Proposal template: technical annex
Research and Innovation actions
Call: H2020-FETOPEN-2018-2020
Topic: FETOPEN-01-2018-2019-2020: FET-Open Challenging Current Thinking

The structure of this template must be followed when preparing your proposal. It has been designed to ensure that the important aspects of your planned work are presented in a way that will enable the experts to make an effective assessment against the evaluation criteria. Sections 1, 2 and 3 each correspond to an evaluation criterion.

Please be aware that proposals will be evaluated as they were submitted, rather than on their potential if certain changes were to be made. This means that only proposals that successfully address all the required aspects will have a chance of being funded. There will be no possibility for significant changes to content, budget and consortium composition during grant preparation.

⚠ A proposal that, according to the evaluators' assessments, does not convincingly satisfy all FET gatekeepers as described in the FET Work Programme will be declared out of scope.

⚠ **Page limit**: Sections 1 to 3 of the proposal should consist of a maximum of 15 A4 pages (no cover page). All tables, figures, references and any other element pertaining to these sections must be included as an integral part of these sections and are thus counted against this page limit. There is no page limit for sections 4 and 5.

The page limit will be applied automatically; therefore you must remove the first 2 instruction pages of this template before submitting.

If you attempt to upload a proposal longer than the specified limit before the deadline, you will receive an automatic warning and will be advised to shorten and re-upload the proposal. After the deadline, excess pages (in over-long proposals/applications) will be automatically made invisible, and will not be taken into consideration by the experts. The proposal is a self-contained document. Experts will be instructed to ignore hyperlinks to information that is specifically designed to expand the proposal, thus circumventing the page limit.

⚠ The following formatting conditions apply.

Each page should include a footnote with the acronym of the proposal.

The reference font for the body text of H2020 proposals is Times New Roman (Windows platforms), Times/Times New Roman (Apple platforms) or Nimbus Roman No. 9 L (Linux distributions).

The use of a different font for the body text is not advised and is subject to the cumulative conditions that the font is legible and that its use does not significantly shorten the representation of the proposal in number of pages compared to using the reference font (for example with a view to bypass the page limit).

The minimum font size allowed is 11 points. Standard character spacing and a minimum of single line spacing is to be used.

Text elements other than the body text, such as headers, foot/end notes, captions, formula's, may deviate, but must be legible.

The page size is A4, and all margins (top, bottom, left, right) should be at least 15 mm (not including any footers or headers).

Deep Knowledge Ledger Network

Deepklen :: Robhoot

1. Excellence

1.1 Radical vision of a science-enabled technology

Public funded science is highly centralized (Günther, 2018; Inhaber, 1977), prone to errors ((Fang & Casadevall, 2011)), difficult to reproduce (Hardwicke. et al., 2019)), and contains many biases (Ioannidis, 2005). Yet, despite rapid advances are occurring to make the science ecosystem less centralized and biased while increasing openness and reproducibility, there is an existing gap in science-enabled technologies automating fully the research cycle in a open, decentralized, and neutral knowledge-inspired technology. In this project we aim to build an automated ledger knowledge network technology to compactly facilitate open-science, decentralization, reproducibility and security in the science ecosystem (Figure 1). Our final aim is to provide real-time open-access neutral data-rule-knowledge to gain informed decisions and help solve complex social, environmental and technological problems.

