

# **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

# **D11.13**

## **Third version**

# **Additional facilities integrated to resources regarding data deposition and access, including user guidelines and documentation**

**Organisation name of lead beneficiary for this deliverable:**

Partner 1 – EMBL-EBI (EGA)

**Contributors:** hPSCreg, Cellosaurus, Metabolights, INFRAFRONTIER, BBMRI-ERIC, RD Connect Sample Catalogue, RaDiCo, CNAG-CRG (RD Connect GPAP), DECIPHER, RD-Connect Registry and Biobank Finder.

**Due date of deliverable:** Month 36

**Dissemination level:** Public

## Table of Contents

<b>1. Introduction .....</b>	<b>4</b>
<b>2. Resources .....</b>	<b>4</b>
<b>2.1 European Genome-phenome Archive (EGA).....</b>	<b>4</b>
I. Resource data flow .....	5
II. Features or facilities added in 2021 .....	6
III. Plans for improvement in 2022 .....	7
<b>2.2 RD-Connect GPAP .....</b>	<b>8</b>
I. Resource data flow .....	8
II. Features or facilities added in 2021 .....	9
III. Plans for improvement in 2022 .....	10
<b>2.3 DECIPHER.....</b>	<b>11</b>
I. Resource data flow .....	11
II. Features or facilities added in 2021 .....	11
III. Plans for improvement in 2022 .....	12
<b>2.4 RD-Connect Registry and Biobank Finder .....</b>	<b>12</b>
I. Resource data flow .....	12
II. Features or facilities added in 2021 .....	13
III. Plans for improvement in 2022 .....	13
<b>2.5 RD-Connect Sample Catalogue .....</b>	<b>13</b>
I. Resource data flow .....	13
II. Features or facilities added in 2021 .....	15
III. Plans for improvement in 2022 .....	15
<b>2.6 BBMRI-ERIC Directory .....</b>	<b>15</b>
I. Resource data flow .....	15
II. Features or facilities added in 2021 .....	16
III. Plans for improvement in 2022 .....	17
<b>2.7 Rare Disease Cohorts (RaDiCo).....</b>	<b>17</b>
I. RaDiCo introduction .....	17
II. Resource data flow .....	18
III. Improvements made in 2021 .....	20
IV. Improvement planned for 2022 .....	20
<b>2.8 hPSCreg.....</b>	<b>20</b>
I. Resource data flow .....	21
II. Features or facilities added in 2021 .....	21
III. Plans for improvement in 2022 .....	22
<b>2.9 Cellosaurus .....</b>	<b>22</b>
I. Resource data flow .....	22

II.	Features or facilities added in 2021 .....	22
III.	Plans for improvement in 2022 .....	23
<b>2.10</b>	<b>INFRAFRONTIER.....</b>	<b>23</b>
I.	Resource data flow .....	23
II.	Features or facilities added in 2021 .....	24
III.	Plans for improvement in 2022 .....	24
<b>2.11</b>	<b>MetaboLights.....</b>	<b>25</b>
I.	Resource data flow .....	26
II.	Features or Facilities added in 2021 .....	26
III.	Plans for improvement in 2022 .....	26
<b>3.</b>	<b>Conclusion.....</b>	<b>27</b>

## 1. Introduction

This deliverable summarizes the additional facilities and features that were integrated during 2021 into the data deposition and access resources participating in EJP RD.

The aim of the EJP RD is to improve the integration, the efficacy, the production and the social impact of research on rare disease (RD) through the development, demonstration and promotion of Europe and worldwide sharing of research and clinical data, materials, processes, knowledge and know-how. To this end, Task 11.3 aims to serve the needs of EJP-RD partners and the overall RD community for the deposition, integration and storage of quality-controlled data and metadata by building on existing resources, including registries, patient cohorts, biobanks, cell lines, mouse models, raw omics data and genome-phenome platforms. Task 11.3 will guide data producers to submit data, making them discoverable through the platform, to suitable public repositories and resources.

The aim of Subtask 11.3.1 is to support the resources and infrastructures, both European and international, relevant to the RD Community. In this Subtask, work is being done to improve and expand deposition capabilities and access mechanisms of resources based on RD use-cases identified through community surveys in EJP-RD. In addition, developments are based on FAIR (Findable, Accessible, Interoperable and Reusable) principles, deploying or expanding user-friendly interfaces to deposit data and metadata using HPO, ORDO, OMIM and/or any other relevant ontology or standard. Moreover, quality mechanisms are being implemented to the datasets by introducing manual curation, automatic metric generation or a mixture of both, and Application programming interfaces (APIs) and graphical user interfaces (GUIs) are being implemented to allow for query functionality and data access.

