EJP RD
European Joint Programme on Rare Diseases

H2020-SC1-2018-Single-Stage-RTD
SC1-BHC-04-2018
Rare Disease European Joint Programme Cofund

Grant agreement number 825575

Del 10.4
Fourth Annual strategic report and Action plan for Pillar 2, including: Systematic surveys reports, QMS of Pillar 2 description, GDPR compliance report and sustainability planning reporting

Organisation name of lead beneficiary for this deliverable:
Partner 1b – INSERM (Orphanet)
contributors: (25) FTELE; (73) BBMRI; (76) ELIXIR/EMBL-EBI.

Due date of deliverable: month 33

Dissemination level:
Public

Table of Contents

1. Pillar 2 Third Annual Strategic Report

1.1. Reminder of Pillar 2 background and concept

Pillar 2 was designed to mainly contribute with one of the major objectives of the EJP RD: To improve the integration, the efficacy, the production, and the social impact of research on RD through the development, demonstration, and promotion of Europe-wide and even world-wide sharing of research and clinical data, materials, processes, knowledge, and know-how. More precisely, Pillar 2 has been set to create an innovative coordinated access to data and services for transformative RD research aiming at rationalizing, optimizing, and increasing potential of existing resources and services, and to address the gaps on data infrastructure that is essential to enable multidisciplinary, holistic approaches for rare disease diagnostics and therapeutics by fostering creation of complete disease pathways.

Pillar 2 is creating the infrastructure for a sustainable and interoperable ecosystem of FAIR resources (the EJP RD virtual platform, or VP in short). It couples robust standards, tools and procedures that infuse FAIR principles into advanced and secure forms of data discovery and analysis by virtual data linkage and sharing. It will allow flexible, real-time access to data under suitably controlled conditions, with supporting tools and services for multiple types of stakeholders (e.g., clinicians, patient representatives, data scientists) serving the ultimate goal of increasing the efficiency and efficacy of RD research. Driven by concrete use-cases and needs arising from the RD clinical, patient, and research community, not least European Reference Networks (ERNs), it will provide the means to harmonise and standardise the way RD relevant data, samples, tools, and other relevant resources are made findable, accessible, interoperable, and re-usable, and the means to query the progressively increasing number of resources and repositories connected to the EJP RD virtual platform. Pillar 2 furthermore provides a central by which end users can explore the VP, and examples of automated analysis workflows including multi-omics network analysis.

The Pillar 2 strategy is to establish a stronger and broader collaboration between the RD community and European Research Infrastructures and global consortia. This will have major mutual benefit and impact. On the one hand, RD research, supported by patient representatives and ERNs, presents an exemplar challenge and opportunity for research infrastructures to create common solutions and stimulate collaboration. On the other hand, progress in RD research depends on the strongest possible infrastructure to address its needs towards efficient information retrieval and analysis across its distributed data resources. The increased capacity of infrastructures and their seamless integration with the RD community will ultimately translate to higher innovation potential and benefit for patients.

The final product of Pillar 2, namely the Virtual Platform (VP), will therefore allow to computationally query and analyse a myriad of heterogeneous resources, as well as build a progressively powerful federated resources and data sources network enabled for fostering RD research by promoting FAIRification of data sources and analysis workflows.

1.2. Methods

1.2.1. Reminder of Pillar 2 thematic structure

Pillar 2 work was organized in 4 Work Packages (WPs) around 4 main themes:

1. *Overall coordination*

WP10: **User-driven strategic planning and transversal activities for Pillar 2 data ecosystem, which** provides the critical 'coordination and navigation' role for the Virtual Platform, where users (especially ERNs) will participate as key leaders and ensure that the work in Pillar 2 (WP11, WP12 and WP13) is synergised and optimised. WP10 further conducts overarching activities towards VP GDPR compliance, quality and sustainability.

