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 10.2
Second 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-ERIC; (76) ELIXIR/EMBL-EBI

Due date of deliverable: month 18

Dissemination level: Public

Table of Contents

1. Pillar 2 Second 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 essential to enable multidisciplinary, holistic approaches for rare disease diagnostics and therapeutics by fostering creation of complete disease pathways.

Pillar 2 will create a sustainable and interoperable ecosystem of resources (the EJP RD virtual platform, or VP in short), coupled to robust standards, tools and procedures that will infuse FAIR principles into advanced and secure forms of data discovery, linkage and sharing. It will allow flexible, real-time access to data (under suitably controlled conditions), with supporting tools and services that serve the ultimate goal of increasing the efficiency and efficacy of RD research. Driven by concrete use-cases and needs arising from the RD clinical and research community, not least 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 through a central facility.

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. 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.

To that point, Pillar 2 partners decided to develop three parallel and complementary axes of work, all three governed by consistent quality, regulations and compliance with agreed standards requirements, and driven by research and clinical user community, ensuring that the Virtual Platform is built for and adopted by this community. The axes are:

- **A common metadata** ontological, machine-readable **model** describing pre-existing **resources** (including catalogues, data repositories, tools and infrastructures) with RD-specific semantic standards and metadata that also encompasses consent and data use conditions. This metadata model will structure a temporary centralized repository (the Linked-Data Platform), as a proof-of-concept, to evolve

to a federated system once resources have exposed their metadata conforming to the metadata model. This is expected to make resources Findable, and their Accessibility conditions to be known. Furthermore, it will provide guidance and services to produce, store and share phenotype-genotype data for researchers, as well as standards for data processing, generation and deposition of multi-omics data, relevant for RD. This axis of work will have the double benefit to make all relevant resources (RD-specific or not) findable and queryable through a single entry point, so optimizing their use by researchers, and to increase normalisation and standardisation of resources' metadata as well as their progressive adaptation to the rare diseases' community needs;

- Federated ecosystem of FAIR-at-the-source resources, in order to enable data discovery, sharing and analysis down to the **record** level. This axis of work organizes partnering with target resources to make them FAIR compliant, as well as the required interdisciplinary collaboration between RD experts and data experts. Beyond achieving FAIRification of resources in the VP, Pillar 2 will offer to the RD community at large the means and support, including training, to enlarge the scope of the federated ecosystem, by building a sustainable FAIRification service.
- Extension of the virtual platform with workflows that allow holistic research based on rare disease data and biological knowledge, by **filling gaps** in data on disease modifiers such as nutrition and metabolism, aspects of lifestyle and exposure to toxicants with the final aim to improve the understanding of the aetiology and progression or rare diseases and to support the discovery of biomarkers and the identification of druggable pathways and targets. This axis of work requires exploitation of previously performed explorative studies and the conduction of proof of principle studies specific for RD. It will be based on important collaboration between European Research Networks and top-level research teams excelling in data mining and interpretation.

The final product of Pillar 2, namely the VP, will therefore allow to centrally query a myriad of heterogeneous resources, as well as build a federated discoverability, query and analysis facility by promoting the progressive FAIRification of data sources, including multi-omics rare disease pathways created by Pillar 2 itself.

1.2. Methods

1.2.1. Pillar 2 thematic structure

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

1.2.1.1. 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 decision makers, and ensures that the work in Pillar 2 (WP11, WP12 and WP13) is synergised and optimised.

1.2.1.2. 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.

1.2.1.3. 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 will develop and apply 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.

1.2.1.4. 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 support finding better diagnostics (for instance process biological based panels) and developing better therapies.

1.2.2. 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). Indeed, since the different components of the VP should be consistently developed towards making resources, records and data findable and queryable in a coordinated manner, organizing the work by Work Foci (WF) allows related tasks to be conducted together in a more efficient and coherent way. WFs do not replace tasks and subtasks as per the Description of Work (DoW) in the Grant Agreement (GA), but make them work together towards a common objective therefore optimizing the use of resources. Tasks and subtasks not fitting a specific WF are conducted as expected as per the GA's DoW.

WF teams (WFTs) work according to the Agile methodology involving both Pillar 2 partners and ERNs designated representatives for each of them. This will allow for cycles of development and testing. Further delineation of WFs will be done as the project evolves in a flexible way. Experiences from field-testing based on use cases ensure that Pillar 2 developments fit the needs of users and have the expected impact

on accelerating research. The Overall Architecture (OA) WF ensures the global consistency of the VP components according to the VP specifications that are thus elaborated collaboratively. 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;
- **Interoperability for registries WF**: mapping expertise and standards between GA4GH, ERNs, and EJP RD partners, with an initial focus on suggestions and guidelines for ERN registries;
- **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;
- **Networks WF:** Collecting and analysing rare disease multi-omics data, creation of networks using prior knowledge in the form of pathways and other database information and experimental data;
- **Pathway creation and curation WF:** Collecting and curating conceptual rare disease pathways for further use in network analysis;

- **Genetic variants WF:** Optimizing mapping of genetic variants to genes, diseases, and other information available on databases, integration of these efforts to network analysis to improve knowledge driven analysis of genetic variants and improve description of variants of unknown significance.
- **Environment/Adverse outcome pathways WF:** Extending rare disease molecular networks with drug targeting and toxicants initiating adverse effects;

Operations at the WP and task levels, including transversal issues such as quality (establishing principles/approaches), governance (focused on GDPR) and sustainability by design, will be monitored using project the EJP RD Microsoft Teams.

1.3. Strategic plan

Following the initial developments in each of the tasks and subtasks in Pillar 2 which resulted in a better understanding of the resources in place, in their improvement to serve specific rare disease research needs, and in the development of common data models, adoption of common standards and exploration of solutions for the Virtual Platform building blocks, next strategic steps are defined towards a consistent, interoperable ecosystem. This implies adopting common governance rules, starting to actually inter-connect the different pieces of the puzzle built in the first 2 years of the project and increasing users' awareness by a series of outreach measures.
The following strategic orientations are the result of discussions held at the First EJP RD General Assembly dedicated Pillar 2 session and at the Second Pillar 2 Annual Retreat, and the actions towards these strategic orientations are detailed in the 3rd EJP RD Annual Workplan.

1.3.1. Increasing visibility and awareness concerning EJP RD available resources to foster RD research

Both year 1 and year 2 showed that researchers have a little knowledge of the existence of research infrastructures and other resources that offer services, data and material for conducting research n RD. In order to increase awareness and to ease identification of appropriate resources that can help addressing a typology of researchers' needs, a webpage will be produced to provide an overview at a glance of EJP RD partner resources, tools and services. Starting with Pillar 2 partner resources, this page can be expanded to other Pillars in the future. This page will be complementary to the EJP RD Helpdesk already available.
Furthermore, a series of webinars and documentation will be produced, in collaboration with Pillar 3.

