Deliverable D2.1

First analysis of cost elements for the setup of 1+MG infrastructure

Project Title Grant agreement no	**Genomic Data Infrastructure** Grant agreement 101081813		
Project Acronym (EC Call)	GDI		
WP No & Title	WP2: Long term sustainability		
WP Leaders	Regina Becker (5. UNILU), Troels Rasmussen (13. NGC)		
Deliverable Lead Beneficiary	13. NGC		
Contractual delivery date	31/01/2024	**Actual delivery date**	29/02/2024
Delayed	[Yes]		
Partner(s) contributing to deliverable	NGC, ISCIII, HRI		
Authors	Troels Rasmussen, NGC, DK		
Contributors	Rob Hooft, Health-RI, NL		
Acknowledgements	N/A		
Reviewers	Regina Becker, PNED, LU, Tommi Nyrönen FI		

GDI project receives funding from the European Union's Digital Europe Programme under grant agreement number 101081813.

Log of changes

Date	Mvm	Who	Description
27/11/2023	0V1	Troels Rasmussen, NGC	First Draft
13/02/2024	0V2	Mercedes Rothschild Steiner (ELIXIR Hub)	Copy circulated to the GDI-MB for review
20/02/2024	0V3	Troels Rasmussen, NGC	Comments addressed
29/02/2024	1V0	Mercedes Rothschild Steiner (ELIXIR Hub)	Final version submitted to the EC portal

Table of contents

Contents

GDI project receives funding from the European Union's Digital Europe Programme under grant agreement number 101081813.

GDI project receives funding from the European Union's Digital Europe Programme under grant agreement number 101081813.

1. Executive Summary

Long-term financial sustainability is vital for any data infrastructure – but it is particularly true for infrastructures like 1+MG that depend on the accumulation of data over time to function. To accommodate the sensitive nature of the data, the computing infrastructure needs to be made available to users in secure processing environments and there needs to be sufficient storage in place. Coordinating and aligning compute hardware infrastructure, the data and services across many stakeholders and countries is challenging. The infrastructure will need to be able to support a large number of different users, working in research, health care and innovation.

This deliverable provides the first step in building the financial underpinnings of this understanding: it defines the functional elements across the infrastructure with their independent functions in operating the infrastructure, and provides lists of cost items for each of these functions including how they scale with regard to the size of the infrastructure. Future work will focus on possible sources of financing that can be appealed to for the maintenance of each of these functions.

Context

Pillar I of GDI is related to Long-Term Sustainability. It includes Work Package 2 which addresses the following objectives related to financial sustainability:

- To determine the cost items associated with setting up, operating, maintaining and further developing a genomic infrastructure
- To identify and evaluate long-term sustainability options and business model for the 1+MG infrastructure covering different aspects of the necessary investment and cost and at various levels (national and central).

Work Package 2 includes three Tasks in this domain:

T2.1 - Cost of the Infrastructure - Determination of the costs associated with the Infrastructure (i.e, IT costs, data hosting costs, data use costs, costs related with general coordination and communication activities), both at national and central level.

T2.2 – Funding sources - Evaluation of the possible funding sources for the costs determined in T2.1 taking into account different stakeholders, including MSs, EC, industry, RIs and other international organisations such as the European Medicines Agency (EMA). Input from relevant sustainable structures (RIs, Joint Actions, projects,...) will be taken into account

T2.3 - Sustainability model & business model - Evaluation of sustainability models to be applied to the 1+MG Infrastructure. Provision of recommendations on long-term sustainability by following a consensus building process among MSs & other relevant stakeholders.

GDI project receives funding from the European Union's Digital Europe Programme under grant agreement number 101081813.

2. Contribution towards project outcomes

With this deliverable, the project has reached or the deliverable has contributed to the following project outcomes:

	Contributed
Outcome 1 Secure federated infrastructure and data governance needed to enable sustainable and secure cross border linkage of genomic data sets in compliance with the relevant and agreed legal, ethical, quality and interoperability requirements and standards based on the progress achieved by the 1+MG initiative.	No
Outcome 2 Platform performing distributed analysis of genetic/genomic data and any linked clinical/phenotypic information; it should be based on the principle of federated access to data sources, include a federated/multi party authorisation and authentication system, and enable application of appropriate secure multi-party and/or high-end computing, AI and simulation techniques and resources.	No
Outcome 3 Clear description of the roles and responsibilities related to personal data and privacy protection, for humans and computers, applicable during project lifetime and after its finalisation.	No
Outcome 4 Business model including an uptake strategy explaining the motivation, patient incentives and conditions for all stakeholders at the different levels (national, European, global) to support the GDI towards its sustainability, including data controllers, patients, citizens, data users, service providers (e.g., IT and biotech companies), healthcare systems and public authorities at large.	Yes

GDI project receives funding from the European Union's Digital Europe Programme under grant agreement number 101081813.

Outcome 6 Communication strategy – to be designed and implemented at the European and national levels.	No
Outcome 7 Capacity building measures necessary to ensure the establishment, sustainable operation, and successful uptake of the infrastructure.	No
Outcome 8 Financial support to the relevant stakeholders to enable extension, upgrade, creation and/or physical connection of further data sources beyond the project consortium or to implement the communication strategy and for capacity-building.	No
Outcome 5 Sustained coordination mechanism for the GDI and for the GoE multi-country project launched in the context of the 1+MG initiative.	No

GDI project receives funding from the European Union's Digital Europe Programme under grant agreement number 101081813.

3. Methods and framework

In the context of this deliverable we focus on mapping cost items that fit into a **financial model.** A financial model is a tool used to project the infrastructure's financial performance over a specific period of time. It includes financial projections such as revenue, expenses and cash flow. A financial model helps organisations to understand the financial implications of its business model and to make strategic decisions based on financial data.

It is important to note that at this point in the project we are not producing a full financial plan: we are at this stage identifying the various **cost elements** that go into a financial plan. In some cases, we are able to make qualified assumptions about the costing level of the different cost items. In others we can present a unit cost, that will give an indication of the scaling of costs over time and activity. In some cases, it is only possible to identify a cost item at this point, the cost of which will be subject to how the infrastructure is organised and scales. Estimating the cost level for such items will require discussion with other parts of the project (such as the use cases and technical implementation) and external stakeholders (such as the members of the 1+MG initiative).

This deliverable is meant to inform the discussion towards a business model, but not to present one. A business model describes how the infrastructure creates value for its users and generates revenue. It includes strategy, potential users, value proposition, revenue streams, cost structure, and key activities and resources required to execute the strategy. Essentially, a business model is a blueprint for how an organisation will operate and be sustainable.