2. *Making resources usable for RD research*

WP11: **Common virtual platform for discoverable data and resources for RD research,** of which the main aim is to tackle fragmentation of data repositories, catalogues, resources and tools, by: (i) building a comprehensive, FAIR-compliant virtual platform extensively describing resources with their metadata (including registries, biobanks, research infrastructures, genome-phenome repositories, methods, standards, etc.) allowing for these resources to be findable online via a central access point and (ii) providing researchers the means to deposit, share and analyse phenotypic, genomic and multi-omics data in a harmonised, standardised manner, building-on and scaling-up existing resources, which will be findable through the virtual platform (VP) as well.

3. *Making record-level data usable for RD research*

WP12: **Enabling sustainable FAIRness and federation at the record level for RD data, patients, and samples,** which develops and applies procedures, standards, and tools, with the RD community to achieve FAIRness at the record level. This will enable clinical and biological researchers to discover useful and usable data with high specificity across resources, assess access restrictions for specific data quickly (e.g., consent, data usage licenses), and develop powerful analysis across multiple resources without delay caused by data incompatibilities.

4. *Making system biology approach data usable for RD research*

WP13: **Enabling multidisciplinary, holistic approaches for rare disease diagnostics and therapeutics**, which objectives are directed at filling the gaps that currently make it hard to perform multi-omics analysis on rare diseases. The aim of multi-omics analysis is ultimately to find better diagnostics (for instance process biological based panels) and to develop better therapies.

1.2.2. Reminder of Pillar 2 operational organization

According to the First Pillar 2 strategic plan (Del. 10.1), tasks and subtasks have been aligned to cross-task teams called Work Foci Teams (WFT) in order to develop the different components of the VP consistently and in a more efficient and coherent way. Tasks and subtasks not fitting a specific WF are conducted as expected as per the Grant Agreement Description of Work. However, stronger synergies between GDPR compliance, quality assessment and sustainability of VP components are sought through a common organisation of tasks 10.3, 10.4 and 10.5.

Since year 3, and according to the strategic plan presented in Del. 10.3, cross-WF tasks are being conducted so as to bring together the first developed VP building blocks. This strategy manifests by common meetings of Metadata WF and Query builder WF,

Resources for sharing data and materials WF ad Metadata WF, and the 3 WF on pathways development and the FAIRification WF, as well as pathways WF and Resources for data analysis and interpretation WF. FAIRification WF and Metadata WF share common activities as well. Below the short description of WFs' objectives:

- **Use cases WF**: setting up research questions by committed stakeholders, including ERN partners that will drive the development of VP components based on real-world needs;
- **Overall architecture WF**: global overviewing the VP components and connections between them;
- **FAIRification WF**: allowing data sources to become progressively FAIR, pertaining to incorporating technical services from Pillar 2 and collaboration with local data stewards, focusing on ERN registries and selected OMICS data resources;
- **Distributed and federated consent control WF**: defining where and how consent control is done based on the state-of-the-art and fitting it into the overall architecture of the VP. Defining other legal bases and definitions of roles (controller vs. processor in GDPR) for entities contributing or interfacing to the VP;
- **Authentication Authorisation Infrastructure WF**: providing Authentication and Authorization Infrastructure (AAI) to be used by other components of the VP. Building on ELIXIR AAI, BBMRI-ERIC AAI and the upcoming LifeScience AAI;
- **Personal data linkage service WF**: identifying datasets which belong to the same person (Privacy-Preserving Record Linkage);
- **Query builder WF:** Developing a federated discoverability and query facility for the VP;
- **Metadata model and alignment service WF**: Developing a computable ontology-based model of interoperable data descriptors for resources and records using semantic standards;
- **Resources for sharing experimental data and materials WF:** Improving, adapting, scaling-up and documenting resources for data and material deposition, access and sharing;
- **Resources for experimental data and analysis interpretation WF:** Prioritising novel user-friendly and cloud analysis functionalities;
- **Pathway creation and curation WF:** Collecting and curating conceptual rare disease pathways, collecting, and analysing rare disease multi-omics data
- **Biological networks analysis methods WF:** Creating networks using prior knowledge in the form of pathways and other database information and experimental data
- **Genetic variants pathways WF:** Collecting and curating conceptual rare disease pathways, collecting, and analysing rare disease genetic variants data
- **Environment/Adverse outcome pathways WF:** Extending rare disease molecular networks with drug targeting and toxicants initiating adverse effects
- **Federated analysis WF:** This is a newly created WF intended to develop a proof-of-consent of federated analysis of FAIR data based on the common work of WP11, WP12 and WP13.