Data security is a concern for many of the resources and infrastructures working in EJP RD that process potentially identifiable human data and is therefore being assessed and aligned with the recommendations of the Global Alliance for Genomics and Health (GA4GH), by using such examples as the Data Security Toolkit, which provides a principled and practical framework for responsible sharing of genomic and health-related data, whilst taking into account the EU GDPR and other national legislations. Where relevant, transparency measures and means of monitoring the re-usability of the submitted dataset will be implemented by the resources. The task also includes the implementation and further development of and interoperability of the federated EGA infrastructure with RD-related databases.

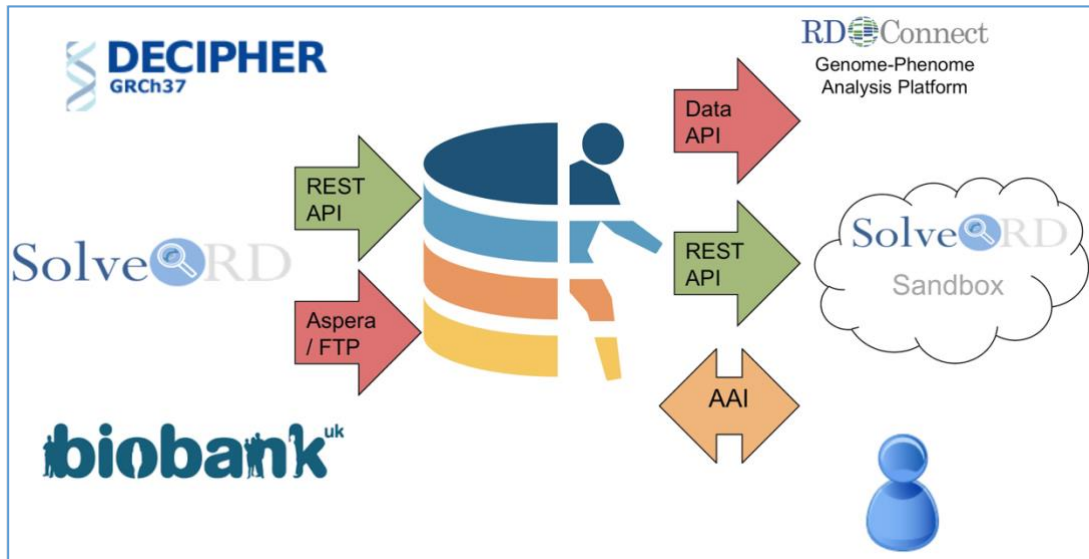
The objective of this Deliverable is to identify each of the current capabilities of the named resources and additional capabilities added during 2021 in terms of data deposition and access to data, to identify areas for improvement that will benefit the RD community and thus the community as a whole. This output will be used to scope and schedule the Work Focus' (WF) work on "Resources for the sharing of experimental data and materials" for the coming year.

## 2. Resources

### 2.1 European Genome-phenome Archive (EGA)

Contributors: Giselle Kerry (EMBL-EBI), Carles Garcia (EMBL-EBI), Mallory Freeberg (EMBL-EBI) Jordi Rambla (CRG), Thomas Keane (EMBL-EBI)

