust knowledge from integrating many parts each containing its own set of methods can generate divergent, fragile and contradictory outcomes. We will develop a flexible research method focusing more in the algorithmic robustness of the deep ledger knowledge network than in the development of robust automated knowledge generation. Our motivation will be to provide a first proof of concept of how the technology works: we will sample the KGs using different deep learning algorithms to estimate the uncertainty of the ruled-based inference obtained by fitting predictions to simulated data (Goal G1). 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 1). We 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 (Goal G2). Despite our focus will be bias towards the side of the algorithmic robustness of the deep ledger knowledge network, we will develop a domain-specific case study, our Robhoot Open Network, to test the robustness of the rule-based inference obtained by fitting each of the generated KG to the empirical patterns (Goal G3). The high risk associated to robustly automate the full research cycle for producing immutable open knowledge is buffered to a great extend because the existing ecosystem of tested and reliable open-source tools: We will combine our own algorithms (i.e., data integration and deep learning algorithms for sampling and automating the KGs) with open-source tools like Renku, Fabric and gitchain. This open-ecosystem will allow us to have a flexible launching of a testnet to collect data to explore the security-scalability-decentralization patterns and the robustness of the generated KGs in the deep ledger knowledge network (Goal G4.)

5.8 The relevant state-of-the-art and the extent of the advance the project would provide beyond it

There are currently automated platforms mostly in the private domain focusing in specific parts or one layer/one path of the research cycle (BigQuery, Modulos, Google AI, Automated statistician; Ghahramani, 2015, Ease.ml; Li, Zhong, Liu, Wu, Zhang, 2017): Novelty of reporting generation following one path or resource allocation when exploring many parts of the research cycle. The science ecosystem, however, still lack a framework automating the research cycle from end-to-end into the scalability-security-decentralization trade-offs of digital ecosystems. Many science and

engineering projects have failed in reproducibility in public-funded science and technology (refs). Yet, despite public institutions are demanding more reproducibility and openness of the data and the scientific cycle (refs), and overall a shifting towards open and reproducible scientific and engineering landscapes, there are not currently open technologies aiming to compactly facilitate and distribute the scientific and engineering cycle in immutable knowledge networks.

5.9 Science-to-technology breakthrough targeted by the project

Closing global sustainability knowledge gaps in the science of bias in neural spiking networks and automation in multilayer networks.

5.10 Interdisciplinarity

ROBHOOT aims to be a hybrid-technology accounting for many features (Tables 1 to 3). Producing such a multi-feature technology requires multidisciplinarity teams making contributions for each of the Robhoot features while integrating all these features in a rapidly evolving digital ecosystem. In this regard, the project aims to put together data and computer scientists (i.e., distributed computing, open-source software development), scientists from 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.

Technologies with the capacity to compactly account for neutral, borderless, immutable, and open-access information in hybrid, trusted-untrusted peer-to-peer interactions, accounting for the multilayer nature of science and engineering are currently not in place. Producing such a technology will require integrating expertise from disparate disciplines like multilayer networks, deep learning, automation algorithmics, and distributed technologies. The integration of these disciplines will require to go beyond domain boundaries. Specifically, we will merge scientists and engineers from data and computer science, the physics of complex systems, artificial intelligence and the biology, ecology and evolution of social, natural and technological ecosystems to develop a "de novo" technology: the synthesis of automated knowledge generation in a neutral, borderless and immutable network synthesized anew from existing open-source projects like Renku, Fabric and gitchain.