1.3.2. Conducting consistent development workflow for VP components: the development life cycle

In order to ensure that individual VP components are developed consistently, a streamlined workflow is in place. It starts with the identification and formulation of new use cases (by the Use case WF) alongside existing use cases already identified. Use

case WF discuss and specify use cases submitted by users or triggered by developers in order to stick to their development workplan. Specifications are structured according to a common agreed template.
Development plans are assessed for functional and non-functional requirements by the Overall Architecture WF team timely enough to ensure appropriate risk management. Assessment is intended at ensuring coherence within the VP architecture, compliance with quality and sustainability criteria, use of standards and compliance with GDPR when appropriate. Assessment steps occurs before going through agile development cycles by Work Foci Teams to enable progress in delivering Minimum Valuable Products (MVPs) that can be tested before going further with development to the final product.
In order to achieve smooth running this development life cycle, dedicated sessions between each individual WF and the Overall Architecture team are regularly scheduled within MS Teams.

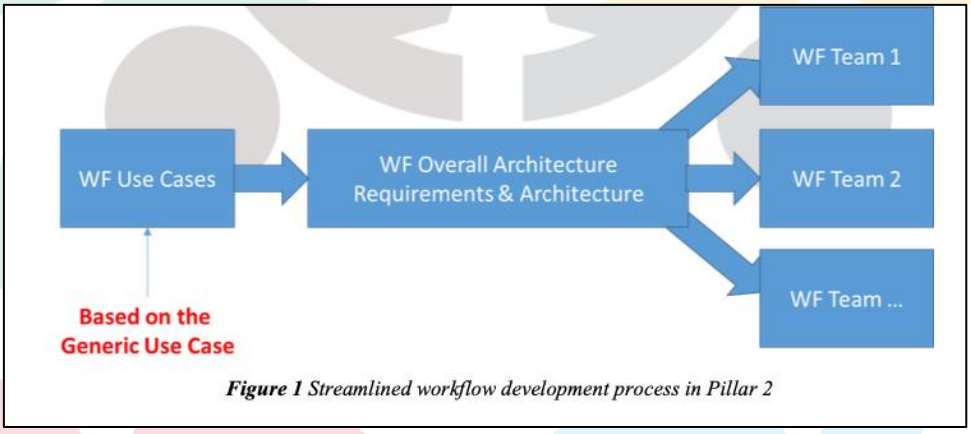
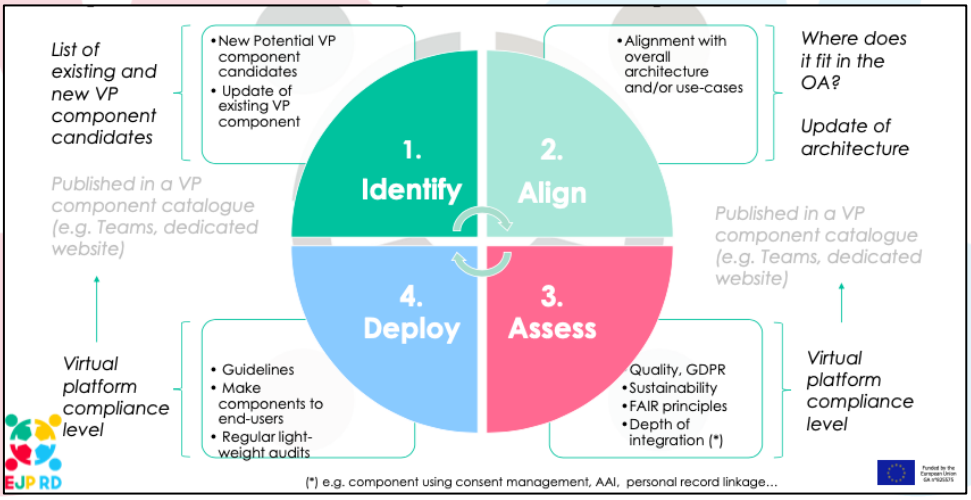


Figure 1 Streamlined workflow development process in Pillar 2



1.3.3. Establishing functional and non-functional requirements for VP components development: Virtual Platform Specifications (VIPS)

Based on the abovementioned workflow applied to a generic use case, design and implementation principles will be agreed upon. These principles will guide achieving formalized, but evolutive, functional and non-functional VP specifications (VIPS). The "counting cases use case" was selected as it is generic enough to mobilise the development of all the components of the VP. In order to obtain a functioning and robust EJP RD Virtual Platform, the VIPS will specify a framework, including implementation and interoperability conditions, to be considered by developers of software components that need to function with multiple resources in the VP infrastructure. Interoperability is here considered in its wider meaning, encompassing its technical, semantic and legal and organisational aspects, including identification, evaluation, selection and endorsement of a growing set of standards.

VIPS will include non-functional requirements. Non-functional requirements encompass criteria VP technical components should be compliant to in order to achieve an increasingly consistent, internally interoperable, Virtual Platform. These requirements include Quality, Sustainability, Scalability, Data Protection (GDPR), Consent Management, FAIRification. Dedicated WFs or working groups, some transversal with non-Pillar 2 working groups (Sustainability, Consent control, GDPR) work at releasing guidelines. Guidelines for non-functional requirements will therefore supplement the Virtual Platform Specifications (VIPS) that will be delivered in year 3.

1.3.4. Continue Pillar 2 efforts in understanding basic and clinical researchers' needs by engaging them in defining use cases

The year 1 survey was intended to capture the needs from ERNs as they are drivers for Pillar 2 developments, and the year 2 survey was conducted in order to capture needs from a broader panel of researchers. The second survey report is in the next section of this document. Consistently with the VP development workflow explained above, and because of the strong interest and engagement researchers within or outside ERNs demonstrate in both surveys, we decided to direct our efforts to identify and engage different types of stakeholders to participate in Pillar 2 agile development cycles. These stakeholders include ERN clinicians, researchers, biobank / registries owners, patient organisations / representatives as well as partners from other Pillars. This will be done through targeted, structured interviews, in order to bring use cases triggering the workflow, and to involve end-users as testers. Outcomes and use cases of the structured interviews will be reported as a part of the annual Strategic Plan.