## I. Resource data flow

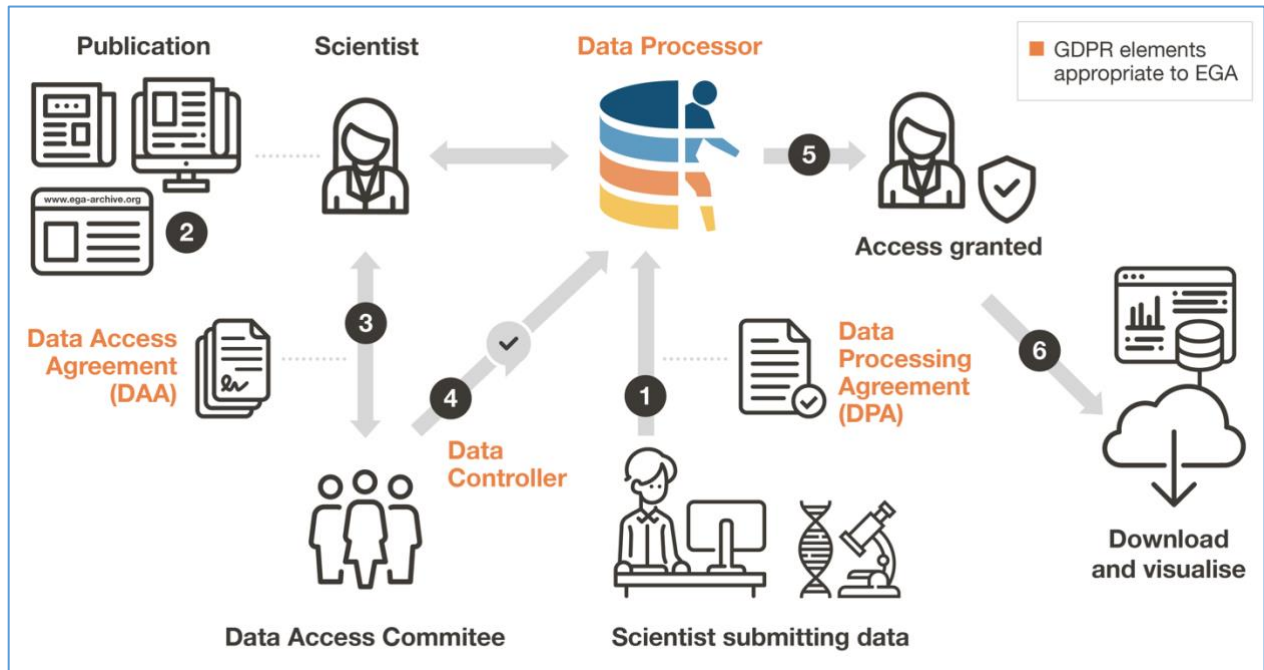


**Figure 1.** Example data flows to and from EGA. Submitters, such as Solve-RD or DECIPHER, submit data to the EGA for archive and distribution. These data are distributed via the EGA Data API to authorised users. Use-cases include distributing data to the RD-Connect Genome-Phenome Analysis Platform, the Solve-RD cloud-based sandbox, or individual users for local analysis. All users must authenticate prior to accessing data.

The EGA provides a service for the permanent archiving and distribution of personally identifiable genetic and phenotypic data resulting from biomedical research. Data submitted to the EGA is collected from individuals whose consent agreements authorise data release only for specific research. Submitters upload controlled access data, which has been encrypted before transmission to the EGA via the EGACryptor, using Aspera or FTP (Figure 1) to a specific submission account. The submitter will then submit open-access metadata, such as details on experimental methodology, file types, and high-level phenotypes via the EGA Submitter Portal or associated REST APIs. Once the metadata has been submitted and validated, the controlled access data is archived ready for distribution. Strict protocols govern how information is managed, stored, and distributed by the EGA, including statements ensuring the submitter has the ethical and legal authorisation to submit the data, recording and auditing of all data movements to and from the EGA, and ensuring the controlled access data is encrypted during transmission and at rest.

The General Data Protection Regulation (GDPR) is a European Union (EU) regulation that legislates how organizations can share and process personal data of EU citizens. Within GDPR, there are two main actors: data controllers and data processors. Data controllers are persons or entities which determine the purposes and means that the personal data may be processed, e.g., companies, researchers, or universities. For EGA, the data controller is ultimately the data producer and the submitter(s) who submit the data to EGA (Figure 2). The data controller also creates a Data Access Committee (DAC) who will decide on data access permissions at EGA. Data processors are the persons or entities which process the data on behalf of a data controller. Regarding GDPR, EGA is a data processor as it processes data as instructed by the data controller. GDPR applies to any organization which accesses personal data from an individual within the EU. Under GDPR, personal data is defined as any data that is identifiable, including names and email addresses as well as health-

related and genetic data. EGA does not accept personally identifiable data except genetic and phenotypic data, so all other data submitted to EGA, such as names and addresses, must be pseudonymized. GDPR requires that data controllers implement data protection principles, such as data minimization, to minimize the risk of data leakage, and protect the rights of the data subjects. As a data processor, EGA has a set of security policies that are followed to minimize the risk of unauthorized data access or data loss.



## II. Features or facilities added in 2021

EGA is continuously working on new features to improve users' data submission. For instance, a file manifest used for submission has been refined, and now it allows users to record all the metadata needed in only one user-friendly spreadsheet<sup>1</sup>, which will ease preparation of large submissions (e.g., thousands of samples, hundreds of experiments). In addition, the Star2xml tool has been developed to transform a submission spreadsheet into corresponding XML files which are used to programmatically submit to EGA. EGA created associated documentation<sup>2</sup> and a

**Figure 2.** EGA facilitates the submission, discovery, access, and distribution of sensitive human data. A researcher submits controlled access human genetic, phenotypic and clinical data to EGA after signing a Data Processing Agreement (1). EGA processes, archives, and releases the dataset to be findable. Another researcher discovers data of interest at the EGA (2). They contact the Data Access Committee for the data of interest and agree to the terms of data reuse by signing a Data Access Agreement (3). The Data Access Committee informs EGA that access is approved (4). The EGA grants access to the requesting researcher (5) who can then download and visualise the data (6). GDPR: General Data Protection Regulation.