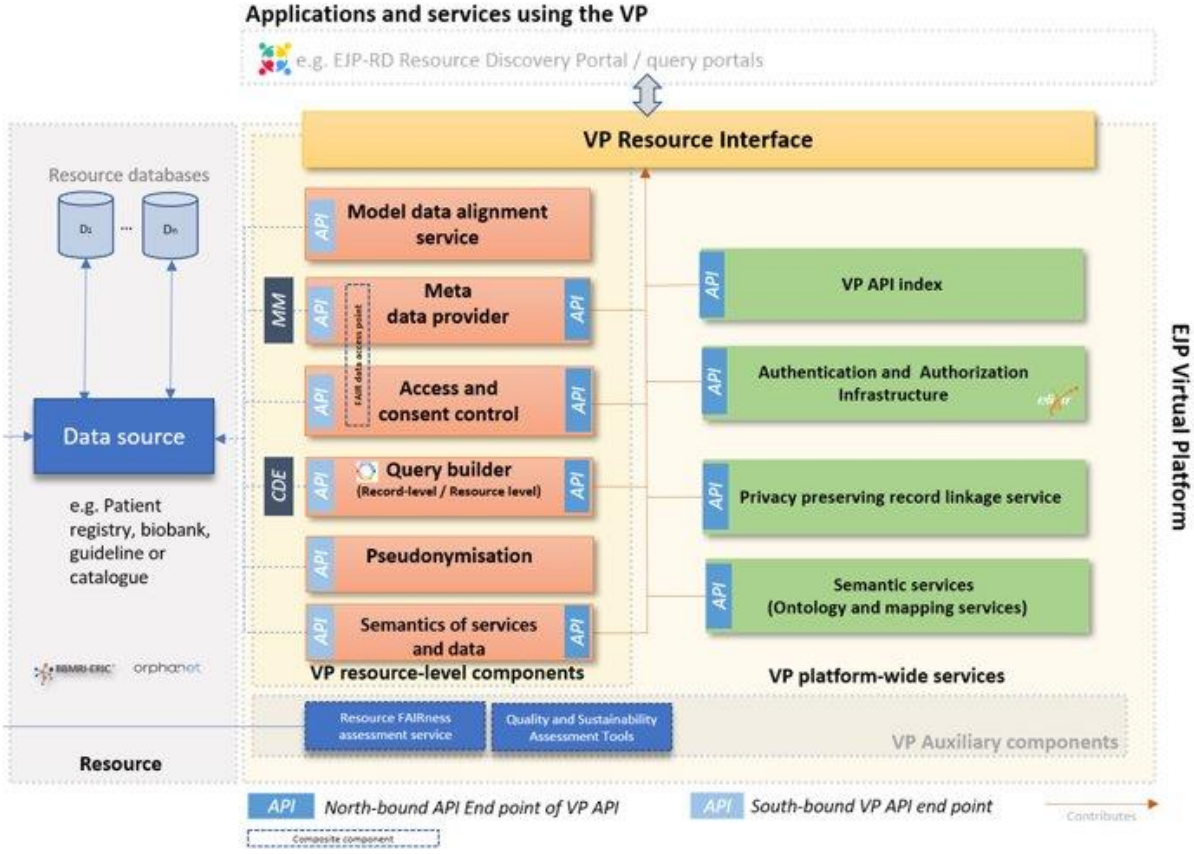
1.3. Strategic plan

1.3.1. Continue increasing visibility of EJP RD available resources through the development of the VP discoverability portal