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

During the first year of work, catalogues of registries, biobanks and tools were prioritized (Orphanet, BBMRI, RD-Connect Biobank and Registry Finder, JRC's ERDRI and ELIXIR bio.tools) to develop a comprehensive suite of catalogues using standardized fine-grained metadata and utilizing existing recommended ontologies and semantic standards (i.e. Orphanet rare disease ontology (ORDO), Human Phenotype Ontology

(HPO) and HOOM, the ontological module making ORDO and HPO be usable together). Following the first Pillar 2 Annual Retreat, resources addressing the ERN researchers' needs for support on translational research have been prioritized. A collaborative work has been therefore engaged with Pillar 4, in particular to extend the VP resource-level metadata model to include metadata descriptors for documents and services. By doing so, Pillar 2 partners contribute building the Innovation Management Toolbox developed within WP19. Based on that, EATRIS and ECRIN documents and services can be made discoverable and queryable trough the VP.

Following the 2nd Annual Retreat, and in order to address rare disease researchers' needs captured by the 2nd survey (see below),Pillar 2 partner resources allowing to find animal models (Infrafrontier), cell lines (Cellosaurus and hPSCreg) and resources for deposition and analysis of genomic and phenomic data (GPAP) are prioritized. The ontological metadata model and the resources-level query builder will be expanded in order to make them findable and queryable through the VP.

1.3.6. Developing the EJP RD FAIRification stewardship programme

According with the 1st Pillar 2 Strategic plan, and the publication of a guidance document on how EJP RD Pillar 2 partners can support increasing interoperability of ERN registries, FAIRification plans are prioritized towards making all ERN registries FAIR, by designing a three-parties process for collaboration, involving:

- An ERN registry steward
- An EJP RD interoperability steward
- A registry software provider

To this end, a FAIRification stewardship program has been decided. It is operating within the FAIRification WF and includes building a FAIR Stewards Team, developing a FAIRification roadmap, and a set of supporting documentation including a mind map summarizing all the FAIRification aspects to be taken into account and for which expertise is to be mobilized within Pillar 2, an inventory of standards and tools and FAIRification procedures. Counterparts at each ERN registry level have already been identified and the work has started.

1.3.7. Improving data deposition and analysis facilities

Improvements in the adaptation of non-RD specific resources (I.e. Infrafrontier, Cellosaurus, hPSCreg, and others) to RD are ongoing. These will go a step further to allow interconnexion between complementary resources and, 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. The facility to query and access resources through a single identification and authorisation service will be developed by implementing Life Science AAI.

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, single-cell RNA sequencing and long-read sequencing, are being assessed. Incorporation of extra layers allowing for multi-omics integration tools will be performed, to create, for instance, a modular pipeline for variant interpretation. This layer consists of a visual representation of omics data, to which modules can be added to filter or prioritize variants based on the (multi-)omics analysis input and identify the variant causal for the patients' phenotype.

1.3.8. Enhancing and expanding RD pathways creation and analysis, and making them findable through the Virtual Platform

X-omics workflows, tools and training materials that have been developed should be included in the Virtual Platform, and interconnected to data deposition and analysis resources (see section above), to progressively build a streamlined research pathways linking phenomics, genomics and X-omics data to enhance findability and usability of data resources in EJP RD. It will allow creating a catalogue of multi-omics integration tools that should help RD researchers to decide which tools to apply in which situation. This integrative work will be performed by leveraging ELIXIR Bundle-services and bio.tools capacities, and will be instrumental in making these different resources working together within the Virtual Platform.

Rare diseases pathway creation and curation according to use cases will continue. In particular, application of the protocols designed in WF AOP/Environment will be applied to a use case. Analysis on multi-omics datasets coming in from the ERNs and from the EJP RD funded researchers will continue. Particular attention will be put towards drug target identification based on integration of experimental datasets with established knowledge databases, including Wikipathways, ChEMBL, and AOPs.

2. Survey report

2.1. Problem/Background

The Community Survey performed in 2019 on European Reference Networks (ERNs) revealed striking information on the status of rare disease research data use and generation, level of awareness in Pillar 2 resources, concept and needs (Deliverable 10.1). A second survey was prepared in 2020 to survey the rare disease researchers in the community at large. Academic scientists and ERNs clinicians perform complementary research on rare diseases, and both are recognised as the major users of the Virtual Platform. Notably, research scientists are identified as those working in universities or research centres and may be funded through the Pillar 1 Joint Translational Calls. These are the two major stakeholder groups representing the RD research community and as such the goal of the year 2 Community survey was to capture the needs of research scientists in the community.

2.2. Methods

The questions from the Year 1 survey on the ERNs served as the backbone of the new survey on Researchers. The Year 1 survey was trimmed to focused on Pillar 2 activities, resulting a survey with 4 sections and 47 questions:

1. About
2. Data generation and storage
3. Data annotation and FAIRification
4. Use of existing resources and infrastructure for research data

The complete list of the questions is attached in Annex 1.
The new set of questions were open for review by Pillar 2 members and the EJP RD coordination office, and set up using the LimeSurvey. The survey was disseminated with an accompanying email through multiple channels including Pillar 1 JTC Call coordinator to JTC grantees, EJP RD Twitter and Newsletter, and EJP RD partners' own channels and networks (CIBER, FTELE, IMAGINE, ORPHANET). In addition, the survey was sent to a number of learned societies including European Society of Human Genetics (ESHG), Society for the Study of Inborn Errors of Metabolism (SSIEM), International League Against Epilepsy (ILAE).

The survey remained open for 25 days between March 27 and April 21, 2020. The whole process on the preparation of the survey is illustrated in Figure 1.

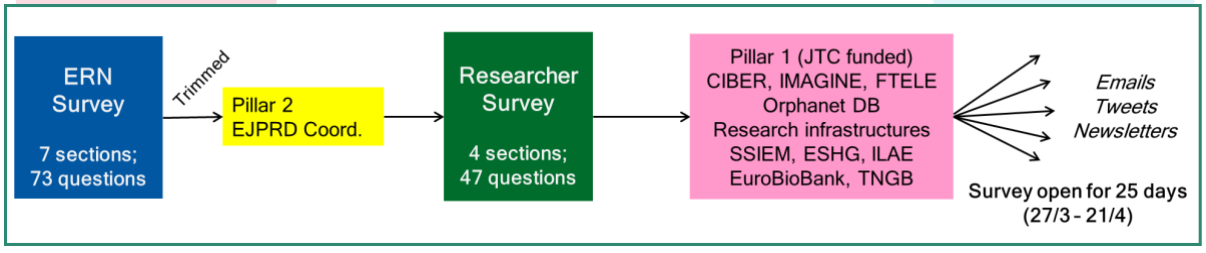


Figure 1. The creation of the 2020 Researcher Survey and its distribution

2.3. Results

2.3.1. Responses to the Survey and respondent research profiles

A total of 463 responses were received and a simple data clean step was applied to remove duplications and responses from persons with matching emails from 2019 ERN survey. Finally, 451 responses were processed for analysis.