3.1 Cost estimation framework

The StR-ESFRI study on Guidelines on cost estimation of Research infrastructures[1] contains a methodology which we will use to determine the central costs of the 1+MG infrastructure. In the rest of this section we will work out the principles of the methodology concretely (in summary fashion) for the 1+MG infrastructure. We will build on this to work out the financial model in section 4.

1) Define the unit of analysis.

The adopted methodology requires that the infrastructure is split up into units for which the financing is to be determined separately. This obviously separates the roles taken by different legal entities, but within a single legal entity it can be useful to separate different functions of the infrastructure into separate units of analysis.

1+MG infrastructure is planned as a federated infrastructure with a central Hub and national nodes. The scope for our financial model is both the Hub and the national nodes. The central hub is foreseen to be running the bulk of the work supporting data access requests, including e.g. the operation of the central discovery portal and the employment of members of the Data Access Committee. Nodes each consist of a National Coordination Point (NCP) either as a single organisation or as member of a group of contributing organisations, supporting the inclusion of

[1] https://www.esfri.eu/latest-esfri-news/new-study-guidelines-cost-estimation-research-infrastructures-str-esfri

GDI project receives funding from the European Union's Digital Europe Programme under grant agreement number 101081813.

data, the storage of data, and one or more Secure Processing Environments (SPE, often part of already existing research infrastructures) to enable data re-use. National nodes themselves may choose to organise as a federation of regional or institutional sub-nodes; that distinction is out-of-scope for our analysis (but we will refer to this in a few places).

2) Adopt a long-time horizon.

The adopted methodology requires cost estimates to be related to the entire lifecycle of the infrastructure, which means considering the costs spanning the entire period of time during which the facility remains useful. Total costs include both investment and operating costs.

When an infrastructure is built from scratch, there is usually a relatively large investment peak during design, preparation and construction.

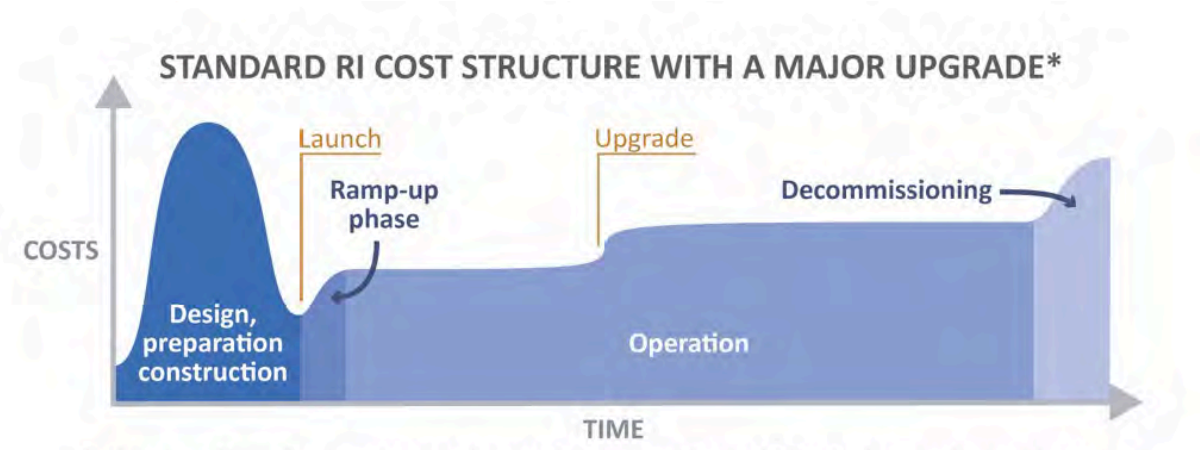

Figure 1. Standard RI cost structure with a major upgrade

The 1+MG infrastructure started with the initiative in 2018, and the GDI project (2022-2026) will take it through its Launch and initial ramp-up. Upgrades in the context of 1+MG could be changes of data standards applied in the infrastructure, such as a change of reference genome. Our financial model will need to have different operational models depending on the expected rate of growth of the total data volume handled by the infrastructure: the costing during the operational phase could be less flat than for the prototypical infrastructure in Figure 1. Decommissioning time is hard to predict, but in any case decommissioning costs for the 1+MG infrastructure are expected to be small, and they will be disregarded for this deliverable.

GDI project receives funding from the European Union's Digital Europe Programme under grant agreement number 101081813.

3) Fix the start date.

The time horizon starts the year when the first resources are deployed (cash or in-kind) for the design and preparation phase of the RI. For 1+MG infrastructure, we choose the date in which the 1+MG infrastructure is formally established as an EDIC, as the reference point for costs that are in the preparatory phase and in the operational phase.

4) Fix the base year.

The base year for the financing of the 1+MG infrastructure is 2024. This is the point-in-time when this cost estimation is made. Past and future costs in the financial model will be represented as present value in 2024.

5) Costs should be expressed in real terms.

Prices must be constant at the base year: future costs are forecasted according to realistic assumptions and net of inflation while past costs will be converted into base year value by applying the inflation index.

6) Only cash outflows are reported.

The cost accounting follows a cash flow method. Depreciation, reserves and other accounting items that are usually reported in balance sheets will not be included. Sources of financing can be used to identify cost items but shall not be mixed or added to them.

7) In-kind contributions must be included.

An in-kind contribution is a contribution of a good or a service other than money. Many RIs may rely on some forms of in-kind support. This can be related to the use of donated scientific equipment or the exploitation of machine time or personnel costs. Although such arrangements correspond to the use of real resources, they do not appear in the budgetary cost as a cash flow of the RI (but can appear in the budget of the donating/participating partner institution). They are however relevant costs and will be considered at their current market price.

8) Costs will be expressed in Euro.

9) Costs must distinguish between investment costs and operating costs.

In addition to the requirements of the followed methodology, in our financing model we classify the costs in different ways:

- Investment costs vs operating costs as following the methodology whereby investment costs are incurred once and provide value to the infrastructure for a longer period of time, and operating costs are costs that are incurred every time a certain event takes place (e.g. monthly, or for every user, or for every included data set).
- Their relevance to different phases in the operation including upgrade and decommissioning.

GDI project receives funding from the European Union's Digital Europe Programme under grant agreement number 101081813.

- Development vs operations vs organisational costs, following an analysis by Roos et al[2]

10) Total costs must be calculated at present value.