<sup>1</sup> [https://github.com/EbiEga/ega-metadata-schema/blob/main/templates/sequence-based-metadata/EGA\\_metadata\\_submission\\_template\\_v1.xlsx](https://github.com/EbiEga/ega-metadata-schema/blob/main/templates/sequence-based-metadata/EGA_metadata_submission_template_v1.xlsx)

<sup>2</sup> <https://github.com/EGA-archive/star2xml>

video-tutorial<sup>3</sup> explaining how “Star2xml” is used. To encourage adoption amongst EGA submitters, EGA has recorded a new video-tutorial that revolves around how to perform programmatic submission to EGA, which allows for metadata to be submitted through an API<sup>4</sup>.

EGA has also created a public GitHub repository<sup>5</sup> to maintain the EGA metadata schemas and related documentation, including templates for metadata submission and descriptions of metadata attributes. These schemas and documentation indicate which ontologies are to be used to select values for attributes, for example to select values from ontologies like HPO, MONDO, and ORDO for the attribute “phenotype”. Making this information more transparent and accessible will enable users to prepare rich metadata more easily for submission to archives like the EGA, increasing the overall FAIRness of data. Maintaining EGA metadata standard in an open way also means other initiatives managing controlled access data can align their standards and specific use cases with the EGA, increasing interoperability of data/metadata.

During 2021 around 4,200 rare disease datasets containing genomic and phenotypic information have been added to EGA. In addition, in 2021 EGA received or updated around 15,000 Phenopacket submissions (an open standard for sharing disease and phenotypic information containing ontologies like HPO, OMIM or ORDO). EGA has worked with rare disease consortia such as EJP RD and Solve-RD to streamline submission, validation and metadata update processes to make more rare disease data available for the community. One example would be the new system for improved findability and interoperability that exports file and sample metadata directly into other data repositories, like the EJP RD Sandbox and RD3 platform (UMCG) from Solve-RD.

Finally, EGA has presented a webinar<sup>6</sup> as part of the “Resources Webinars” series organized by EJP RD, with the aim to explain to the Rare Disease community what the EGA is, and how can it be useful for clinicians/researchers involved in rare disease studies.

### III. Plans for improvement in 2022

In 2022 EGA will continue to work on improving efficiency and quality of data deposition through improving metadata representations, while continuing efforts to improve data access.

Work will be done to improve the Submitter Portal and submission tools, which is basis for maintaining the high quality of the data submitted to the platform. For instance, EGA will add ontology validation tools in the workflow, like OLS<sup>7</sup> (or the OxO service<sup>8</sup> which will be evaluated), to support submitters in the validation of the ontology terms provided. In addition, tools involving DACs (Data Access Committees) will be developed to make data management in the platform easier.

The EGA was one of the first deposition resources to implement the GA4GH Researcher Passports standard, and in 2022 we will continue the planning and work to be compatible with Lifescience AAI<sup>9</sup> when it becomes available (to replace ELIXIR AAI),

---

<sup>3</sup> <https://embl-ebi.cloud.panopto.eu/Panopto/Pages/Viewer.aspx?id=58d299c7-7e25-4b48-8fc3-ad18011ef0b4>

<sup>4</sup> [https://ega-archive.org/submission/sequence/programmatic\\_submissions](https://ega-archive.org/submission/sequence/programmatic_submissions)

<sup>5</sup> <https://github.com/EbiEga/ega-metadata-schema>

<sup>6</sup> <https://www.ejprarediseases.org/event/european-genome-phenome-archive-ega/>

<sup>7</sup> <https://www.ebi.ac.uk/ols/index>

<sup>8</sup> <https://www.ebi.ac.uk/spot/oxo/>

<sup>9</sup> <https://zenodo.org/record/3386307#.YaTuuVOnwll>

which will expand user access across the platform by enabling users to federate their identity.

During 2022 EGA will continue to work on projects involving rare disease data to improve the processes and tools employed for adding them into the platform. One example are Phenopackets, which have been integrated in the EGA submission pipeline, and for which EGA plans to develop additional user-facing documentation to encourage widespread adoption by rare disease submitters.

EGA will work with the Query Builder and Metadata Work Foci to evaluate its integration into the EJP-RD Virtual Platform. Finally, EGA will continue to work on the development of the Federated EGA (FEGA), a network of human data repositories across different countries that allows for federated discovery and access of sensitive human data while allowing the data itself to be stored locally under its governing jurisdiction. The FEAGA will continue to expand to include additional functional "nodes" in the network, providing a technical and policy framework by which rare disease data can become more findable and accessible, rather than remaining in silos where it is more difficult for researchers to find.