Researchers from 31 different countries responded to the survey: Austria, Brazil, Belgium, Canada, Croatia, Cyprus, Czech Republic, Denmark, Finland, France, Germany, Greece, Hungary, Japan, Ireland, Israel, Italy, Latvia , Netherlands, Poland, Portugal, Romania, Slovakia, South Korea, Spain, Sweden, Switzerland, Turkey, Uganda, United Kingdom, USA. Interestingly, a number of responses came from outside of Europe. The countries with the highest number of responses were Spain, Italy, France, Germany and UK (Table 1). The levels of responses appear to be slightly different, but comparable, to the top five responding countries of the 2019 ERN Survey (Italy, Netherlands, France, Germany, Spain).

Table 1: The countries with top responses on the 2020 Researcher Survey.

Country	No. of responses
Spain	104
Italy	103
France	65
Germany	50
UK	18

The majority of the 2020 Researcher Survey identified themselves as Principle Investigators (83%). Many of the researchers agreed for recontact to further exploration of their research needs for Pillar 2 VP development (68.1%; 307 persons). The researchers work on all types of research, with basic research (70%) being the most common, followed by translational (65%) and clinical (53%) research. Most of the researchers own a digital identifier number such as ORCID (88%) and/or another researcher digital identifications, versus 9% who said they did not own any. This information may be relevant for the establishment of VP user authentication process.

A question was included in the survey to map the rare diseases the researchers are working on. Altogether, the respondents mentioned 629 rare diseases and 109 disease groups. These diseases and groups map to 21 disease classifications out of the 35 Orphanet Disease Classification. More than half of the respondents (54.5%) are working in 3 major disease classifications: Rare neurologic diseases, Rare developmental defect during embryogenesis, Rare inborn errors of metabolism (Figure 2). This data suggests that researchers are working on a small percentage of all rare diseases.

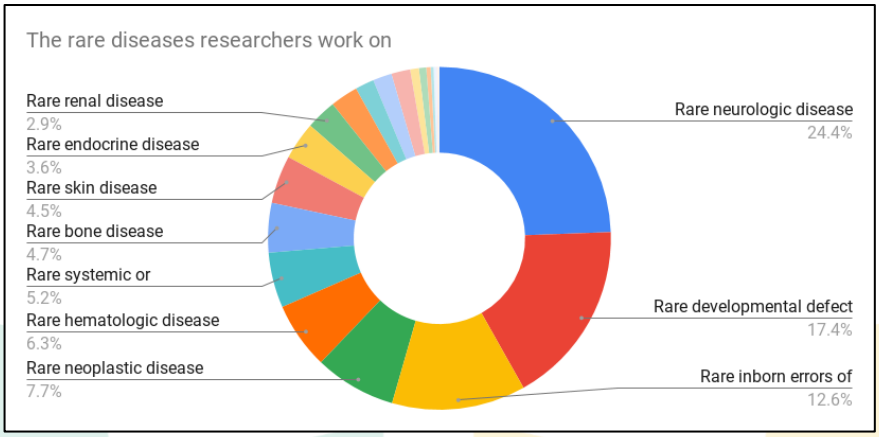


Figure 2. The rare diseases researchers are working on according to Orphanet Disease Classification

2.3.2. Data generation and storage

RD researchers collect and use multiple types of data, where the highest ranked data type use or collected are biological samples, phenotypes and genotypes. This ranking is different from the ERN most used and collected data types: patient information, natural history / follow up and biological samples. The researchers use fewer data types related to patient health, and the difference most likely arises from the different nature of the research performed by clinicians and researchers.

Over half of the researcher respondents said they generate or use –omics data (55%). Some 27% of respondents indicated they have multi-omic datasets originating from the same samples. The most common – omics generate/used are *genomics* and *transcriptomics* data, with the data most frequently generate from whole exome sequencing and RNA-Seq respectively. A very small percentage of respondents (< 5%) used the listed tools or resources for data deposition, sharing and analyses, such as RD-Connect GPAP, EGA, ArrayExpress, PRIDE or Metabolights.

On the collection or use of data related to environmental factors, only 20% indicated yes compared to 64% who indicated no. Around 27% of respondents said they have rare disease data and 13% on metabolic signalling for pathways. Although a third of respondents with these data indicated they were not able to judge whether data could fit an existing pathway or structured into a new one.

In terms of general research data storage, the most common method is having files or databases at host institutions' servers and/or personal computers. Only 21% of respondents deposit and/or share -omics data in open or controlled access resources. The rest of the respondents either indicated no sharing (38%) or did not answer the question. The major hurdle that limit data sharing was the complexity of use of resources.

2.3.3. Data annotation and FAIRification

The responses in this section were very similar between researchers and ERNs. Only approximate 35% of researchers indicated use of ontologies or standards to annotate data (compared to 32% who said no and 33% who did not answer the question). The most used standards were in this order: OMIM, ICD-10, HPO and ORDO (the order of

use for ERNs were ICD-10, OMIM, ORDO, HPO). Of the respondents who said they do not use standards/ontologies, the most indicated reason was because they did not know standards / ontologies existed.

Like the ERNs, 45% of the researcher respondents feel that they do not sufficiently understand what FAIR research data means. No resources available for FAIR effort was the most cited reason as a barrier to FAIR (70% of those who were interested but faced barriers).

Taken together, the responses in this section identified a need to improve awareness in data standardisation and its significance.

2.3.4. Use of existing resources and infrastructure for research data

Researchers selected facility to find biobanks / biosamples / cell lines and support to conduct translational research as the utmost important services they need for research (indicated by >50% of respondents). This was very different from the ERNs, who indicated facility to find patient registries and support to conduct clinical trials as the utmost important services. Researchers and ERNs require different services for research, which again is most likely due to the differences in the type of research they perform. Of note, significantly fewer researchers and ERN clinicians felt facilities to implement semantic standards or facilities to deposit –omics data as of utmost importance (< 20% of respondents).

In terms of existing resources and research infrastructures, only 10-20 % of respondents indicated the listed rare diseases resources on registries, biobanks / biological materials, mouse models, bioinformatics and translational/clinical research support as of importance for research. Strikingly, 40-55% of respondents replied they did not know about these resources and 30-40% did not reply to the questions. The responses suggested a lack of awareness on the existing infrastructures and resources with which they had previously indicated as important services for their research.

2.3.5. Key findings of the 2020 Researcher Survey

- A major pan-European surveys on rare disease researchers with more than 450 responses.
- -OMICS: *Genomics* and *transcriptomics* are the most used/collected data.
- Researchers & ERNs can have different research needs or priorities.
- Lack of general understanding on standards and data FAIRification concepts – but there are interests.
- Experience general challenges finding reusable data, sharing, analysing data and finding expertise
- Lack of general awareness on Pillar 2 resources or EU research infrastructures.