The Virtual Platform (VP) is a network of resources and data sources technically empowered to interoperate with each other. One of the nodes of this virtual network is a discoverability portal on the EJP RD website, allowing to discover and query all the resources connected as a VP, and that are partners in the EJP RD. Based on the initial beta version of a portal allowing for federated queries on onboarded resources in the VP, which allows for querying a first set of resources by rare diseases names and ORPHAcodes, new query possibilities will be added on a growing number of onboarded resources. These will include, non-exhaustively, the use of semantic mapping services, the use of disease hierarchies based on the Orphanet classification of RD, research by genes, research by other metadata describing non-disease specific resources. A dedicated Task Force will develop the user interface (UI) in order to improve users' experience. Developments are done according to Agile methodologies, involving back-and-forth interactions with end users with the help of the Use Cases WF. Subsequent versions of the portal will be delivered every 6 months starting in December 2022.

1.3.2. Consolidate the technical components of the VP as an integrated architecture, and update the Virtual Platform Specifications (VIPS) accordingly

As the development of the VP components evolves, it is crucial to ensure internal consistency of the VP architecture in order to achieve a well-urbanised, integrated system that complies to quality standards defined during the project, that is sustainable, and that allows for further evolutions following the evolution of needs. As the VP matures, a set of standards, technologies and methods is defined and consolidated in the VP architecture, schematized in the figure below. A series of workshops will be conducted in order to refine the choices of the final elements in the VP architecture, that will be reflected in an updated version of the VP specifications document (VIPS).



1.3.3. Documenting GDPR compliance, quality assessment and technical sustainability of VP components, and set up a governance for the VP

In order to contribute to build a consolidated, sustainable VP architecture, it is crucial to make sure that its components are consistently assessed for their quality and technical sustainability, as well as for their GDPR compliance when relevant. After having set up the assessment criteria and provided the assessment tools, a common work will be conducted in order to optimize VP components assessment though a series of dedicated interviews with developers and owners. This work will be further extended to the onboarded resources in order to describe their GDPR compliance, quality and technical sustainability characteristics as metadata, expanding if necessary, the VP metadata model.

Based on the sustainability and overall architecture discussions, and after setting up the methodologies for onboarding new resources in the future, a discussion will be engaged within Pillar 2 in order to establish the VP governance after the EJP RD project ends. Discussions will encompass maintenance of the core elements of the VP, as well as the criteria governing the inclusion of new resources and the interactions with them.

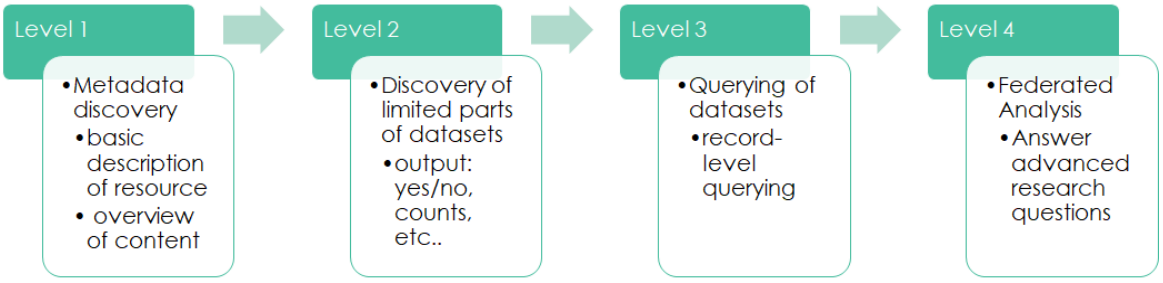
1.3.4. Facilitate onboarding of resources in the VP as a network for an improved discoverability and queryability

In order to be part of the VP as a network of resources and data sources for RD research, there is a number of requirements for a resource to comply with that increase the findability, accessibility, interoperability, and reusability of the resource with the rest of the ecosystem. These requirements are related with the metadata and data

standardisation and harmonisation, and with the technical components that will allow the resource to be discoverable and queyrable with other resources in the network by external end-users through the VP discoverability portal and computationally on behalf of multiple types of stakeholders. To this end, a portfolio of possibilities will be offered to the resources in order to ease their integration to the VP according to their current status, so as to provide enough flexibility to accommodate different situations, and to allow resources' holders to make the right choices. Several actions will be therefore engaged to this end, including the publication of an onboarding methodology, and the provision of trainings (see below). A common effort will be engaged across the Metadata WF, FAIRification WF, Query builder WF, Overall architecture WF, Use cases WF and Resources for sharing experimental data and materials WF, by organising online meetings and Face-to-face workshops.