5.11 Scientific and technological contributions to the foundation of a new future technology

Peer-to-peer interactions composed by trusting and untrusting peers abound in social, economical, natural and technological ecosystems. Many studies in such systems are producing an immense gain in detailed knowledge about scalability, security and decentralization trade-offs (refs; TON network; Fabric ledger OS network). Automation and AI technologies is the other angle from which many advances are rapidly occurring. 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. Most studies about these trade-offs have considered one-level networks. Yet, information generation usually comes from the interactions within- and between-layers, and the feedbacks occurring among layers in these systems have provided new unexpected behaviors that are difficult to anticipate when exploring one layer alone (refs). In biological systems, the genetic architecture of functionally important traits feedback throughout the genotype-phenotype map producing variation in phenotypes that are functionally important to understand the evolution of genotypic and phenotypic variation like growth rates and the immune system that ultimately determine the frequency of the

phenotypes and the interaction centrality patterns in natural populations (refs). In science and engineering, many steps within- and between-layers occur to generate information (Figure 1a) Similarly to biological systems, interactions including intra- and inter-layer feedbacks are not easy to reproduce if not properly accounted for. One of the main facts when accounting for more than one layer is that the interactions and feedbacks to each other produce a dynamics that significantly differ from the one-layer approach (refs). Accounting for levels and scales in many systems using multilayer networks have provided a framework to explore how the microdynamics of peer-to-peer interactions might connect to the macroscopic properties of the ecosystem like the centralization and and the sensitivity to attacks within and between layers (refs).

Science and technology ecosystems are in need of accounting for the uncertainties, reproducibility and immutability related to the complexity of the research process (Table 1). Such needs are 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 (Tables 2 and 3). Reducing knowledge-gaps at global scales in knowledge-inspired societies bring 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 and adaptive research method focusing step by step in increasing levels of complexity (i.e., from *ROBHOOT* v.1.0 to v.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 (*ROBHOOT* v.1.0) to sample the KGs (*ROBHOOT* v.2.0) contrasting deep learning algorithms to estimate the uncertainty of the ruled-based inference obtained by fitting predictions to simulated data (*ROBHOOT* v.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* v.4.0 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.

Despite our focus will be bias towards the algorithmic robustness during the four stages of development, we will implement a domain-specific case study, a “Biodiversity, Global Change and Sustainability Research”, to test the robustness of the rule-based inference obtained by fitting the KGs 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).

5.12 Potential for future social or economic impact or market creation

The following are the general and the specific impacts according to our objectives, working packages and deliverables:

- Automated knowledge-based network technology
  The integration between open-source data integration and inference schemes, the interlayer automation (O1: Multilayer), will allow for the systematic exploration of robust knowledge-based patterns when exploring the population of KGs. This is in sharp contrast to existing AI technologies mostly oriented to prediction without knowledge-based understanding (refs). Despite open-source ETLs are rapidly evolving towards accounting for many aspects of data integration (formats, historical-real time, storage, dimensions, size, bias and spatiotemporal resolution), there is a missing component in quantifying the robustness of knowledge that integrated data can provide. Automated populations of KGs connecting cutting-edge open-

Word	Meaning
Reproducible knowledge graph	Algorithms accounting for the research cycle to make it fully transparent
Evidence-based knowledge gap	Factors limiting scientific transfer to benefit society
Research-based knowledge gap	Factors limiting access to reproducibility in science
Question-based knowledge graph	Algorithms detecting questions poorly explored but important for multidisciplinarity in science
Automation	Functionally interdependent algorithms targeting minimal human-driven interference
Knowledge-inspired society	Open access reproducible reports for global society
Neutral-knowledge generation	Open reproducible reports accounting for the many biases of the scientiic process

Table 2: Glossary of terms.

source ETLs to inference classification schemes can provide the quantification of robustness in knowledge-based patters for future predictive technologies.