Future costs are discounted while past costs are capitalised (in addition to inflation, as explained at point 5) with an appropriate discount factor. Year of reference is 2024 (see also point 4).

[2] https://docs.google.com/document/d/1Ng0G7XnMQwjY1Kax5fgK0w4aME-SUUeRWEFqhMLR5LI/edit?usp=sharing

GDI project receives funding from the European Union's Digital Europe Programme under grant agreement number 101081813.

4. Cost framework for 1+MG EDIC

To get to an estimation and also a monitoring of the infrastructure costs, it is important to have the commitment from the national nodes to share and facilitate financial information and establish a procedure to define which information will be analysed and monitored. To gather all the data that feeds into the overall financial model is not in scope at this point as it would be time-consuming, and fruitless without a better understanding of the distribution of tasks and business model. At the initial stages, we focus on separating the different costs into Cost Units – AND we will discuss some of the challenges and opportunities that should feed into the overall financial model. This is particularly important, since the financial model needs to enable long term sustainability - which can only happen if we incorporate the extremely dynamic nature of the area in which the infrastructure is operating.

Below are the overall Cost Units that have been identified. What we focus on is what the functional elements of the infrastructure value chain are across the EDIC.

Hub: Central coordination and operation

- Hub coordination and management
- Operation of central data search and request services
- Handling of data access requests
- Further development of the central infrastructure

Node: National coordination and operation, Cost Units

- National coordination and management
- Data inclusion (including storage)
- Handling of data access requests
- Serving user-requests (computing)
- Further development of the national infrastructure

Each of these Cost Units contains several cost components that are required in order to deliver services. These will be presented in the following sections. Given that this report is the initial cost report, it is likely that additional cost elements will be added (or removed), as the project develops. The important part is that the Cost Units are exactly that separate units - areas of operation that need to function as a whole, within an overall organisational framework AND within the overall life cycle of the infrastructure. Furthermore, we categorise Cost units into three overarching activity areas, using Roos et al' s categorization:

GDI project receives funding from the European Union's Digital Europe Programme under grant agreement number 101081813.

1. **Development of Services:** Any activity that can change what is exactly offered to users. Typically project based funding coupled with in-kind contribution (hours).
2. **Operational Costs:** Any activity that is needed to offer the services to external customers. Capital investment and operational staff - sensitive to level of activity.
3. **Organisational Costs:** Any activity to maintain an organisation that runs service development and operations. Long term operational budget. Some sensitivity to activity.

Distributing costs into Cost Units and into areas of activity that provides a foundation for establishing funding model and business model for the overall infrastructure. This foundation relies on the following principles:

Matching activity with funding sources: It is crucial to match the costs and activities with the requirements and framework from different funding sources.
Timing of activity and costs: Funding must fit the different phases of the RI's lifecycle, be it one-time capital investments or in-kind contribution, project based developmental work or long term operational budget.
Marginal cost sensitivity: Some costs - mainly operating the infrastructure - are activity sensitive: costs increase with the level of activity; the activities have a high marginal cost. If funding for such activity is not linked to marginal costs, there is an inherent risk of "success syndrome" - an inability to expand as activity grows.
Distributing costs and duties between different stakeholders: Focusing on different cost elements within cost units allows for a transparent discussion on the distribution of tasks between the central and local level. In a federated system, it is not unlikely that some stakeholders, while being part of a local node, may indeed provide operational services on behalf of the central hub, and also that the hub will have coordination duties towards development and operational activities on node level.
Creating transparency on dependencies: There are likely resource dependencies outside the control of the infrastructure itself. These cost items are important to address, even if they are not part of the direct finance model. Examples could be national efforts to develop Personalized Medicine, adaptation towards EHDS and costs related to data and data quality and hardware investments, legal and organisational aspects – that might influence 1+MG or be co-designed to accommodate the 1+MG infrastructure.

4.1 Central Hub cost units

In the following we will list the different activity elements for the different cost units, and separate them into phases and activity categories. It represents a long list of cost elements within the overarching cost units, as it can be envisioned at this stage in the project. There are likely to be changes made to this over time, as the project develops and a more detailed cost mapping can be done in cooperation with the GDI partners.

GDI project receives funding from the European Union's Digital Europe Programme under grant agreement number 101081813.

Organisational costs - Hub Coordination and Management

Activities required to maintain the 1+MG central organisation. Some marginal cost sensitivity.

Cost Element	**EDIC life cycle phase**	**Comment**
General management	Construction + Operation	Leadership and core staff needed before legal entity and budget in place.
Stakeholder Engagement / Communication	Construction + Operation	Same
Financial and legal management	Construction + Operations	Central Hub must be able to maintain a positive cash flow and have a working capital for the initial ramp-up phase. Hub is legally liable for its operation.
Fundraising and project management	Operations	Marginal costs: scale of fundraising.
Information Security monitoring/management	Construction + Operations	Marginal Cost: Level of activity and number of nodes. Change management. Should ideally begin operations before users are allowed on the RI.
Administration, housing, travel	Operations	
Documentation system	Construction+ Operations	System to manage contracts, credentials, fees, approvals, and must allow national level veto and resource entitlement. The documentation system must be able to monitor progress, licences, AAI eligibility check as well as an automated scanning/flagging tool.
Data subject information portal	Construction + Operations	Interaction and role of RI and National systems needs to be discussed.
Quality assurance - Services	Operations	Marginal costs: level of activity and number nodes. Discussion: which expertises does the central hub need to engage with nodes and users? (Data management, Life science, HPC, clinical services).

GDI project receives funding from the European Union's Digital Europe Programme under grant agreement number 101081813.

Operational costs - Central data and user services

Activities needed at central level to offer the services to external users. This entails both capital investments, operational resources and staff. Some cost elements have a high marginal cost.

Cost Element	EDIC Lifecycle phase	Comment
User Portal	Operations + Upgrade	Improving the user portal will require constant development work - cooperating with SPE's.
Capacity building and Training	Operations	Centrally coordinated activities towards users, data providers, SPE specialists.
Metadata catalogue	Operations	Distribution of roles between Node and Hub to be discussed.
AAI management	Operations	Must follow the requirements and developments throughout Europe on national and/or EU level data passports.
Help Desk	Operations	Very activity sensitive - Depends on the number of users and the effectiveness of information system. EDIC needs to ensure users get access to help with their inquiries at the right level. Ecosystem from central to local level of experts on ELSI, technical and data matters.
DAC Review	Operations	Very activity sensitive - scales with the number of requests. Requires Ethics and legal competence - must interact with national level.
Resource management and ticketing	Operations	Allocation of Help desk, data and HPC/Storage resources with local SPE's - marginal cost sensitive.