1.3.5. Enlarging the VP to other prioritised resources in a federated manner

In order to get EJP RD partner resources represented in the VP metadata model, and to allow these resources to be discoverable and queryable through the VP, a new set of resources are prioritized each year. As the onboarding work has been either initiated or completed with the majority of EJP RD partner resources, the strategy for the 4th and 5th years of the project will be directed to finalise the onboarding process by defining the level of queryability each resource will achieve within the VP, represented in the figure below. A common work of Metadata WF, Query builder WF and FAIRification WF, with ad-hoc involvement of other WFs, will be engaged to this end.



1.3.6. Continue expanding the EJP RD FAIRification stewardship programme

The FAIRification stewardship program initiated to accompany ERNs' registries development in a standardized, VP-compliant way, will be extended to other resources such as genomes (e.g., for the Beyond one-million genome initiative – B1MG), cell lines, stem cells, and analysis (workflows, results, provenance). In particular, the data stewards will collaborate further with the Metadata WF team in developing the methodology for onboarding new resources in the VP based on the methodologies and experience already developed to accompany data sources.
This program is operating within the FAIRification WF and includes building a FAIR Stewards Team, developing a FAIRification roadmap, and a set of supporting documentation, an inventory of standards and tools and FAIRification procedures, including a Smart guidance for FAIR data stewards, as well as learning modules and materials.

1.3.7. Tackling the secure accessibility and reusability challenge

To allow querying data elements in data sources and to provide detailed results (i.e., number of patients with a combination of data elements for charting purposes) it is necessary to put in place authentication and/or authorisation technologies allowing for enabling or rejecting these types of deep-querying. LifeScience AAI will be implemented in the VP and proposed to connected data sources to interact with it. In addition, data access and use conditions of resources will be described in the metadata model for machine readability based on Common Conditions of use Elements / Data Use Conditions (CCE/DUC) profiles or in an ontological model using the Open Digital Rights Language (ODRL).

The Common Informed Consent Form (ICF) should also be made machine-readable and guidelines for the implementation of Privacy-preserving record-linkage technologies should be adopted. Their implementation allows for advanced and accurate federated analysis.

Tackling the secure accessibility and reusability challenge is key for enabling resources and data sources for queryability levels 3 and 4.

1.3.8. Improving and developing further data deposition and analysis facilities

Improvements in the adaptation of non-RD specific EJP RD partnering resources to RD have been made and will continue to allow interconnection between complementary resources (i.e., allowing to query a resource from another one) with the goal to streamline researcher's search experience. Data deposition and analysis resources will be made also queryable from the VP, in the frame of the integration of WFs developments. In particular, resources will act as reviewers of the methodological documents produced within the VP, and participate to the VP development workshops and trainings.
The development of a cloud-based solution for custom analysis will continue by expanding the EJP RD Virtual Cluster, in close collaboration between WP11 and WP13. In particular, Variant interpretation, whole exome, and multi-omics analysis methods will be established as separate pipelines (using a standard pipeline language and containerization) to automate future analyses. These developments will be made publicly available and registered to be findable in the UMCG Virtual Cluster Environment (VRE). New analysis capabilities and tools will be developed in EJP RD analysis platforms in response to users' needs collected during webinars. Furthermore, WF related to resources for data deposition and sharing, and to Resources for experimental data and analysis interpretation will engage in the new federated analysis WF activities.