2.4. Exploitation & next steps

The outcomes of the 2020 Researcher Survey were presented in several meetings: Pillar 2 Retreat 4 - 6 May 2020, ELIXIR All Hands 8-10 June 2020, EJP RD ExCom Meeting 7 July 2020, and EJP RD GA 14 – 17 September 2020. In addition, video registration of the Retreat presentation was shared with IRDiRC Operating Committee to stimulate ideas for the IRDiRC Roadmap. Dissemination and exploitation of the survey results are important as needs and gaps identified in the surveys cannot be met by efforts of Pillar 2 alone through technical development. Integrated actions from multiple work packages and parties are required to encourage a systemic practice in standardising and sharing of rare disease data as a part of the research process. These actions could include guidelines from funding agencies, targeted training programs, increased interactions with Pillar 2 partners and communication/dissemination of Pillar 2 products.

To capture specific needs of ERNs and researchers in the next phase, structured interviews to interested stakeholders would be carried out throughout the year as a part of the WF Use Cases. The selection of the use cases will in part driven by the results of the surveys and in part based on functional requirement needs for VP development.

3. GDPR implementation in Pillar 2 report

3.1. Background

Compliance with GDPR and data protection regulations in broader sense is one of the clear goals of the EJP RD Virtual Platform (VP). The overall regulatory oversight in the project is provided by the AREB. The purpose of this task is to prepare practical guidance for the Pillar 2 for regulatory compliance and ensure consistency across the Pillar 2 how the compliance is achieved.

3.2. Methods

In this monitoring period we have focused on development of the Pillar 2 guidelines with respect to the GDPR, focusing on the main aspects: (1) identifying roles with respect to GDPR for different entities of the ecosystem of the VP, (2) clarifying role of data controller and guidance to institutional vs. personal responsibilities, (3) clarification of various practical aspects of data controllership, and (4) explanation of consequences of anonymization. Furthermore, the guidance is provided for the two main parts of the data managements cycle: (a) for the initial project preparation phase where the legal situation is set up and (b) for the routine data management phase covering also situation after the project finishes. The whole guidance document has been thoroughly revised in the beginning of 2020, with further feedback collected via AREB in the Q1-Q2. The document has been finally updated based on the feedback from AREB at the end of Q2.

3.3. Achievements

The guidance document has been delivered and is currently prepared for journal publication, so that it can be reference from within as well as from outside of the EJP RD project; this strategy has been recommended by AREB.

3.4. Next Steps

- Definition of roles and data
- Processing personal data
- Execution of data subject rights
- Anonymization and processing anonymous data
- Compliance of infrastructure

4. Quality oversight report

4.1. Problem

To assess the quality (fitness for purpose) of the tools and resources in Pillar 2 applying the quality criteria identified in year 1.

4.2. Background

To provide information of the quality of the various tools and resources included in the Virtual Platform (VP) to ensure the they are fit for purpose and to give users confidence in the results.

4.3. Methods

Design a webform to capture the relevant information from the Lead of each tool or resource identified as being of value to the Virtual Platform. This form should be straightforward and unambiguous to encourage maximum engagement from participants.

To assess the effectiveness of this form initially on tools and resources identified in the ELIXIR Rare Disease (RD) Bundle (link), using information taken from the public domain (website etc).

4.4. Achievements

The draft form was completed using the list in the Rare Disease Bundle (**Annex 2**) and the initial results presented at the EJP-RD General Assembly (**figure 3**). The majority of the required information for those tools listed in the RD Service Bundle is available online. This suggests that the selected tools and resources are of acceptable quality, as would be expected. The outstanding information will be sought to complete the form.

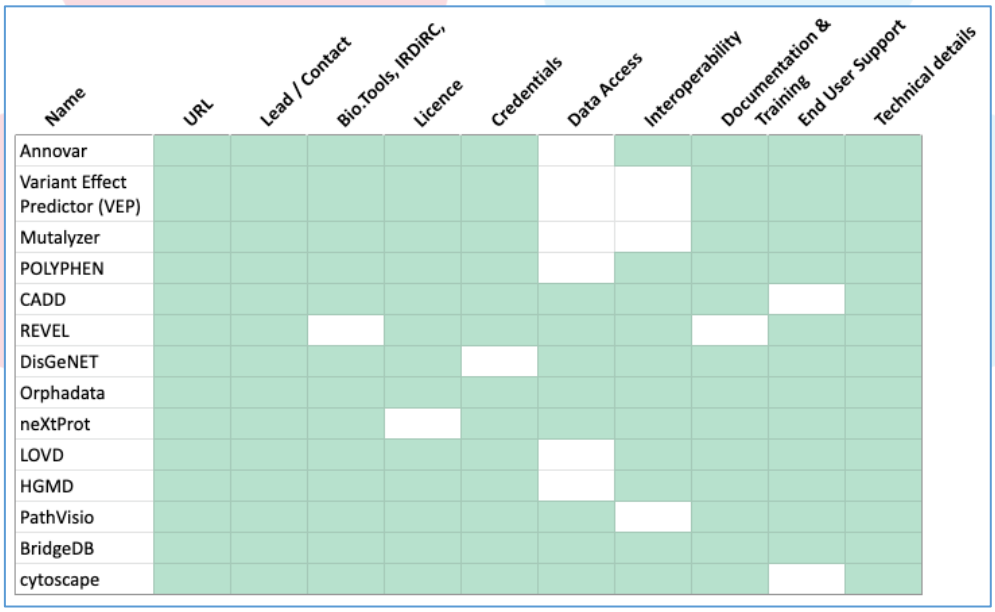


Figure 3. Publicly available quality evaluation results of the ELIXIR Rare Disease (RD) Bundle tools and resources

4.5. Next Steps

- Refine the form questions based on initial results (e.g. Licensing [academic / commercial]; AAI; GA4GH specifications; How best to present information)
- Send pre-filled form (public information) to RD Bundle Leads for completion
- Share results with VP for comment / evaluation
- Contact the wider group of tools and resources contributing to the VP

5. Contribution from Sustainability report

5.1. Problem

Many of the outputs of the EJP RD, including Pillar 2's, will have value when the project ends. Each of these outputs will require explicit thoughts on how they will be sustained after the project. Sustainability in a public health program is important for four main reasons: (1) sustained programs can maintain their effects for a long time, (2) there is often a latency period between the beginning of program-related activities and their effects on population health, (3) absence of sustainability can lead to an investment loss for the organisations and people involved, and (4) discontinuation of program-related activities may bring disillusion to participants and make subsequent community mobilisation difficult. In order to give a long-term perspective to actions contributing to visibility, registration, surveillance, and knowledge dissemination of RD, there is a need to use legal and funding tools

5.2. Background

Without a proper sustainability plan, any services or other resources that are created within a project may no longer be systematically developed and maintained after the project's end. In that case, they will be losing their value, and potentially have to be redeveloped in new projects. To prevent such thing from happening to the EJP RD VP and its services, financing models to sustain and scale up the development and use of the VP are being studied. Program implementation and sustainability should not be distinct and successive phases but should be concomitant processes in order to take into account the recursive or reflexive character of sustainability and learning or of the continuous adjustments that shape the sustainability process

5.3. Methods

We plan to organise reviews between the participants in the architecture Work Focus and experts on sustainability in WP3 "*Sustainability strategy and business plan*", in order to make sure that sustainability aspects are considered whenever architectural decisions are made.