Development of Services - Central infrastructure

Activities to improve service provision to users - Since these costs are the foundation for the operational activities of the Central Hub fitting these activities into the RI life cycle phase is extremely important. Some elements may require integration with national/EU resources.

Cost Element	EDIC Life cycle Phase	Comments
Clinical Use interface	Upgrade	Extension of the infrastructure from a pure research infrastructure in the first instance to add services for clinical goals (e.g. diagnostics) by medical staff. This is a significant upgrade project - new users, new interfaces and legal framework and risk management.
Automation of User interface	Upgrade	Scaling the infrastructure will depend on the ability to develop automation.

GDI project receives funding from the European Union's Digital Europe Programme under grant agreement number 101081813.

User Portal	Construction	Must be operational from day one - and will need continued development resources.
Documentation system	Construction	System to manage contracts, credentials, fees, approvals, and must allow national level veto and resource entitlement. The documentation system must be able to monitor progress, licences, AAI eligibility check as well as an automated scanning/flagging too.
AAI management system	Construction + upgrade	Development needs to be aligned with EU development of tools, and integration of clinical users eventually. Must follow the requirements and developments throughout Europe on national and/or EU level data passports.
Resource management and ticketing	Construction	Needs further discussion on how users get allocations - including cost recovery /Business models and interface with national nodes.
Information Security monitoring/management system	Construction + Upgrade	HUB must build a certified Information security management system as a legal entity and data controller. The federated infrastructure and data processing is a complicating factor. Significant investment. Interface and roles at national node level to be investigated.

4.2 Cost units with Local/national nodes

Organisational costs - National coordination and management

Activities required to maintain the 1+ MG national node organisation. Some marginal cost sensitivity

Cost Element	**EDIC Life cycle Phase**	**Comment**
General management	Construction + Operations	Leadership and core staff needed before legal entity and budget in place.
Stakeholder Engagement / Communication	Construction + Operations	Same
Financial and legal management	Construction + Operations	National node must be able to maintain a positive cash flow and have a working capital for the initial ramp-up phase. Legal liability to be identified - service provisioning, Data protection, SLA. Legal person.
Fundraising and project management	Operations	Marginal costs: scale of fundraising - distribution of roles with HUB and national stakeholders to be discussed.

GDI project receives funding from the European Union's Digital Europe Programme under grant agreement number 101081813.

Quality assurance - Services	Operations	Marginal costs: level of activity and number nodes. Discussion: which expertises is needed at national level to engage with nodes and users? (Data management, Life science, HPC).
Communication and engagement	Operations	Support and dialogue with data providers and data subjects interface Central Hub – data, users, ISM, etc.
Information security and Data Protection System	Construction + Operations	Requirements / SLA according to overarching EDIC Data Protection Impact Assessment and Risk management system.
Consent management - citizens	Construction + Operations	Interaction and role of EDIC and National systems needs to be discussed.
Data subject information portal	Operations	Interaction and role of EDIC central level and national systems needs to be discussed.

Organisational costs related to data inclusion (including storage) at national level

Cost related to building up 1+MG data resources. Data storage costs are directly linked to data volumes. Quality requirements and controls will significantly impact costs. Long term funding stability required to ensure capacity build up. It is likely to be organised differently at national level depending on interface and modus with data holders. It is also likely to depend on resources not directly under EDIC control.

Cost Element	**EDIC Life cycle Phase**	**Comment**
Data quality control and approval	Operational + ramp up	Requires discussion on scope and requirements vs. reality. Likely to be a gradually increased effort - particularly important towards clinical users.
Standardisation Framework, Monitoring of uptake	Operational	Same as above + adaptation of GA4GH standards nationally.
Guidelines, training and capacity building	Operational	Key element to engage with data holders, particularly within health care. Requires input from GOV partners on interface. [Guidelines and training materials t in English may be provided centrally]
Semantics: tools for inclusion of present and future data types	Operational + ramp up	Support function to help data holders towards inclusion of data. [Tools may also be provided centrally]
Persistent Identifier Archive	Operational	Requires discussion on FAIR and reusability of sensitive data sources.
Long term data	Operational +	FEGA requirement - Storage + Off-site backup. Extremely Cost sensitive

GDI project receives funding from the European Union's Digital Europe Programme under grant agreement number 101081813.

storage	Upgrade + decommissioning	scales with total amount of data and retention time, pending data inclusion parameters and productivity. Major upgrade expected towards clinical data, Needs to address continuity of data sources in case of Decommissioning, node exit etc. Discussion on distribution of data storage facilities at national level required - many solutions available.
Maintaining Metadata catalogue	Operation	Distribution of roles between Node and Hub to be discussed.

Organisational costs related to User Services

Costs related to user services that precedes actual usage of resources. Low marginal costs but high initiation costs.

Cost Element	**EDIC Life cycle Phase**	**Comment**
AAI - User licencing and administration	Operations	Subject to data protection requirements and EU development.
Service Ticket management	Operations	Needs further discussion on how users get allocations - including cost recovery /Business models and interface with national nodes.
1+MG Software tools operation and upgrading	Operations and upgrade	Not development - but continued maintenance and update of software kit. Life cycle management- replacement/depreciation, Devops.
Tool Kit API integration on platform	Construction + Operational	As above -
AAI management system	Construction + upgrade	Development needs to be aligned with EU development of tools, and integration of clinical users eventually. Must follow the requirements and developments throughout Europe on national and/or EU level data passports.

Operational costs - SPE Serving user requests (computing)

Servicing users on the SPE platform is the most marginal cost sensitive cost unit of the RI. Likely to rest on considerable national funding and/or in kind contribution. Scaling according to the number of users and use cases will present a challenge.

Cost Element	**EDIC Life cycle Phase**	**Comment**
Computational capacity	Operations + Upgrade	Extreme activity sensitive - Capacity needs to scale with users, data and job types. Scalability in a European context may be hard to plan for, and ensure transparent cost

GDI project receives funding from the European Union's Digital Europe Programme under grant agreement number 101081813.

		recovery. Expanded in the next chapter.
SPE operational hot storage	Operations	Extreme activity sensitive - Capacity needs to scale with users and need.
User data Ingest platform	Operation	Connectivity from long term storage and users own data to hot storage, subject to information security protocols.
Results extraction platform	Operation	Results and users own data extraction platform, subject to information security protocols.
1+MG Software kit operation and upgrading	Operations	
User Interface and resource management	Construction + operation	Activity sensitive , related to number of users and service requests.