1.3.9. Enhancing and expanding RD pathways creation and analysis based on case studies, and making them findable through the Virtual Platform

The strategy for the results of the work done over the first years creating and curating of RD pathways and networks is directed towards increasing their usability by containerizing them, further FAIRifying them and exposing them through FAIRDataPoint, and contributing to the Federated analysis WF. In addition, a series of hackathons and workshops will contribute to this strategy by modelling toxicology for RD and working with researchers in "bring your own data" workshops. External outreach and connexions will be ensured by aligning with ELIXIR tools, and by

collaborating with Solve-RD project to improve genetic diagnosis. Finally, a general "How to approach a new(rare) disease in systems biology" workflow will be published.

1.3.10. Preparing for federated analysis on FAIR data

As more data sources in the VP become FAIRer and VP compliant *at source*, and thus describe themselves and their content using the machine-interpretable metadata and data models adopted in the VP, the VP itself becomes progressively a FAIR virtual store of ontologically qualified data. Based on the low-hanging fruits built up during the previous years in the EJP RD, a new work focus team will be launched to design a proof-of-concept of federated analysis on FAIR data, paving the way towards AI-readiness within the VP. This WF will encompass two teams, one on the architecture of federated analysis and another on exploitation of FAIR data for analytics. This WF will deliver the EJP RD methodology for FAIR-based federated data analysis and demonstrate the concept of using machine actionable, interoperable data. In addition, it could demonstrate possibilities of current VP specifications/FAIR artefacts with early adopters and example data based on real resources in the VP.

1.3.11. Putting a dissemination/communication plan into actions, including the provision of training for VP contributors and end-users

As the VP becomes operational, a dissemination strategy will be designed and implemented based on engagement of contributors (resources and data sources) and end-users (researchers, clinicians, patients, and all interested stakeholders). This plan will include, non-exhaustively, the publication in the EJP RD websites of use cases' "Flash cards" to exemplify usability of the VP, seeking at interactivity with end-users in order to capture their feedback; a label "VP connected resource" to be displayed in contributing resources websites and dissemination channels including social media; a series of trainings, in collaboration with Pillar 3, intended to VP contributors to helping onboarding, and to end-users, to foster the utilisation of VP facilities.

1.3.12. Expanding Pillar 2 external collaborations to promote an interoperable RD research network of resources in Europe and beyond

Partnerships with ongoing initiatives that are complementary to Pillar 2 developments, in order to increase VP interoperability capabilities and ultimately contribute at creating the conditions for RD research both at the European level and internationally. As in the previous years, partnerships will intend to develop common use cases. In particular, collaborations with the Critical Path Institute (C-Path), Beyond 1 Million Genomes project (B1MG) and the Genome Data Infrastructure project (GDI), X-eHealth, ERICA, EOSC-Life, and the European Health Data Space will be initiated or reinforced. They will aim at harmonising the use of semantic and technical standards, at developing data model transformation methods, and/or at developing further a federated framework across projects.

2. Users survey report

In order to feed the discussions preparing the future governance of the VP, an internal survey will be launched in year 4 in order to identify the gaps in terms of resources connected to the VP to respond to end-users' needs; to establish the criteria for onboarding new resources in the future; as well as to obtain feedback on the long-term sustainability aspects to be considered after the project ends.
Results of the survey will be presented and discussed at the VP governance workshop that will initiate drawing the plans for the future.

A second survey will be organised during the last year of the project, reproducing the EJP RD initial survey towards VP end-users. We want to understand and document the changes in the awareness and knowledge that researchers have about resources available for them at the end of the project compared with the situation before EJP RD starts.

3. Oversight and monitoring of non-functional elements of the Virtual Platform

Strategic actions on GDPR compliance, quality assessment, and sustainability planning which were listed in section 1.3.3 were revisited in a series of WP10 meetings between May and September 2022. The aim was to ensure holistic long-term planning that aligns with the technical developments toward the release of the upcoming versions the Virtual Platform. This collaborative effort between tasks 10.3, 10.4, 10.5 and WP3 resulted in an oversight plan that will cover the following:

3.1. Identification and prioritisation of Virtual Platform technical components for assessment

Following the updates in technical development and the alignment of technical approaches used for connecting to the VP, the initial list of technical components was reviewed and categorised by function. The list included technologies and standards (building blocks of the VP) developed and deployed in EJP RD.