## **2.2 RD-Connect GPAP**

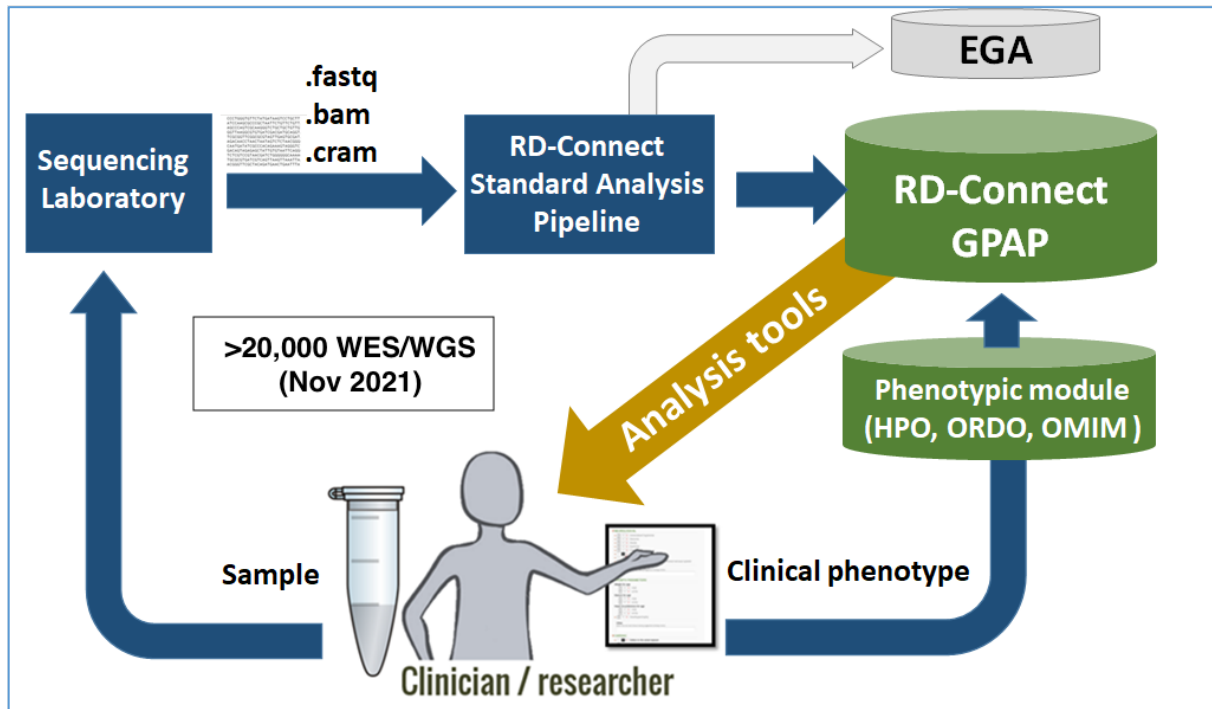
Contributors: Sergi Beltran (CNAG-CRG), Steven Laurie (CNAG-CRG), Davide Piscia (CNAG-CRG), Luca Zalatnai (CNAG-CRG)

### **I. Resource data flow**

The RD-Connect GPAP is a sophisticated and user-friendly online analysis system for RD gene discovery and diagnosis. The RD-Connect GPAP is an IRDiRC recognized resource hosted at the CNAG-CRG.

De-identified phenotypic data is collected using HPO, ORDO and OMIM ontologies through custom templates implemented through the RD-Connect GPAP-Phenostore module. Pseudonymized experiment data (exomes and genomes) and metadata are collected in the RD-Connect GPAP and processed using a standardized analysis and annotation pipeline. Integrated genome-phenome results are made available to authorized users for prioritisation and interpretation of genomic variants in the RD-Connect GPAP. Raw genomic data is deposited at the EGA for long-term archive and controlled access.