Development of services of the national infrastructure

This is a long list of individual Development tracks, which is considered under one Cost Unit in order to serve an overarching discussion on timing, and per-project funding.

Cost Element	**RI Life cycle Phase**	**Comment**
HPC Next generation	Upgrade (major)	Long term perspective - moving life science into the exascale computing domain. Training of Life science algorithms towards next generation computing.
Data and metadata models	Construction + Ramp up	The availability of data and agreement on metadata will decide the level of ambition.
Clinical Use interface	Upgrade	National component of the developments towards serving users in clinical care. This is a significant upgrade project - new users, new interfaces and legal framework and risk management.
Automation of User interface	Upgrade	Scaling the infrastructure will depend on the ability to develop automation.
Resource management and ticketing	Construction	Part of GDI project REMS product - Tuning towards requirements.
Information Security monitoring/management system	Construction + Upgrade	HUB must build a certified Information security management system as a legal entity and data controller. The federated infrastructure and data processing is a complicating factor. Significant investment. Interface and roles at national node level to be investigated.
AAI management system	Construction + upgrade	Development needs to be aligned with EU development of tools, and integration of clinical users eventually. Must follow the requirements and developments throughout

GDI project receives funding from the European Union's Digital Europe Programme under grant agreement number 101081813.

		Europe on national and/or EU level data passports.

5. Initial discussion on cost elements

In this section some of the key cost factors as well as needed discussion at GOV level are presented. They relate to the funding requirements of the cost units, but also highlight some of the baseline cost calculations that need to go into considerations regarding scaling key components of the infrastructure.

5.1 Central Hub Staffing

Salary costs estimation: Initial considerations One FTE with social costs: Variation in salary depends on the staff categories, the general salary level of the country/legal entity in which the staff is employed as well as pension rates and how other social costs are calculated. For budgeting purposes, it is assumed that an employer must add other costs (social costs, pensions, etc) at a rate of 25-45% of the net salary).

Central Hub staffing: While it is difficult to estimate the number of staff (FTE's) at this point, it is likely that the central staffing requirement of the 1+MG EDIC is comparable to other existing data- and technology-infrastructures built within the ESFRI framework, which typically have a central staffing of 25-35 FTE's - some of which are funded directly by the consortium and others being funded through grants. There is no reason to assume that the requirements will not be equal or likely even higher with the EDIC, at a point in time where the EDIC has reached a mature operational stage, managing a considerable number of user requests across health care and research. The EDIC is situated in an area that is complex to manage and coordinate on matters like ELSI, Secure Processing Environments, Data Inclusion, Access Management and Data Protection which will require significant expertise. It should also be noted, that whereas other RI's typically seek to provide access to either data resources OR technical platforms, the 1+MG EDIC needs to do both under stringent data protection measures catering to a very broad spectrum of users within research, healthcare and innovation, in a federated delivery system.

At some point in the construction phase there will need to be dedicated staff that can bring the preparatory work into operation – particularly bringing the stakeholders together to negotiate the details that goes into establishing a new public entity, build up capacity and act on behalf of the project, even before there is an organisation in place and secure funding and sign on from key stakeholders – this requires funding even before the infrastructure is operational.

In the transition from construction to ramp-up towards operational stage the funding requirement for staff will increase towards full operational stage. Securing funding for at least 10 years that allows for the infrastructure to grow will be crucial.

The high level of coordination and build-up of capacity required, will likely mean that ***there is a very steep ramp-up phase*** – which translates into a considerable amount of capital to work with from the

GDI project receives funding from the European Union's Digital Europe Programme under grant agreement number 101081813.

start. If the EDIC does not have an operational User Portal, data resources and compute capacity available there will be no users. Functionalities need to be in place and staff to handle this. The critical mass needed can be expected to be fairly high, before a steady state operational level can be achieved. This means that significant developmental investments as well as access to compute infrastructure (see section 5.2 below) must be secured.

In a complex public funding environment one key problem is that being successful might present scaling problems within fixed staff budgets. There is a risk that the number of requests the EDIC receives require additional staff resources. If the number of interactions increase continuously, as it hopefully should, the EDIC must be in a position to increase staff level and/or develop systems to automate as much as possible those interactions. At the point where the EDIC bridges into clinical utilisation, it is likely that there will be a need for a significant upgrade in terms of both staffing and skills required.

The Central Hub does require some leverage in order to engage with the NCP's and SPE's and others to have uniform data, support, SPE's and compliance in place. Lessons could be learned from ELIXIR, which uses some of its core funding to support development work at the national nodes through commissioned services.

The User Portal will likely need a significant investment to become functional. At this point it is not decided if the HUB runs all its operations on its own or through a SLA with product owners. The EDIC will be a collaborative effort that requires the skills of the specialists who work very close to the users, as well as a high degree of coordination and oversight from a central level in order for the infrastructure to function.

The 1+MG data catalogue depends on rich aggregated metadata and the ability to create a clear data characterization. This requires deep integration and resources at node level as well as central level to function. The Metadata catalogue requires built-in safeguards and granularity limitations.

In accordance with FAIR principles there should be a system in place that creates a persistent identifier[3] (PID) for the dataset used by the user, this (sometimes called "virtual") dataset will be a specific combination of a subset of data subjects and data items, as created by a specific query of the infrastructure at a specific time[4]. Due to the nature of Genomic research, there are unique challenges in developing an effective system that will allow users to test findings years later based on time stamped data and code used without compromising data protection. This is not an easy task, since it requires stringent documentation of data and code. On the other hand, if the EDIC is to bridge into health care, it is vital to establish a system that allows researchers, industry and clinicians to test the consistency of findings.

[3] Compatible with the EOSC, see http://dx.doi.org/10.2777/926037 - This may use a subscription rather than our own development.
[4] RDA has created recommendations for citing dynamic data; see http://dx.doi.org/10.15497/RDA00016

GDI project receives funding from the European Union's Digital Europe Programme under grant agreement number 101081813.