The list was prioritised by identifying core components that enable the VP ecosystem to function as a network in accordance with the updated Virtual Platform Specifications document and the overall architectural design of the network of resources. Stakeholders were identified for each of the prioritised components.

3.2. Grouping tools used for GDPR compliance, quality assurance, and sustainability planning into a kit ready to be used to assess the prioritised components

The tools developed in Y3 are harmonised and grouped into an online kit ready to be deployed by the stakeholders of the technical components identified in section 3.1. The kit comprises the following:

3.2.1. GDPR Compliance Monitoring Questionnaire

The questionnaire, offered by Molgenis, is based on a data model that was implemented in Y3. It covers aspects related to data protection, such as information on legal persons operating services, roles of the legal persons with respect to the data (controller/processor) types of data being processed, legal bases used for processing for data controllers, and information on application of privacy enhancing technologies.

3.2.2. Quality Assessment Tool

The tool was developed to assess the quality and fitness for purpose of the different technical components of the VP. In Y4, the quality criteria were refined following an initial trial and was incorporated into a draft Quality Policy. Following approval this will be added to the VIPS updated document as a next step.

Multiple rounds of data gathering started using the tool; which was expanded to include licensing data. The collected data supplements information from the regular presentations by different development teams and resources as part of the regular Non-Functional calls. This has contributed to the Molgenis data model described in

section 3.2.1 above; expanding the pilot phase to create a more comprehensive model.

3.2.3. Sustainability Oversight Reporting and the Business Canvas Model

During the Annual Retreat of Pillar 2 in May 2022, the partners involved in the session on sustainability and long-term planning agreed on extending the work of technical and financial sustainability assessment to include all technical components of the Virtual Platform (refer to section 3.1). This work is being carried in parallel with the assessment of maturity and sustainability of resources connected to the Virtual Platform during the onboarding process. Through the two methods of assessing the technical components as well as the connected nodes within the network, we can ensure that the whole ecosystem is sustained beyond the timeline of the EJP RD. Deliverable 10.3 detailed the reasoning behind the expansion of the analysis and the motivation to thoroughly cover the following aspects of sustainability in close alignment with WP3:

- Internal dependencies
- External dependencies
- Licensing and ownership
- Contingency
- Credentials
- Technical approach and respective specifications
- Interoperability with other components and/or standards
- Other sustainability aspects: sustainability plan; advisory board; documentation & trainings

3.3. Conducting interviews and meetings with stakeholders of each component to collect the data in the kit

The series of extensive interviews and meetings are scheduled between October and December 2022 to review the progress in data collection using the kit that comprises the tools detailed in section 3.2. The interviews will be conducted with identified stakeholder of the VP technical components as per the prioritisation plan (section 3.1).

This step is adaptive and will start with the core components that are key to the Minimum Viable VP in its first version. With the technical developments, the characteristics of some components will become well-defined at the beginning of Y5; which will enable us to refine the information collected via the Business Canvas Model.

Further interviews will be planned in Y5 for elements that were identified as non-core and were not prioritised in the first step (for example, record-level access authorisation).

3.4. Analysing the results and liaising with WP3

The results of the questionnaires and the outcomes of the interviews will be curated and populated per component. The analysis will identify any gaps and enable the alignment of financial, legal, and technical long-term plans. The analysis will be

conducted by the Overall Architecture Work Focus and experts on sustainability to ensure the implications of architectural and technical decisions are considered in the review of sustainability, GDPR compliance, and quality monitoring.

The results will be discussed in the Y5 Annual Retreat scheduled in March 2023 and the outcomes will be formulated in the next action plan and disseminated to the resources, technical teams overseen by the Overall Architecture Work Focus, and the supporting teams that conduct documentation and training functions.