5.4. Achievements so far

The work on sustainability in WP10 is done in close collaboration with the transversal WP3 dedicated to sustainability. In WP3 a survey was run for all WPs asking them to describe elements used for or developed that will require preservation. Over 90 elements have so far been identified, and prioritization of the activities surrounding their sustainability is being undertaken. Several of the elements are coming from pillar 2 and are components of the virtual platform that is under development.
In WP10, since October 2019 a few meetings dedicated to the sustainability of the virtual platform were held. In the meantime, these have now been merged with the

Quality oversight, GDPR and other so-called *non-functional* aspects into a biweekly meeting on the non-functional requirements of the virtual platform.
We have discussed about the best way to stimulate interaction between the developers and architects of the Virtual Platform and experts on GDPR, quality and sustainability. A template presentation is being worked out, which we will ask the architects and developers of the components of the virtual platform to fill (Annex 3).

5.5. Next steps

In the coming time, we will complete the template presentation, and ask the groups delivering components to the virtual platform to fill that template. In a series of review meetings, WP3 along with EJP RD respective legal services (including those of the partner 86-Inserm-transfert) will consult the architecture of the virtual platform about sustainability aspects.

6. Annex 1 - List of questions in the 2020 Survey for the RD researchers.

About	1. Are you a researcher working on rare disease(s)?
	2. Your institution is located in the following country?
	3. Which rare disease(s) do you work on?
	4. You are involved in which type of research- [Basic research, including genetic research] [Translational/applied pre-clinical research] [Clinical research]
	5. The main purpose of your current research is: [to develop disease models] [to improve diagnostics] [to Identify disease modifiers (including natural history and biomarker studies)] [to develop novel therapies] [Other]
	6. Please indicate your current position: [Graduate Student] [PhD Student] [Postdoctoral fellow] [Research Assistant / Technician] [Laboratory Manager] [Principal Investigator] [Other]
	7. Do you have a researcher digital identifier - [I do not have a researcher digital identifier] [ORCID] [ResearcherID] [Scopus Author ID] [ISNI] [Other]
	8. To perform your research on rare diseases, do you face challenges in any of the following? [Finding (reusable) data for research][Finding suitable facilities / expertise] [Sharing research data] [Analysing research data][Storing research data] [Other]
	9. What are your research needs other than financial/funding considerations?
	10. From where did you hear about this EJP RD Research Community Survey? [EJPRD Website/Newsletter/Twitter] [Funding agency] [European Reference Network] [Learned society / professional association] [Research Infrastructure] [Word of mouth] [Other]
	11. Are you familiar with EJP RD actions and support services?
	12. Are you willing to be contacted by EJPRD partners for a detailed analysis of your research needs?
	12.1 If yes, please leave your email here:
A	A1. Which of the following data types do you collect or use? [Phenotypes] [Genotypes][Medical images] [Microscopy images] [Patient information][Electronic medical/health records] [Treatment of disease] [Natural History / Follow-up][Biological samples] [Other]
	A2. Do you generate or use –omics data in your work?
	A2.1 Genomics or other DNA level genetic variation data:
	A2.2 Transcriptomics:
	A2.3 Proteomics:
	A2.4 Metabolomics:
	A2.5 Epigenomics
	A3. Do you have different types of –omics data that originate from the same sample, from multi-omic studies?
	A4. Do you collect or use data related to environmental factors (e.g. nutrients, lifestyles, pollutants)?
	If yes, please list
	A5. Do you have any data on diseases / risk factors / drugs / metabolic / signalling pathways?
	A5.1 Which data?

	A5.2 Do you think this data could be added into an existing pathway or structured into a new one?
	A6. What best describes your current research data storage method? [Paper forms] [Information stored in a personal computer; excel sheets, SPSS/SAS-files, MS Access database] [Files hosted on my institute's IT infrastructure] [Files hosted in a cloud infrastructure] [Files hosted in a public resource/infrastructure] [Database on my institute's IT infrastructure] [Database hosted in a cloud infrastructure] [Database hosted in a public resource/infrastructure]
	A7 Do you use any of the following resources as your primary data deposition mechanism? [European Genome-phenome Archive (EGA)][RD-Connect Genome-Phenome Analysis Platform (GPAP)] [ArrayExpress] [PRoteomics IDEntifications (PRIDE)] [MetaboLights] [Other]
	A8 Do you use any of the following resources as your primary data sharing mechanism- [European Genome-phenome Archive (EGA)] [RD-Connect Genome-Phenome Analysis Platform (GPAP)] [ArrayExpress] [PRoteomics IDEntifications (PRIDE)] [MetaboLights] [Other]
	A9 Do you use any of the following resources as your primary data analysis mechanism- [European Genome-phenome Archive (EGA)] [RD-Connect Genome-Phenome Analysis Platform (GPAP)] [ArrayExpress] [PRoteomics IDEntifications (PRIDE)][MetaboLights] [Other]
	A10. Do you usually deposit and/or share your -omics data in open or controlled access resources?
	A11. What are the issues that are more limiting for you to deposit / share your data in such open or controlled access resources? [Complexity of use][Research Use Only][Research Purpose][Clinical Care] [No Ancestry Research][Geographical Location][Legal Jurisdiction][Not for Profit Only][Time-limited Use][Collaboration Required][Publication Required][Publication Moratorium / Embargo][No adequate resource for my data type to use][Other]
	A12. Have you used a Data Transfer Agreement (DTA) and/or a Material Transfer Agreement (MTA) that are compliant with GDPR for your research activities?
	A12.1 Can you describe your experience in processing such DTA/MTA in 3 words?
B	B1. Do you use ontologies or standards to annotate your data?
	If yes, which of these ontologies or standards do you use? [International Classification of Diseases (ICD)][Online Mendelian Inheritance in Man (OMIM)][Orphanet Rare Disease Ontology (ORDO)] [Monarch Disease Ontology (MONDO)][Human Phenotype Ontology (HPO)] [SNOMED CT][Experimental Factor Ontology (EFO)] [Data Use Ontology (DUO)][Logical Observation Identifiers Names and Codes (LOINC)][Minimum Information About Biobank data Sharing (MIABIS)] [Ontology for Biobanking (OBIB)][Other]
	If no, why do you not use standards / ontologies? [I did not know standards / ontologies existed][Too burdensome to adopt standards][Standards do not meet my needs; they are difficult to use][Standards do not meet my needs; I do not find the terms I need][Other]
	B2. Please describe your interest in making your research data FAIR* [I don't sufficiently understand what "FAIR research data" means.][I am currently not interested in this kind of activity.] [I am interested, but I am facing some barriers to get involved.] [I am in the process of FAIRifying my data.] [Not applicable]
	What could raise your interest- [Seeing added value in FAIRifying my research Data] [Having the advantages of making my data FAIR outweigh the drawbacks (such as the resources to be committed, etc.)][Other]