5.2 Considerations on hardware investments and cost recovery models

The core model for 1+MG is that the data is computed upon at national level, not leaving the 1+ MG infrastructure. . That will require a certain level of hardware investment and capacity, so that the nodes can deliver data analysis. In some countries there are already computer facilities available to provide this integral part of the platform. In other countries there will need to be a decision on how to ensure allocation of suitable compute resources towards the purpose of the 1+MG infrastructure. In either case, eventually upgrade of the required infrastructure and new investments have to be made, as the needs change and a new generation of hardware becomes available. It is important to have a very transparent business model and a clear picture of the revenue streams, for the SPEs to make investments that will scale towards European usage. It is also important to acknowledge that it may be very difficult to predict the amount of hardware needed to match usage. A key element will be to discuss how to alleviate some of the operational risks (downtime, idling) and risks that goes with planning investments and/or contracts towards an assumed future usage of capacity.

Whether or not 1+MG compute resources are directly allocated to that mission, or the resources are allocated as part of a larger national/local HPC system or through third party contracts, depends a lot on the actual framework in the individual countries offering SPE services. However, the infrastructure needs to be adapted to life science high throughput computing, while also maintaining an extremely high level of data protection and digital security, building on a general framework set by the EDIC, including security regime and change management. This level of integration *may* be difficult to achieve if the SPE depends on compute resources which are technically and/or organizationally part of a larger all-purpose compute infrastructure or commercially available resources.

5.2.1 Calculating compute costs

It must be recognized that what the 1+MG EDIC seeks to achieve, it does so in an area that has significant computational scaling challenges

At its core a federated infrastructure means users need to access the data where it happens to be, even if that means multiple places each to access part of the data, rather than collecting it in one place for analysis. There must be an effective analytic platform in place that can connect to the data resources. The platform must have a capacity to do large scale high throughput computing, meaning it needs capability to run an effective workload management system. The system must be capable of storing large amounts of data, not only the genomic data itself, but also the data and software the user wishes to bring to the analysis. This in terms requires a data-in and data-out infrastructure that is compliant towards general risk management paradigmes and systems. Obviously the computing resources must be sufficient to do the analysis on, and there needs to be a very tight compliance system in place.

Scientific computing infrastructure is investment heavy, requiring a significant one-off investment to purchase hardware, with traditionally only limited ability to recover those costs during the life cycle

GDI project receives funding from the European Union's Digital Europe Programme under grant agreement number 101081813.

of the hardware. Most scientific HPC infrastructure are installed based on an investment and operating grant – from research funders/ministries and/or EU, with a mix of in-kind contribution, co-funding and user fee from institutional sources and only a fraction of the costs are recovered through research grants awarded to the users. In itself users and funders are accustomed to only see the partial picture, not realising the full cost of offering compute services, since a lot of costs may be provided as in-kind: back-office services, interface services, physical space and power consumption may not be clearly visible.

Since the EDIC is likely to rely on resources acquired by someone else, the EDIC relies on someone else's funding decision, which may present a challenge in terms of developing long-term service consistent towards 1+MG users. This challenge feeds into the long term sustainability of the infrastructure. There is an inherent risk that the EDIC becomes "too successful" meaning that at some point the number of user requests cannot be accommodated on the infrastructure available; this risk is stronger if the use of the infrastructure is subsidized and not synchronized with upgrade cycle investments.

In order to address these scaling challenges and calculate the costs of actually computing on data, there needs to be a baseline cost item definition for the system and at some point, a carefully budgeted assumption about the number of users and jobs that system needs to accommodate.

Baseline cost assumption:

The cost calculation is based on the DNGC and CSC costs and price models, which are based on recovering real costs of services of a public entity: How many core hours are offered, and what does it cost to run that system, giving its' capacity, and depreciated value over the life time of the system.

Node days estimated cost 40-50 EUR[5]**:** Computer cores are the minimum cost items that a computer system can be divided into from a user perspective. How many cores you need, then depends on the scale and complexity of the job you wish to compute, the time allocated for the job – and the computational speed and storage capacity of the system. Cores are typically bundled into server nodes, linking computer cores to a processing memory and storage system. Typically an HPC Node consists of between 16 to 96 cores depending on the specific architecture needed.[6]

Active hot storage 25 EUR/ Month / 1 TB: Data that forms part of the actual data processing needs to be moved into hot storage, that allows the data to be fed to the computers per the requirements of the analysis - both 1+MG data and any data the user wishes to bring into the SPE.

[5] Based on DNGC and CSC costs and price models, both of which are based on recovering real costs of Service of a public entity: https://research.csc.fi/billing-units#buc

[6] For reference DNGC operates most of its 400 nodes/16.000 core system as thin nodes with 40 cores operating at 2.1 GHz with 196 GiB/1.9 TB SSD. In addition, DNGC offers Fat servers of 40 cores operating at 2.1 GHz with 1.536 TB RAM, 3.8 TB SSD. Some thin nodes also offer GPU accelerators (NVIDIA Tesla v100 16GB).

GDI project receives funding from the European Union's Digital Europe Programme under grant agreement number 101081813.

Examples of jobs and costs on DNGC's 400 node/16.000 core system (Not including overhead for, technical support):

Job size	Nodes	Days	Storage TB	Price EUR	#jobs in a year
Small job	1	½	0	20	292.000
Minor job	1	5	0	200	29.200
Medium job	4	20	20	5.300	1.825
Large Jobs	16	60	100	43.400	152
Disclaimer: The examples above are highly theoretical. No system can effectively launch jobs at 100% utilization. Traditionally effective compute utilization will be around 80% allowing for downtime and lag time between jobs on the system. The DNGC system used as baseline is also currently being updated, since the system from 2019 is at its end of lifecycle. Every update usually means a significant increase in performance and cost effectiveness.					
Running job	2	365	20	35.200	200

What the examples above show is how much the size of the job affects the costs and the capacity of the system – and very importantly - the potential distribution of jobs possible on a single system. This is important to understand in order, to have further discussions on the resource availability and the potential use cases for those resources. It will be vital for the development of the EDIC business model and the financial model to have a solid understanding of the number of users expected to use the infrastructure over time and for what types of jobs.

From a business perspective, one might consider offering free services or a flat-rate access fee for smaller jobs – since a high number of small jobs could add significantly to the cost recovery of the infrastructure without being prohibitable expensive for the users – charging 50-100 EUR for 6-24 hours of access to a 40-core computational node, could for some applications be very attractive and flexible. For other applications that put a significant drain on the available resources, there might be a need to cover those costs on a case-by-case basis.