- Open immutable knowledge in untrusted digital peer-to-peer ecosystems
  The open access of immutable accumulation of knowledge in untrusted digital peer-to-peer ecosystems: Social, environmental and economic impact to facilitate global access to transparent knowledge. ETLs are rapidly evolving towards accounting for many key aspects of data integration: Data manipulation across formats (CloverDX), merging of historical and real-time streaming data pipelines (i.e., Kafka) and data structures facilitating the storage and access of large amounts of data (i.e., clickhouse.) Our research methodology will be focused on developing an automated workflow using the geographically distributed cloud on computing and storage to test the robustness of data integration metrics across gradients of simulated data containing dimensions, biases, sizes, formats, temporal and spatial resolution (should we be more domain specific here? Or should we stay general and thinking broadly about simulated data with along complexity gradients and explore data integration metrics? How is the SDSC dealing with data integration for Renku? We anticipate implementation of an automated end-to-end research cycle within an open ledger to facilitate real-time open-access neutral data-rule-knowledge to gain informed decisions to help solve complex social, environmental and technological problems. This facilitation might occur for local, regional and global problems in many fronts. Specifically, open deep ledger knowledge networks might have an impact in the following five areas
- The identification of gaps in research paths not explored consequence of lack of synthesis in interdisciplinary research
  The creation of new markets opportunities obtained from exploring these gaps and the development of comparative method in the science of science and citizen data science.
- The merging of prediction and explanatory power in open science to gain synergy between AI open predictive tools and ruled-based pattern inference creating a more balanced pattern and process inference interaction. Recent examples of AI algorithms playing chess and go using brute force deep learning models or rule-based algorithms have discovered the power...: The integration between prediction and understanding power to facilitate explanatory synthesis.
- The automation of reproducible open knowledge will facilitate the reusability, repeatability, and replicability of research outputs. The open access knowledge for governance transparency.
  Measures to maximise impact a) Dissemination and exploitation of results 1. G4 will launch a

testnet to help disseminate the main results of the deep ledger knowledge network. The launch will have invited NGO's and GO across disciplines and social, economical and technological sectors. 2. The Robhoot Open network will be launched as a Biodiversity research network to integrate the existing public databases and crowdsource data collections into the automated KGs and ledger network to facilitate NGOs, GO and other organizations transparency and governance in Biodiversity management. 3. The project aims to publish its main findings in top open scientific journals to communicate the global impact of a deep ledger knowledge network for transparency and governance across social and economical sectors. b) Communication activities 1. The contribution in communication of the Swiss Data Science Center, Switzerland 2. Contribution of the Wyss center 3. Contribution of Ifisc, Spain

5.13 Create new market opportunities, strengthen competitiveness and growth of companies, address issues related to climate change or the environment, or bring other important benefits for society

6 Impact

6.1 Expected impacts

The following are the general and the specific impacts according to our objectives, working packages and deliverables:

- Automated knowledge-based network technology

  The integration between open-source data integration and inference schemes, the inter-layer automation (O1: Multilayer), will allow for the systematic exploration of robust knowledge-based patterns when exploring the population of KGs. This is in sharp contrast to existing AI technologies mostly oriented to prediction without knowledge-based understanding (refs). Despite open-source ETLs are rapidly evolving towards accounting for many aspects of data integration (formats, historical-real time, storage, dimensions, size, bias and spatiotemporal resolution), there is a missing component in quantifying the robustness of knowledge that integrated data can provide. Automated populations of KGs connecting cutting-edge open-source ETLs to inference classification schemes can provide the quantification of robustness in knowledge-based patters for future predictive technologies.
- Open immutable knowledge in untrusted digital peer-to-peer ecosystems

  The open access of immutable accumulation of knowledge in untrusted digital peer-to-peer ecosystems: Social, environmental and economic impact to facilitate global access to transparent knowledge. ETLs are rapidly evolving towards accounting for many key aspects of data integration: Data manipulation across formats (CloverDX), merging of historical and real-time streaming data pipelines (i.e., Kafka) and data structures facilitating the storage and access of large amounts of data (i.e., clickhouse.) Our research methodology will be focused on developing an automated workflow using the geographically distributed cloud on computing and storage to test the robustness of data integration metrics across gradients of simulated data containing dimensions, biases, sizes, formats, temporal and spatial resolution (should we be more domain specific here? Or should we stay general and thinking broadly about simulated data with along complexity gradients and explore data integration metrics? How is the SDSC dealing with data integration for Renku? We

anticipate implementation of an automated end-to-end research cycle within an open ledger to facilitate real-time open-access neutral data-rule-knowledge to gain informed decisions to help solve complex social, environmental and technological problems. This facilitation might occur for local, regional and global problems in many fronts. Specifically, open deep ledger knowledge networks might have an impact in the following five areas