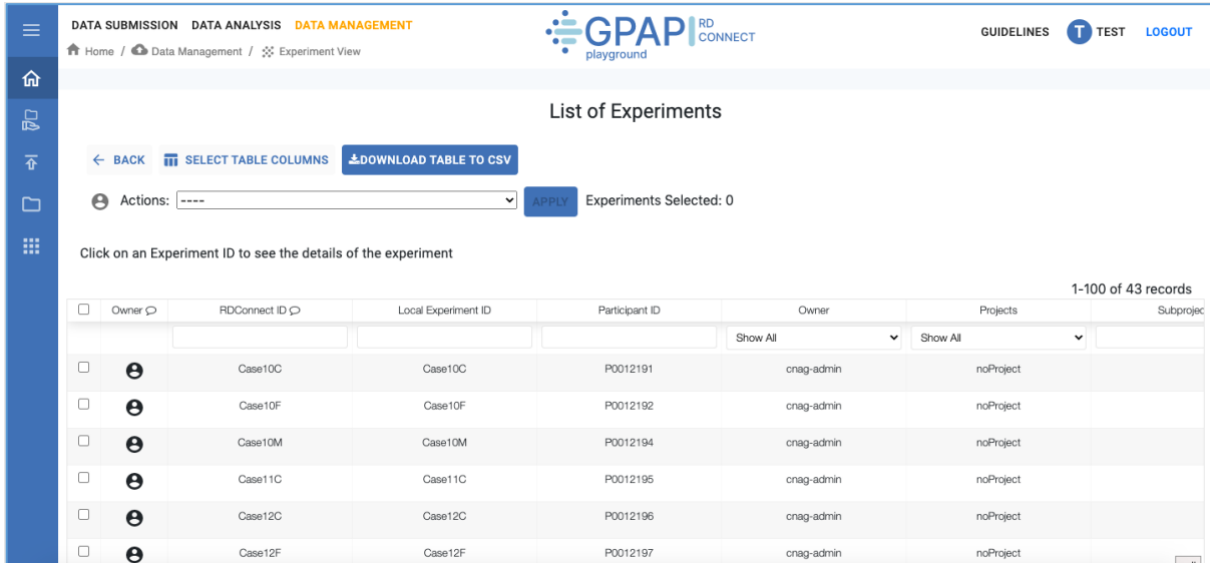
**Figure 3.** Data flow for RD-Connect GPAP

## II. Features or facilities added in 2021

In 2020 we finalized the implementation of Keycloak as the main User Management System for all RD-Connect GPAP modules. This has allowed us to implement ELIXIR AAI in 2021, which we have done for the RD-Connect GPAP Playground instance (<https://playground.rd-connect.eu/genomics/>). We will implement AAI in the production instance once the use cases within the Virtual Platform have been clearly defined.

The new RD-Connect GPAP Data Management (DM) Client module has been rewritten in SPA (Single Page Application), and the look and feel is now aligned with the other GPAP modules (see Figure 4 below). This has resulted in a marked improvement in the response time for metadata related queries at the level of an experiment as there is no need to reload the page when changing views. We have also introduced a new view within the experiment view, where a user can see all the information related to a particular experiment and the status of the experiment within the workflow (e.g., whether variant calling has been completed, and whether the experiment has been uploaded to ElasticSearch).

DM administrators can now see which files related to an experiment have been uploaded to the RedIRIS Aspera server. The information is reported by group and provides a useful overview for the platform administrators of the status of raw data upload. Administrators can also view which experiments are currently being processed, whether the gVCFs have been generated successfully, and identify for which experiments, if any, there has been a problem in processing.



The screenshot shows the 'List of Experiments' interface. At the top, there are navigation tabs for 'DATA SUBMISSION', 'DATA ANALYSIS', and 'DATA MANAGEMENT'. The breadcrumb trail is 'Home / Data Management / Experiment View'. The interface includes a sidebar with navigation icons, a top navigation bar with 'GPAP RD CONNECT playground', 'GUIDELINES', 'TEST', and 'LOGOUT'. Below the navigation, there are buttons for 'BACK', 'SELECT TABLE COLUMNS', and 'DOWNLOAD TABLE TO CSV'. An 'Actions' dropdown menu is set to '----' with an 'APPLY' button and 'Experiments Selected: 0'. A note says 'Click on an Experiment ID to see the details of the experiment'. The table shows 1-100 of 43 records. The table columns are: Owner, RDConnect ID, Local Experiment ID, Participant ID, Owner (with a 'Show All' dropdown), Projects (with a 'Show All' dropdown), and Subproject. The data rows are as follows:

Owner	RDConnect ID	Local Experiment ID	Participant ID	Owner	Projects	Subproject
Case10C	Case10C	Case10C	P0012191	cnag-admin	noProject	
Case10F	Case10F	Case10F	P0012192	cnag-admin	noProject	
Case10M	Case10M	Case10M	P0012194	cnag-admin	noProject	
Case11C	Case11C	Case11C	P0012195	cnag-admin	noProject	
Case12C	Case12C	Case12C	P0012196	cnag-admin	noProject	
Case12F	Case12F	Case12F	P0012197	cnag-admin	noProject	

**Figure 4.** New look of RD-Connect GPAP Data management client module

New User-guidelines have been generated for the DM module which explain the data submission process to the user and how to use the DM application. They are accessible through the RD-Connect GPAP Homepage: <https://platform.rd-connect.eu/>.

The RD-Connect GPAP has implemented internal changes to accept CRAMs as raw sequencing data for the processing pipeline.

We have defined the experimental Quality Control metrics and functionalities to be reported for each exome or genome processed, and we are currently working on adding the corresponding fields to the database and developing the user interface to display this information.