5.2.2 Free-at-the-point of use vs. user fee

Many HPC systems rely on a mix of funding including user fees. User fees are challenging for two main reasons: 1) it can act as a blocker, disallowing scientists and clinicians access to compute and data resources and 2) it can potentially add a time delay from the moment an application has been sent, to the data becomes available, since paying for services adds some form of contractual element to the engagement. On the other hand, having resources free-at-the-point-of-use presents its own sets of challenges. CSC Finland operates a model in which scientists get a free resource entitlement based on their scientific need. That requires a very extensive and scalable system with strong central funding to work. And while that model might work on a national level, it is difficult to see how this could work across Europe. The "success trap" mentioned earlier often happens where resources are free, but not unlimited, which are then resolved by adding a form of excellence review allowing to prioritise one project over another, which is both an administrative burden, and difficult to

GDI project receives funding from the European Union's Digital Europe Programme under grant agreement number 101081813.

do fairly across different scientific or clinical domains. It is important to note, that past experience in operating public scientific compute infrastructure at national or EU level have not been able to produce a business model that relies on user fees to cover costs of operating scientific compute infrastructure - at best only the marginal costs.

It must also be recognized that national funders are generally not inclined to accept a funding commitment without a funding cap - which is a significant challenge when dealing with marginal cost sensitive areas of an infrastructure. An ability to contain costs and to link those costs to national priorities will be vital for a positive sign-on from countries. How cost-recovery for users' resource burn is structured across countries will be essential in that, particularly since there will be huge disproportionalities in the attractiveness of data sources across the different SPE's.

5.4 Data Inclusion

Data curation is the process of capturing and cleaning data and integrating data into a database in a standardised format, providing metadata for users. Data curation is extremely important since the 1+MG dataset is collected from many different research and clinical sources in different countries. The 1+MG dataset must ideally be as uniform as possible, including having minimal metadata standards attached to them, in order to be scientifically valuable. The most cost-effective way to ensure this, is if data providers employ the same set of centrally decided ontologies and standards across Europe. However, since data is captured specifically fit for purpose there may not be a clear focus on data quality and/or documentation needed for reuse. It will be particularly hard to implement pan-European/global standards into uniform clinical practice, since the EDIC will have no formal authority to do so, and only limited ability to engage directly with data providers.

The actual costs associated with Data inclusion is no doubt substantial. On one hand the EDIC intends to improve on current practise, but cannot take full responsibility nor fund basic data curation activities at data provider level – this must be fed into a bigger discussion on FAIR and Clinical Genomic at both EU and national level. Data curation costs are mainly staff costs and will mainly rest with the data providers – and outside the scope of the EDIC funding model.

Considerable development costs must be expected in order to design and develop tools to support data curation tasks with the data providers. The 1+MG / GA4GH framework analytical work on best practice on sequencing methodology and health economics will inform decisions on data curation and basic data quality. In that respect, coordination and support action costs should be expected and distributed across the Central Hub and the NCP's in order to engage with the data providers.

For now, it will suffice to say that the EDIC will need a strong staffing on data management, in order to drive discussions, lead quality controls and coordinate efforts with the national level. At NCP/national level there must be competent staff dedicated fully to the EDIC 1+MG mission that can liaise with the data providers, in order to support the implementation of 1+MG standards and create feedback loops to the EDIC. In addition, it is likely that there initially also needs to be staff dedicated to curating data to 1+MG standards in the NCP – however, this is a short-term, costly and non-scalable solution – the

GDI project receives funding from the European Union's Digital Europe Programme under grant agreement number 101081813.

fundamental aim is to support putting standards into clinical practise – not taking over the clinical data curation tasks.

Assumption

- Data curation costs are extremely difficult to calculate, as they are directly related to the quality standards that are employed and the amount of data generated.

- Since there will be organisational and cost issues with post-curating existing data to EDIC standards, it is probably not feasible to curate large amounts of existing data to EDIC standards. It is probably a more viable long-term solution to build the next generation 1+MG dataset from scratch using new data, and accept a larger degree of heterogeneity in existing data sources.

- Generally speaking, responsibility for the basic data curation should be outside the EDIC costing scope – however mechanisms to coordinate and support are most likely needed.
- The long-term ability of the EDIC to impose standards on clinical labs rests on the ability of the EDIC to become a trusted and integral part of clinical practice.

Discussion

- It needs to be clear WHO is the 1+MG data curator. Is it the individual PI or clinical lab or is it the NCP/SPE, who receives the data from the data providers that curates the data further? Having someone else doing data curation (to 1+MG and/or GA4GH standards) other than the actual data provider is not optimal in terms of quality and it will be time-consuming and costly. However, the EDIC will likely not have a working dataset anytime soon without data curation funding and doing this inside the infrastructure. But this will not be a long term scalable solution, nor something that should be encouraged through the funding model. This is a discussion that needs to happen at national level - What mechanisms could motivate data providers / national stakeholders to increase basic data quality and provide solid metadata for European integration?

- There needs to be a discussion on "where is the sweet spot?" between reusability/ data quality and cost/realism. One dilemma is that standards are probably easier to put into research practice, than clinical practice – on the other hand clinical production data are likely more uniformly captured and quality controlled than many research datasets. There needs to be a careful analysis on the cost/benefit of where the EDIC puts its energy.

- This also relates to the currently undecided discussion whether the infrastructure will place demands on the data (quality criteria) or whether it sets targets, and demands proper documentation of the targets that are reached and missed by a certain dataset. If we end up with hard inclusion criteria, these can probably be checked technically, such as by inclusion

GDI project receives funding from the European Union's Digital Europe Programme under grant agreement number 101081813.

scripts (e.g. using something like SHACL for validating whether the metadata satisfies constraints). That puts the burden that scales with the data volume outside of the EDIC; it leaves the burden to facilitate the process technically (lowering the costs per dataset).

5.5 Data Storage

In this section we separate the data storage from data curation, compute and general data management. Data storage refers specifically to costs relating to storing data on a medium. It is an important cost element of building up the EDIC infrastructure. Data storage must be divided into hot storage (able to feed data directly into the processing environment) and storage for data at rest - including off site backup. Storage must be scalable to accommodate the influx of data over a long time period and requires a long-term commitment.