– The identification of gaps in research paths not explored consequence of lack of synthesis in interdisciplinary research: The creation of new markets opportunities obtained from exploring these gaps and the development of comparative method in the science of science and citizen data science.
– The merging of prediction and explanatory power in open science to gain synergy between AI open predictive tools and ruled-based pattern inference creating a more balanced pattern and process inference interaction. Recent examples of AI algorithms playing chess and go using brute force deep learning models or rule-based algorithms have discovered the power...: The integration between prediction and understanding power to facilitate explanatory synthesis.
– The automation of reproducible open knowledge will facilitate the reusability, repeatability, and replicability of research outputs. The open access knowledge for governance transparency.

6.2 Measures to maximise impact

* Dissemination and exploitation of results
  1. G4 will launch a testnet to help disseminate the main results of the deep ledger knowledge network. The launch will have invited NGO's and GO across disciplines and social, economical and technological sectors.
  2. The Robhoot Open network will be launched as a Biodiversity research network to integrate the existing public databases and crowdsource data collections into the automated KGs and ledger network to facilitate NGOs, GO and other organizations transparency and governance in Biodiversity management.
  3. The project aims to publish its main findings in top open scientific journals to communicate the global impact of a deep ledger knowledge network for transparency and governance across social and economical sectors.
* Communication activities
  1. The contribution in communication of the Swiss Data Science Center, Switzerland
  2. Contribution of the Wyss center
  3. Contribution of Ifisc, Spain

7 Implementation

7.1 Research methodology and work plan – Work packages, deliverables

3. Implementation 3.1 Research methodology and work plan – Work packages, deliverables We will use a hybrid modular approach (refs for selecting project management) containing four objectives, eleven work packages and X deliverables. The following is the research methodology and timing of the work packages and the connections among the packages within and across goals: O1: Multilayer WP1 (ADAI): Automated data acquisition and integration. Open-source ETLs are rapidly evolving towards

accounting for many key aspects of data integration: Data manipulation across formats (CloverDX), merging of historical and real-time streaming data pipelines (i.e., Kafka) and data structures facilitating the storage and access of large amounts of data (i.e., clickhouse.) Our research methodology will be focused on developing an automated workflow using the geographically distributed cloud on computing and storage to test the robustness of data integration metrics across gradients of simulated data containing dimensions, biases, sizes, formats, temporal and spatial resolution (should we be more domain specific here? Or should we stay general and thinking broadly about simulated data with along complexity gradients and explore data integration metrics? How is the SDSC dealing with data integration for Renku? MapReduce Golem network Resource distribution in data storage and simulating data (Ease.ml constraints and novelty and how Fluence network can help to solve it) WP2 (PROPANCE): Process pattern automated inference. Automated classification scheme to explore many inference methods across the population of KGs. WP3 (VISUR): Intralayer visualization and reporting. We will integrate and develop new algorithms to merge distinct database into one large db using mySQL, clickhouse or similar db-open-source software. WP4 (XX) Multilayer and KG integration: Existing knowledge of neural networks (refs by Luis) and deep knowledge-based algorithms to obtain the KGs (integrating WP1-3.) automated ledger knowledge network technology to compactly facilitate open-science, decentralization, reproducibility and security in the science ecosystem O2: Ledger WP5 (): DeepKlen consensus protocol (DCP). The ledger represents the DeepKlen universe at a given point in time. It contains the KGs list and all the orders in the distributed network. We will define and implement a protocol to which KG set to apply to the last ledger. WP6 (): Scalability, decentralization and security protocols. WP7 (): Ledger automation O3: DeepKlen WP8 (): Testnet WP9 (): Mainnet O4: Robhoot WP10 (): Testnet Showcase of the julia packages to be integrated within and between layers and the existing gaps. WP11 (): The Robhoot Open Network (RON) for Biodiversity research