	What are these barriers- [I do not have resources available for this effort][I do not understand what standards to use] [I do not have software or servers to make the data safely and accurately discoverable/sharable][I do not have consent/authority to make my data more widely accessible and useful by others][Other]
	In what stage is your FAIRification process- [Initiation: defining objectives, scope, purpose and deliverables][Planning: Allocating time and resources to tasks] [Execution: tasks undergoing towards FAIR compliance] [Closed: my data is already FAIR (i.e. FAIR for humans and machines)][Other]
	What type of data does your FAIRification involve- [Patient information] [Electronic medical/health records][Natural history / follow-up][Other]
C	C1. Which of these services are of utmost importance for your research- [Facility to find patient registries][Facility to find biobanks / biosamples / cell lines][Facility to find software and tools][Facility to find animal models][Facility to deposit your genomics, phenomics and other 'omics data][Support to conduct translational research][Support to conduct clinical trials][Facility to implement semantic standards (to describe phenotypes, diagnosis, genomics)][Databases of re-usable data related to rare diseases (metabolomics, proteomics, genomics)][Tools for personal data anonymisation / pseudonymisation] [Other]
	C2. Are these research infrastructure resources of utmost importance for your research- [BBMRI-ERIC Directory of Biobanks] [BBMRI-ERIC Negotiator][RD-Connect Sample Catalogue][RD-Connect Registry and Biobank Finder] [EuroBioBank][Orphanet / Orphadata] [ELIXIR bio.tools][ELIXIR European Genome-Phenome Archive (EGA)][INFRAFRONTIER][EATRIS] [ECRIN]
	C3. Which of these databases and catalogues are of utmost importance for your research- [European Rare Disease Registry Infrastructure (ERDRI)][Rare Disease Cohorts Programme (RaDiCo)] [Orphanet's catalog of RD expert resources][hPSCreg] [Cellosaurus]
	C4. Which of these data management and analysis resources are of utmost importance for your research- [Orphanet nomenclature (ORPHAcodes) and ORDO] [Human Phenotype Ontology (HPO)][RD-Connect Genome-Phenome Analysis Platform][PhenoMeNal][neXtprotKB][DECIPHER][European Patient Identity Services (EUPID)]
	C5. Any other resources that you need and use that were not listed in the previous questions?

7. Annex 2 – Form designed to capture the quality of the various tools and resources included in the Virtual Platform (VP)

EJP-RD Quality 10.4

A survey to assess the level of maturity, stability, reliability and usability of the software tool or service, giving users necessary information to make a decision.

Please tick all that apply, use 'Other' for any further information, links etc.

* Required

1. Name of the tool, service or product: *

Enter your answer

2. URL: *

Enter your answer

3. Name & email address for the lead person or main contact point. *

Enter your answer

4. Available in Bio.Tools or similar registry *

- Included in Bio.Tools. (if "Yes" the form will skip a number of questions)
- Included in IRDiRC Recognised Resources. (https://irdirc.org/research/irdirc-recognized-resources/)
- Included in ELIXIR Core Data Resource or Recommended Interoperability Resource list
- Included or planning to include in a similar registry (eg. Dockstore / BioContainers etc)
- Other

5. Purpose / Outline (if not available via website. Max 100 words.)

Enter your answer

6. Licence (does it matter what the licence is? Is it enough to know that there is a licence in place and link to it?) Note: this overlaps with a question on Sustainability presentation.

- Public Domain (Code can be used & modified without any restrictions)
- Free or Open Source (FOSS)
- Creative Commons (eg. CC by 4.0)
- Copyleft (aka reciprocal or restrictive licenses, eg GPL)
- Proprietary (Pay for licence, code may not be modified or redistributed)
- Other

7. Credentials

- Have established a Scientific Advisory Board or similar oversight body
- Published specification subject to peer review (eg. EuropePMC)
- Standalone publication (eg. F1000)
- Part of the work of a clearly identified user community
- Is seen as default / norm (i.e. accepted to the point of not being cited)
- Guidelines available for referencing this in publication
- Other

8. Further details (eg. Link to publication or the community mentioned above). Is this helpful?

Enter your answer

9. Accessibility of data

- Data Access Committee to manage access & processing of data
- Sign-in via LifeScience AAI and/or ELIXIR AAI
- Sign-in via alternative process (OAuth, Google etc)
- Guidelines available for use of data (to meet GDPR, consent or other requirements)
- Does not apply
- Other

10. Interoperability (for further details please use 'Other')

- Uses existing and standard input/output formats
- Use / supports data using a published ontology
- Marked up in line with bioschemas.org
- Other

11. Documentation & Training

- Documentation (eg. Read The Docs)
- Release notes (versions)
- Training materials
- Webinar, video or similar online resource
- Materials available in TeSS (tess.elixir-europe.org)
- Other

12. End User Support

- Point of contact (Lead or other individual)
- Dedicated support desk (Team)
- User community
- Consultant
- Other

13. Technical details (this feels extensive, are all necessary?)

- Has a persistent global, unique identifier for each released version
- Source code available (eg. GitHub, GitBucket, SourceForge)
- Published release cycle (even if more aspirational than actual)
- Development plans including 'Feature wish list'
- Performance data is available (eg. Average uptime)
- Generation of Errors, Warnings and Log Files
- Regular software testing
- Other

14. EJP-RD is a GA4GH Driver Project involved in the development, implementation and championing of the GA4GH Standards. (Further details on these are available in the ga4gh.org/genomic-data-toolkit)

	Has implemented	Are planning to implement	Working with GA4GH in developing standard
Data Use Ontology (DUO) github.com/EBISPOT/DUO	○	○	○
Data Repository Service (DRS) tinyurl.com/GA4GH-DRS	○	○	○
GA4GH Passports tinyurl.com/GA4GH-Passport	○	○	○
Phenopackets tinyurl.com/Phenopackets	○	○	○
RNAget API tinyurl.com/RNAget	○	○	○
ga4gh-service-info	○	○	○
Service Registry	○	○	○
Tool Registry Service API (TRS)	○	○	○
Beacon API	○	○	○
CRAM File Format	○	○	○
Crypt4GH	○	○	○
Family History Tools Inventory	○	○	○
htsget API	○	○	○
refget API	○	○	○
SAM/BAM File Formats	○	○	○
Variant Benchmarking Tools	○	○	○
Variation Representation	○	○	○
VCF v4 / BCF v2 File Formats	○	○	○
Workflow Execution Service (WES) API	○	○	○