We are working together with the EGA on linking datasets between the two platforms. At the level of alignment visualisation (using the GA4GH htsget standard), the RD-Connect GPAP stores the EGA ID for the files to allow BAM slices to be retrieved on-the-fly from EGA. At the level of raw data (FASTQ), the GPAP has created the necessary fields in the database to allow this connection. Further work is required to put the system into production.

We are working with hPSCreg to connect the two platforms through an API so that GPAP users can discover relevant cell lines in hPSCreg for genes identified as being of interest in experiments they are analysing in the GPAP. We currently have a draft implementation, but it has not been made available to users yet as it needs some further work.

### III. Plans for improvement in 2022

IBMT has been working with RD-Connect GPAP to link with hPSCreg in 2021, and it is planned that the interaction will be extended to INFRAFRONTIER in 2022 via a use case. The connection GPAP-INFRAFRONTIER is expected to be re-launched during 2022.

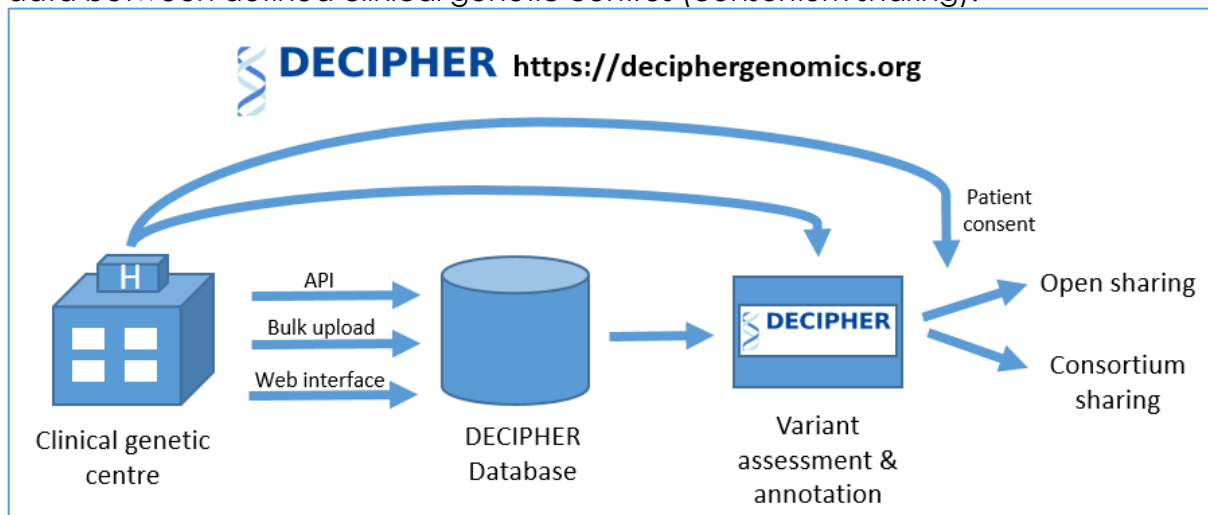
RD-Connect GPAP will work with the Query Builder and Metadata Work Foci towards its integration into the EJP RD VP. Also related to data discovery, RD-Connect GPAP will work on the implementation of the new Beacon standard (V2).

## 2.3 DECIPHER

Contributors: Helen Firth (DECIPHER), Julia Foreman (DECIPHER)

### I. Resource data flow

DECIPHER is a web platform that helps clinical and research teams to assess the pathogenicity of variants and to share rare disease patient records. Patient genotype and phenotype data is uploaded by academic clinical genetic centres worldwide, using the web interface, via bulk upload or through a deposition API. The DECIPHER web interface provides a suite of tools to assist users in assessing the pathogenicity of variants. Registered DECIPHER users at the depositing centre annotate the variants using the tools provided in DECIPHER. With explicit patient consent, the patient record is shared openly through the web portal. DECIPHER also supports the sharing of patient data between defined clinical genetic centres (consortium sharing).



**Figure 5.** DECIPHER Data Flow

### II. Features or facilities added in 2021

During 2021 a major new version of DECIPHER was released. This major release included the move from visualising genomic data in genome assembly GRCh37, to GRCh38. All deposited patient data was lifted over and re-annotated against the GRCh38 genome build, and all reference data was updated. Visualising the data in GRCh38 allows DECIPHER to use the most up-to-date gene build information and transcripts, to enable accurate variant interpretation. DECIPHER now promotes the use of MANE Select transcripts ([https://tark.ensembl.org/web/mane\\_project/](https://tark.ensembl.org/web/mane_project/)), a high-quality transcript which is 100% identical between Ensembl and Refseq. Deposition using GRCh37 coordinates is still supported, with DECIPHER lifting over variants to GRCh38. Interfaces to visualise the difference in builds are available, including GRCh37 and GRCh38 comparative genome browsers and a liftover mapping genome browser track. This year DECIPHER has also introduced the option for depositors to record for *de novo* mosaic variants, the proportion of cells carrying a variant for different tissue or sample types. This information is important clinically as this can affect variability of clinical symptoms. Online user guides and video tutorials have been updated as new features have been added to the website.