7.2 Management structure, milestones and procedures

- Describe the organisational structure and the decision-making (including a list of milestones (table 3.2a))
- Explain why the organisational structure and decision-making mechanisms are appropriate to the complexity and scale of the project.
- Describe any critical risks, relating to project implementation, that the stated project's objectives may not be achieved. Detail any risk mitigation measures. Please provide a table with critical risks identified and mitigating actions (table 3.2b) and relate these to the milestones.

7.3 Consortium as a whole

The individual members of the consortium are described in a separate section 4. There is no need to repeat that information here. • Describe the consortium. Explain how it will support achieving the project objectives. Does the consortium provide all the necessary expertise? Is the interdisciplinarity in the breakthrough idea reflected in the expertise of the consortium? • In what way does each of the partners contribute to the project? Show that each has a valid role and adequate resources in the project to fulfil that role. How do the members complement one another? Other countries and

international organisations: If one or more of the participants requesting EU funding is based in a country or is an international organisation that is not automatically eligible for such funding (entities from Member States of the EU, from Associated Countries and from one of the countries in the exhaustive list included in General Annex A of the work programme are automatically eligible for EU funding), explain why the participation of the entity in question is considered essential for carrying out the action on the grounds that participation by the applicant has clear benefits for the consortium.

7.4 Resources to be committed

Please make sure the information in this section matches the costs as stated in the budget table in section 3 of the administrative proposal forms, and the number of person months, shown in the detailed work package descriptions. Please provide the following: • a table showing number of person months required (table 3.4a) • a table showing 'other direct costs' (table 3.4b) for participants where those costs exceed 15% of the personnel costs (according to the budget table in section 3 of the administrative proposal forms)

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] Matías E Mastrángelo, Natalia Pérez-Harguindeguy, Lucas Enrico, Elena Bennett, Sandra Lavorel, Graeme S Cumming, Dilini Abeygunawardane, Leonardo D Amarilla, Benjamin Burkhard, Benis N Egoh, Luke Frishkoff, Leonardo Galetto, Sibyl Huber, Daniel S Karp, Alison Ke, Esteban Kowaljow, Angela Kronenburg-García, Bruno Locatelli, Berta Martín-López, Patrick Meyfroidt, Tuyeni H Mwampamba, Jeanne Nel, Kimberly A Nicholas, Charles Nicholson, Elisa Oteros-Rozas, Sebataolo J Rahlao, Ciara Raudsepp-Hearne, Taylor Ricketts, Uttam B Shrestha, Carolina Torres, Klara J Winkler, and Kim Zoeller. Key knowledge gaps to achieve global sustainability goals. *Nature Sustainability*, 2 (12):1115–1121, 2019. ISSN 2398-9629. doi: 10.1038/s41893-019-0412-1. URL https://doi.org/10.1038/s41893-019-0412-1.

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

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

[9] 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.

[10] 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/.

[11] 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.

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

[13] 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.

[14] 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.

[15] Christian Steinruecken, Emma Smith, David Janz, James Lloyd, and Zoubin Ghahramani. The Automatic Statistician. pages 161–173.

[16] Roger Guimerà, Ignasi Reichardt, Antoni Aguilar-Mogas, Francesco A. Massucci, Manuel Miranda, Jordi Pallarès, and Marta Sales-Pardo. A bayesian machine scientist to aid in the solution of challenging scientific problems. *Science Advances*, 6(5), 2020. doi: 10.1126/sciadv.aav6971. URL https://advances.sciencemag.org/content/6/5/eaav6971.