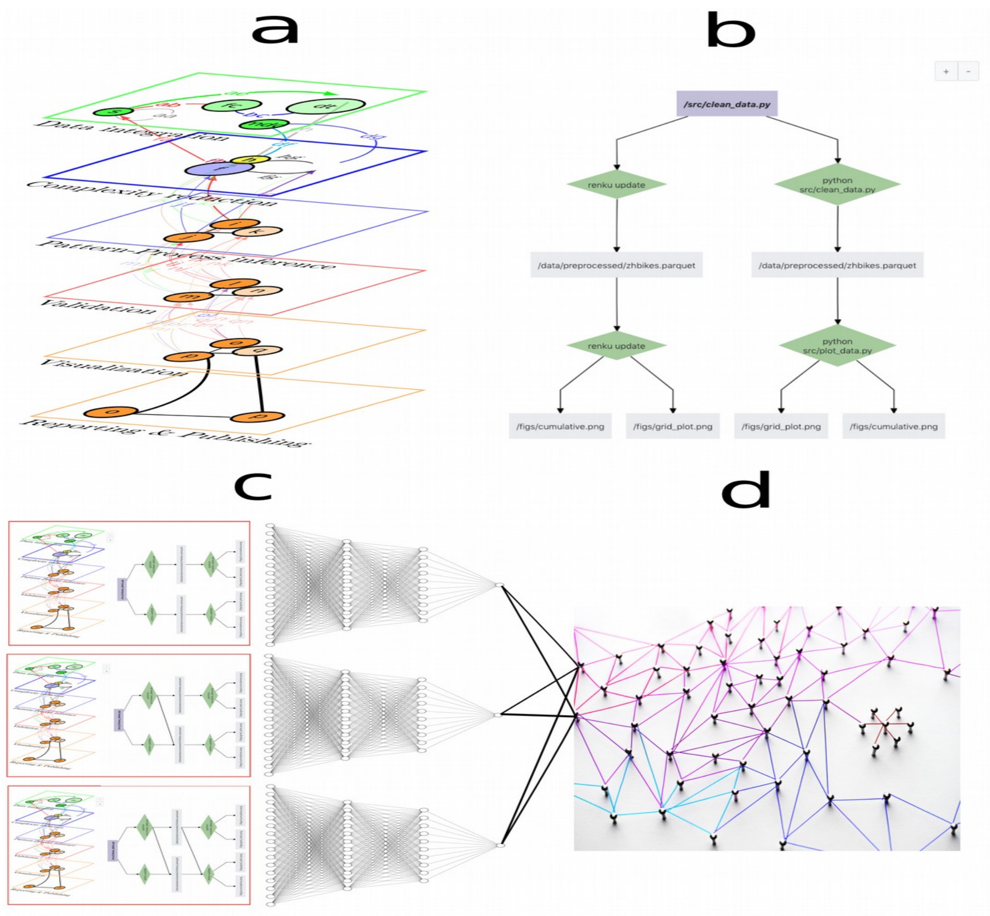

Figure 1: Deep ledger knowledge-based *network technology. **a)** End-to-end research cycle from data integration to reporting.* ***b)*** The knowledge graph (KG) tracking one research path of **a** (i.e., *Renku* open-source code). ***c)*** *Deep knowledge-based network can* automatically explore a population of KGs to gain rule-based knowledge of the data. d) The population of all KGs can be stored in a ledger distributed network of mutually trusting/untrusting peers with every peer maintaining the population of the KGs (i.e., decentralized P2P git network like *Gitchain.*)

The science ecosystem is error prone with multiple steps requiring information transfer among trusted/untrusted peers. In this ecosystem, immutable peer-to-peer architecture with a mixture of trusted-untrusted peers is key to have open access to full reports to gain informed decisions in complex societal, environmental and technological problems. The overall objectives of the proposal with the specific goals of each one are the following:

G1: Deploy an automated *knowledge-based network technology* accounting fully for the research cycle in a lineage client-tracker to produce a population of *Knowledge Graphs* (KGs) (Figures 1a and 1b.)

G1.1: Intralayer automation of data integration, inference, and validation (Figure 1a.)

G1.2: Intralayer automation of visualization and reporting generation (Figure 1a.)

G1.3: Deep intra- and inter-layer automation with a lineage client-tracker to sample paths in the multilayer network (Figure 1a.)

G2: Deploy an *end-to-end permissioned-permissionless distributed ledger technology* to guarantee decentralization, open-access and security of the KGs in the science ecosystem (Figures 1c and 1d.)

G2.1: Blockchain implementation accounting for consensus algorithms and smart contracts among trusted-untrusted peer-to-peer events.

G2.2: Scenarios to minimize scalability-security-decentralization trade-offs.

G3: *Robhoot Open Network in Biodiversity Research as a c*ase study to facilitate accessibility to open science to provide real-time open-access data-rule-knowledge to gain informed decisions and help solve local and global environmental problems.

G3.1: Biodiversity datasets integration (Figure 1a top layer.)

G3.2: Interaction with G1.2 to test Biodiversity patterns in a deep ledger knowledge network (Figure 1.)

G4: *Launch a testnet* (or prototype) to collect data to explore the security-scalability-decentralization patterns and the robustness of the generated KGs in the deep ledger knowledge network.

G4.1: Interaction with G2.2 to test the scalability-security-decentralization trade-offs.

G4.2: Interaction with other platforms/decentralized networks to facilitate real-time open-access data-rule knowledge to general case studies.

1.2 Science-to-technology breakthrough that addresses this vision

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). 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).

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.

1.3 Interdisciplinarity and non-incrementality of the research proposed

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.

1.4 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 l*aunching 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.)

2. Impact

2.1 Expected impacts

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

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

2. 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:

3. 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.

4. 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.

5. The automation of reproducible open knowledge will facilitate the reusability, repeatability, and replicability of research outputs. The open access knowledge for governance transparency.