Since storage costs are very linear according to data size, it is possible to approximate the size of the data set and thus the cost of data storage fairly precisely, given today's hardware costs. Costs will vary to some degree across countries according to salary levels, electricity costs and the ability to create economies of scale – for instance tapping into existing clinical or research storage facilities as well as mechanisms to recycle heat generation. Given those variables and dependencies it is estimated that ***Storing one Million whole human genome sequences constitutes a yearly cost of 35 MEUR.***

This estimation is based on the following:

- 1 WGS (30X coverage with current sequencing techniques) as CRAM and vcf files constitutes 80 GB of data. In practical terms this translates into approximately 250 GB of data, since some doubling of data is necessary as well as back-up.
- 1 Petabyte active storage and off-site backup requires a hardware investment of 400.000 EUR, Including overhead over three years.
- 1 Petabyte of storage will be able to accommodate 4000 WGS's . 1 Million WGS's will require 250 PB data storage, at a cost over 3 years at 100 MEUR or roughly 35 MEUR a year.
- ***Unit cost for storing one WGS for a year is 34 EUR***

Note: It is likely that data storage costs will decrease over time, as it has done historically. 250 GB per WGS depends on the requirements specified, particularly accessibility on demand and does not consider new CRAM formats potentially lowering storage per WGS. Any other formats (FastQ, Spring comprised etc.), will drastically increase storage requirements. Further guidelines on how to design cost effective storage needs to be discussed.

Countries need to decide how to organise long term storage and how to link/integrate that with the SPE locally/nationally? How this is done will decide organisational complexity and thus costs, and the ability to honour service level requirements for the EDIC. This depends a lot on the existing

GDI project receives funding from the European Union's Digital Europe Programme under grant agreement number 101081813.

funding options and/or existing research data storage facilities but eventually also the development of storage solutions in relation to national Genomic Medicines initiatives in health care. Some countries already use commercial cloud services to store and process genomic data. Whereas commercial cloud provides on-demand scalable resources, it is unclear if depending on cloud services will function within the EDIC – both in terms of technical integration, data security and cost recovery models. The implication of integrating commercial cloud providers into the ecosystem needs to be analysed both technically, legally and funding-wise.

5.6 Information security management

The biggest liability of any organisation that deals with digital (personal) information is compromising data protection and Intellectual Property Rights (IPR). This is an inherent risk of handling digital information, and it is especially true for a complex organisation, using multiple access points, infrastructures and business processes and staff distributed across multiple sites in a digital environment. In order to deal with those risks, organisations employ an information security management system (ISMS) to mitigate those risks, the key components of which are a) the establishment of a information security management system, for documenting risks, mitigation activities, controls and incident reporting and b) deploy a Data Processing Impact Assessment (DPIA) process designed to identify risks arising from processing personal data and to minimise these risks as far and early as possible[7].

An important design parameter of the EDIC is moving away from a trust-based security system into a security-by-design, meaning that the data delivery system, soft- and hardware are as free of vulnerabilities as possible in design and with built in safe-guards and monitoring systems. This is the essence of providing data access in a secure processing environment, which requires the EDIC to develop and operate a holistic system for information management and data protection, given the purpose and context of the data processing on genomic information.

At central level there will need to be staff qualified to engage in Information security management work, depending on the distribution of roles and liabilities. A formal entity that is managing a high risk data processing enterprise, needs to have the capacity to identify and handle risks pertaining to data processing. In order to do this, the central level needs to employ at least a Data Protection Officer (DPO) and a Chief Information Security Officer (CISO) and potentially other competences that bridges information security, data protection and ELSI. Also, there will likely need to be a Change manager, in order to ensure that a high level of coordination and control on ISM tasks are uniformly implemented across the different SPE's – including hardening, software version control etc.

[7] Under the GDPR, a DPIA is mandatory where data processing "*is likely to result in a high risk to the rights and freedoms of natural persons*". This is particularly relevant when a new data processing technology is being introduced.This means that while there has been no risk assessment done for the actual usage, there is a duty to screen for risk factors that point to the potential for a widespread or serious impact on individuals. Considering the organisational and technological aspects of the 1+MG EDIC a DPIA is therefore considered mandatory.

GDI project receives funding from the European Union's Digital Europe Programme under grant agreement number 101081813.

At NCP/SPE level there must be at least an CISO and a DPO to handle data security monitoring, and to conduct DPIA's and to engage with the EDIC and users. To which extent these persons need to be fully dedicated to the EDIC or work within a broader institutional framework, will likely depend on national set-up and scale of operations.

For the EDIC to function within the highest possible standards, operations should ideally be ISO27001 and ISO27701 certified. It should be noted that this is a very big and complex task in terms of mapping and assessing risks, documentation and developing protocols to build a ISO 27+ certified Information Security Management System. For reference, DNGC who is ISO27+ certified has spent a***t least 1.5 MEUR in consultancy fees, and between 4-5 FTEs for a year*** solely dedicated to build the ISMS system and achieve certification (although for all operations within a highly integrated national clinical and research service system). Currently there are four FTEs dedicated directly to ISMS operational tasks within the DNGC, not including all the operational staff members involved in implementing the systems and controls.

Whereas ISO certification is an extremely exhaustive and costly endeavour, the EDIC should initially aim for the SPEs to be ISO27+ compliant (not certified), and having ISEA 3000 audits conducted annually. Whereas an ISO27+ certification will indicate that an entity has an effective ISM system, it does not evaluate the actual data protection. An ISEA 3000 Type 2 audit, will give a direct opinion on the state of the data protection measures put in place. **An ISEA 300 Type 2 audit will cost around 15-20.000 EUR for each SPE and an ISO27+ recertification around 20-25.000 EUR,** both subject to national differences (DNGC platform as reference).

GDI project receives funding from the European Union's Digital Europe Programme under grant agreement number 101081813.

References

- OECD Global Science Forum - BUSINESS MODELS FOR SUSTAINABLE RESEARCH DATA REPOSITORIES. Chapter 4. SUSTAINABLE BUSINESS MODELS
  2017
  https://one.oecd.org/document/DSTI/STP/GSF(2017)1/FINAL/En/pdf

- StR-ESFRI Study GUIDELINES ON COST ESTIMATION OF RESEARCH INFRASTRUCTURES
  This publication was developed for StR-ESFRI - Support to Reinforce the European Strategy Forum on Research Infrastructures – by CSIL – the Centre for Industrial Studies.
  2019
  https://www.esfri.eu/sites/default/files/StR-ESFRI2_STUDY_RIs_COST_ESTIMATION.pdf

- TEHDAS - Preliminary study on funding sources and costs of secondary use of health data in the EU
  https://tehdas.eu/results/tehdas-identifies-funding-options-for-secondary-use-of-health-data/

- 2022 ELIXIR Annual Report
  ELIXIR releases 2022 Annual Report | ELIXIR (elixir-europe.org)

- 2020 BBMRI Annual report
  Annual_Report_BBMRI_2020_PRINT.pdf (bbmri-eric.eu)

GDI project receives funding from the European Union's Digital Europe Programme under grant agreement number 101081813.