15. Comments & thoughts regarding this Survey

Enter your answer

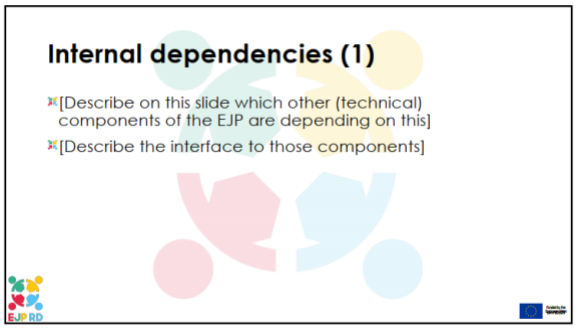
Submit

This content is created by the owner of the form. The data you submit will be sent to the form owner. Never give out your password.

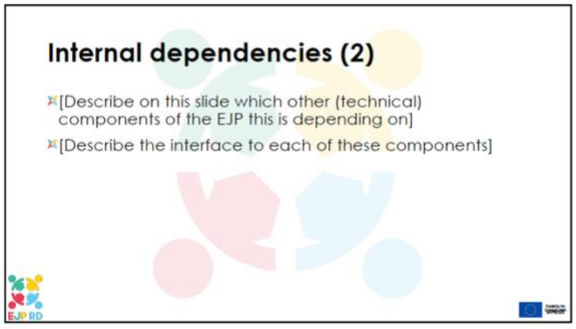
Powered by Microsoft Forms | Privacy and cookies | Terms of use

8. Annex 3 – Sustainability Template to be filled by the architects and developers of the Virtual Platform Components

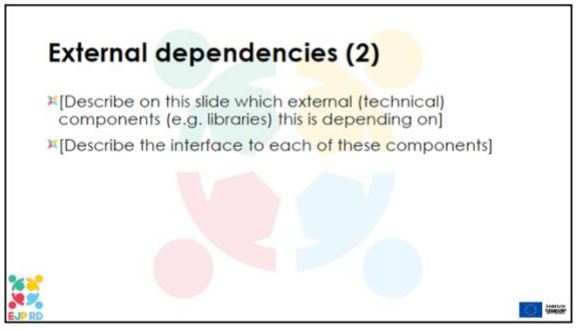
1



2

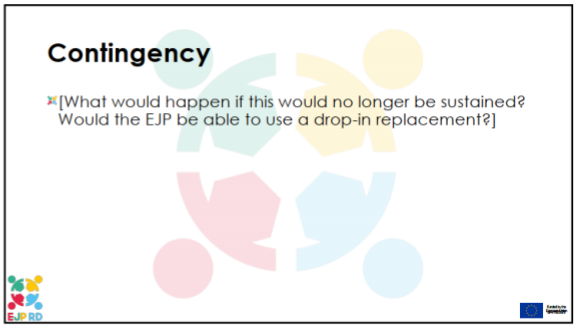


3



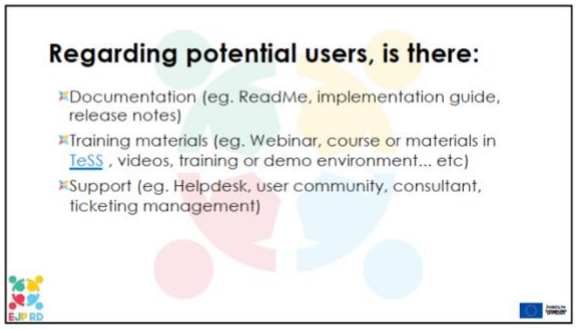
4

5

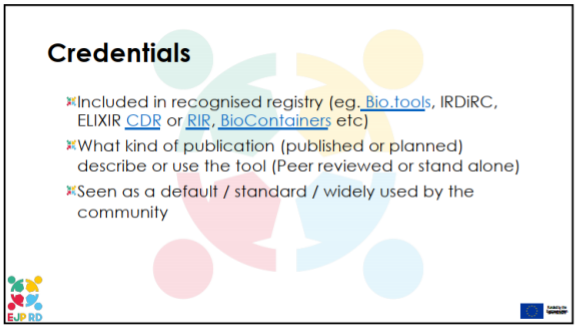


6

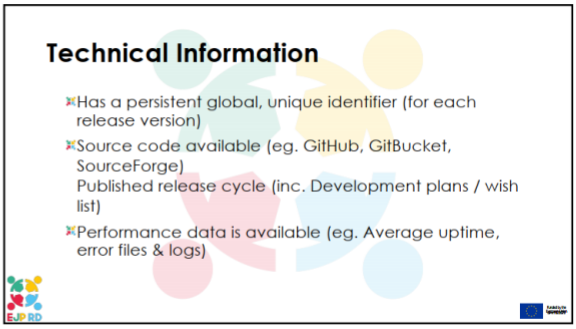
7



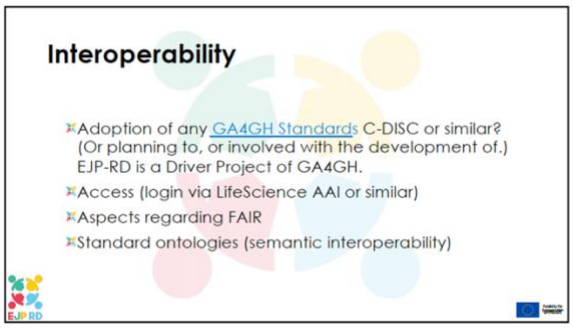
8


Credentials
Included in recognised registry (eg. Bio.tools, IRDiRC, ELIXIR CDR or RIR, BioContainers etc)
What kind of publication (published or planned) describe or use the tool (Peer reviewed or stand alone)
Seen as a default / standard / widely used by the community

9


Technical Information
Has a persistent global, unique identifier (for each release version)
Source code available (eg. GitHub, GitBucket, SourceForge)
Published release cycle (inc. Development plans / wish list)
Performance data is available (eg. Average uptime, error files & logs)

10


Interoperability
Adoption of any GA4GH Standards C-DISC or similar? (Or planning to, or involved with the development of.) EJP-RD is a Driver Project of GA4GH.
Access (login via LifeScience AAI or similar)
Aspects regarding FAIR
Standard ontologies (semantic interoperability)

11