2.2 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

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

O3: DeepKlen

O4: Robhoot

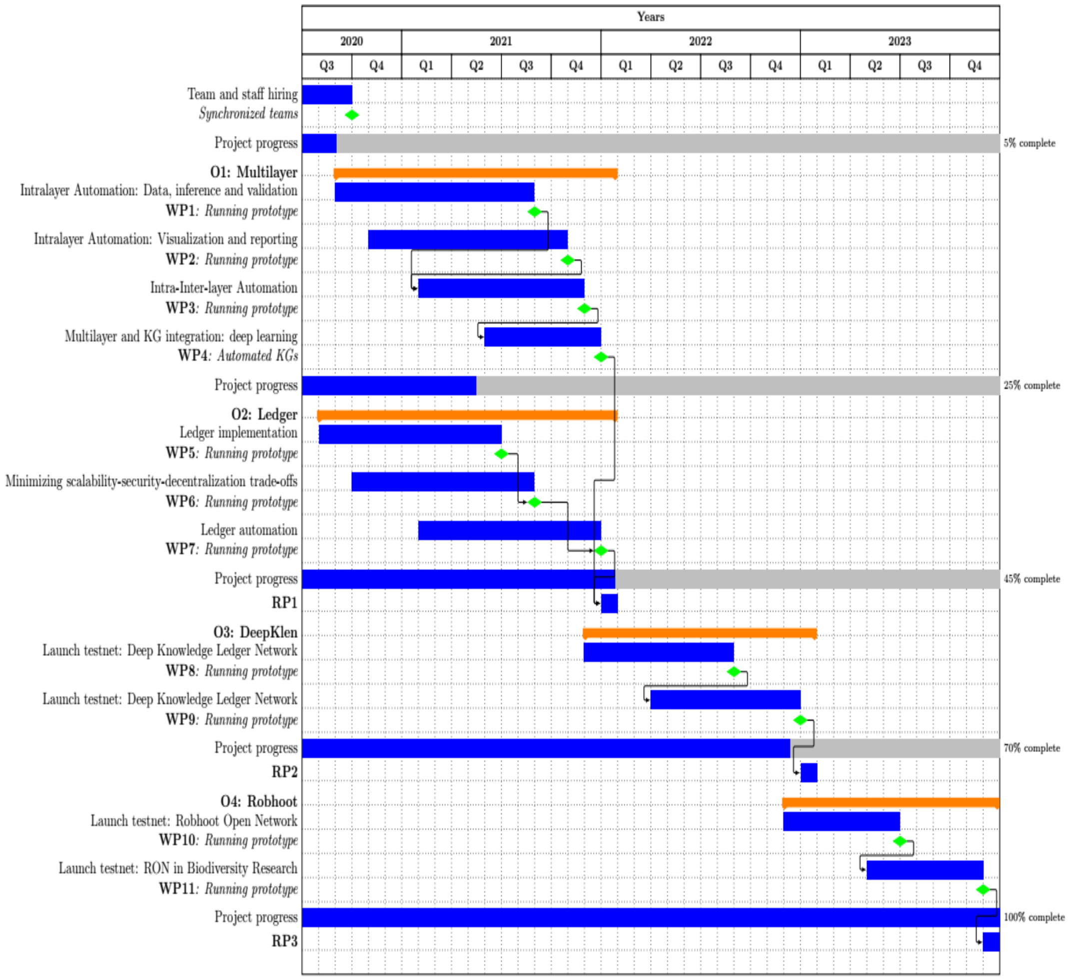

Figure 2: Gantt chart showing the structure of the work plan, the timing of the working packages and the connections among them. The total proposed length of the project is 42 months with three reporting periods (RP1: 18; RP2: 30, and RP3: 42.)

Table 3.1a: List of work packages

Work package No	Work Package Title	Lead Participant No	Lead Participant Short Name	Person-Months	Start Month	End month
WP1	Valia	1	Melian	1	3	14
WP2						
WP3						
WP4						
				Total person-months		

Proposed length of the project (months)	RP1 duration (months)	RP2 duration (months)	RP3 duration (months)
48	12	18	18
42	12	12	18
36	12	24	
30	12	18	
24	12	12	

- detailed work description, i.e.:
  - a list of work packages (table 3.1a);
  - a description of each work package (table 3.1b);
  - a list of all deliverables (table 3.1c);
- Graphical presentation of the work packages showing how they inter-relate (Pert chart or similar).

⚠ *Give full details. Base your account on the logical structure of the project and the stages in which it is to be carried out. The number of work packages should be proportionate to the scale and complexity of the project.*

⚠ *You should give enough detail in each work package to justify the proposed resources to be allocated and also quantified information so that progress can be monitored, including by the Agency*

⚠ *Resources assigned to work packages should be in line with their objectives and deliverables. You are advised to include a distinct work package on 'management' (see section 3.2) and to give due visibility in the work plan to 'dissemination and exploitation' and 'communication activities', either with distinct tasks or distinct work packages.*

⚠ *You will be required to include an updated (or confirmed) 'plan for the dissemination and exploitation of results' in both the periodic and final reports. This should include a record of activities related to dissemination and exploitation that have been undertaken and those still planned. A report of completed and planned communication activities will also be required.*

⚠ *If your project is taking part in the Pilot on Open Research Data, you must include a 'data management plan' as a distinct deliverable within the first 6 months of the project. A template for such a plan is given in the guidelines on data management in the H2020 Online Manual. This deliverable will evolve during the lifetime of the project in order to present the status of the project's reflections on data management.*

Definitions:

'Work package' means a major sub-division of the proposed project.

'Deliverable' means a distinct output of the project, meaningful in terms of the project's overall objectives and constituted by a report, a document, a technical diagram, a software etc.

3.2 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.

Definition:

'Milestones' means control points in the project that help to chart progress. Milestones may correspond to the completion of a key deliverable, allowing the next phase of the work to begin. They may also be needed at intermediary points so that, if problems have arisen, corrective measures can be taken. A milestone may be a critical decision point in the project where, for example, the consortium must decide which of several technologies to adopt for further development.

3.3 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.

3.4 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)

Tables for section 3.1

Table 3.1b: Work package description

For each work package:

Work package number		**Lead beneficiary**					
Work package title							
Participant number							
Short name of participant							
Person months per participant:							
Start month			**End month**				

Objectives

Description of work (where appropriate, broken down into tasks and deliverables), lead partner and role of participants.

Deliverables (brief description and month of delivery)

Table 3.1c: List of Deliverables[1]

1 If your action is taking part in the Pilot on Open Research Data, you must include a data management plan as a distinct deliverable within the first 6 months of the project. This deliverable will evolve during the lifetime of the project in order to present the status of the project's reflections on data management. A template for

Deliverable (number)	Deliverable name	Work package number	Short name of lead participant	Type	Dissemination level	Delivery date (in months)

KEY

Deliverable numbers in order of delivery dates. Please use the numbering convention <WP number>.<number of deliverable within that WP>.

For example, deliverable 4.2 would be the second deliverable from work package 4.

Type:
Use one of the following codes:
R: Document, report (excluding the periodic and final reports)
DEM: Demonstrator, pilot, prototype, plan designs
DEC: Websites, patents filing, press & media actions, videos, etc.
OTHER: Software, technical diagram, etc.

Dissemination level:
Use one of the following codes:
PU = **Public,** fully open, e.g. web
CO = **Confidential, r**estricted under conditions set out in Model Grant Agreement
CI = **Classified,** information as referred to in Commission Decision 2001/844/EC.

Delivery date
Measured in months from the project start date (month 1)

such a plan is available in the H2020 Online Manual on the Participant Portal.

Tables for section 3.2

Table 3.2a: List of milestones

Milestone number	Milestone name	Related work package(s)	Due date (in month)	Means of verification

KEY

Due date
Measured in months from the project start date (month 1)

Means of verification
Show how you will confirm that the milestone has been attained. Refer to indicators if appropriate. For example: a laboratory prototype that is 'up and running'; software released and validated by a user group; field survey complete and data quality validated.

Table 3.2b: Critical risks for implementation

Description of risk (indicate level of likelihood: Low/Medium/High)	Work package(s) involved	Proposed risk-mitigation measures

Definition critical risk:
A critical risk is a plausible event or issue that could have a high adverse impact on the ability of the project to achieve its objectives.

Level of likelihood *to occur:* Low/medium/high
The likelihood is the estimated probability that the risk will materialise even after taking account of the mitigating measures put in place.

Tables for section 3.4

Table 3.4a: Summary of staff effort

Please indicate the number of person/months over the whole duration of the planned work, for each work package, for each participant. Identify the work-package leader for each WP by showing the relevant person-month figure in bold.

	WP1	WP2	WPn	Total Person-Months per Participant
Participant Number/Short Name				
Participant Number / Short Name				
Participant Number / Short Name				
Total Person				**Total Person Months**

Months per WP				

Table 3.4b: ‘Other direct cost’ items (travel, equipment, other goods and services, large research infrastructure)

Please complete the table below for each participant when the sum of the costs for Travel, Equipment, and Other goods and services exceeds 15% of the personnel costs for that participant (according to the budget table in section 3 of the proposal administrative forms).

Participant Number/Short Name	Cost (€)	Justification
Travel		
Equipment		
Other goods and services		
Total		

Please complete the table below for all participants that would like to declare costs of large research infrastructure under Article 6.2 of the General Model Agreement[2], irrespective of the percentage of personnel costs. Please indicate (in the justification) if the beneficiary’s methodology for declaring the costs for large research infrastructure has already been positively assessed by the Commission.

Participant Number/Short Name	Cost (€)	Justification
Large research infrastructure		

2 Large research infrastructure means research infrastructure of a total value of at least EUR 20 million, for a beneficiary. More information and further guidance on the direct costing for the large research infrastructure is available in the H2020 Online Manual on the Participant Portal.