### III. Plans for improvement in 2022

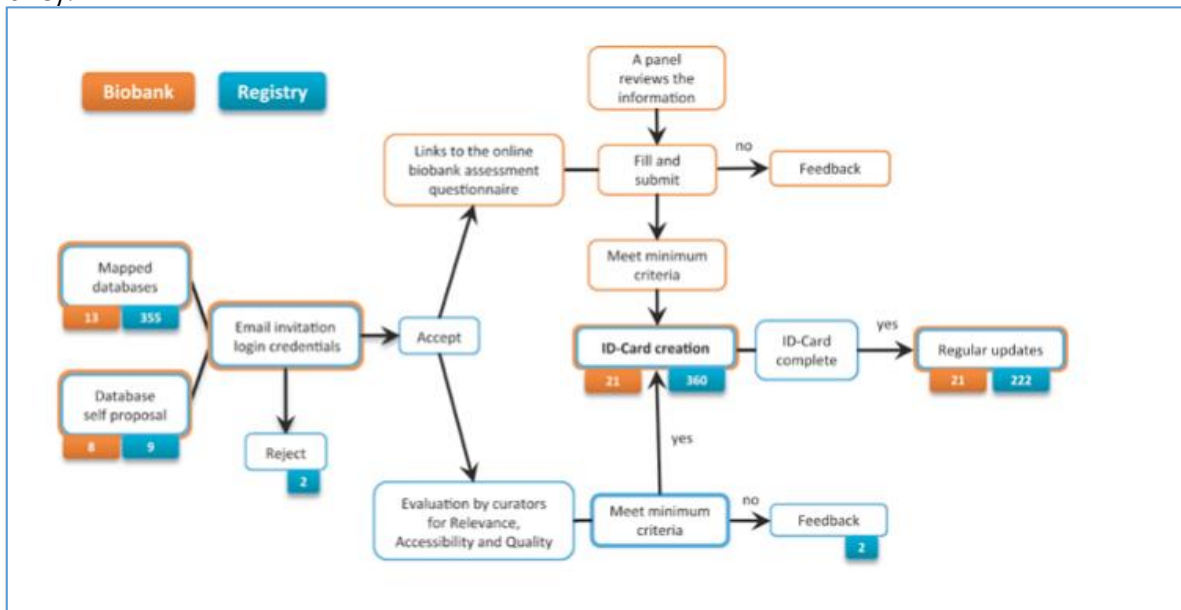
DECIPHER will continue to ensure the high quality of data deposition and that it meets the needs of the rare disease community. Online user guides and video tutorials will be updated throughout 2022 as new features are added to the website.

## 2.4 RD-Connect Registry and Biobank Finder

Contributors: Esther van Enckevort (UMCG), Mary Wang (FTELE), Heimo Müller (BBMRI), Vittorio Meloni (BBMRI-CRS4), Robert Reihls (BBMRI-MUG)

### I. Resource data flow

The initial data in the RD-Connect Registry and Biobank Finder was collected from several existing online resources such as the Orphanet Catalogue. Biobanks and registries were then invited to join the Finder. Next to this initial inclusion workflow the system also allows registration of new Biobanks and Registries through the Suggest a Biobank/Registry form. The biobank / registry is requested to provide general information about the institute, the disease focus, available data and/or samples and related documents such as SOPs and Consent forms through an online questionnaire. All registries and biobanks are assessed by a panel and if they meet the minimal requirements for inclusion an ID-Card is created (See workflow from Gianotti, et. al., 2018).



**Figure 6.** Inclusion of biobanks and registries in the Registry and Biobank Finder. The process of inclusion and evaluation of biobanks and registries in the Registry and Biobank Finder (mapped and self-proposed)

Gainotti, S., Torrer, P., Wang, C. M., Reihls, R., Mueller, H., Heslop, E., ... Taruscio, D. (2018). The RD-Connect Registry & Biobank Finder: A tool for sharing aggregated data and metadata among rare disease researchers. *European Journal of Human Genetics*. <https://doi.org/10.1038/s41431-017-0085-z>

## **II. Features or facilities added in 2021**

We worked on two aspects of the RD-Connect Registry and Biobank Finder: (a) we continued to develop the pilot for the federated querying of biobanks and registries. We added support to other query parameters, and we updated it to the v3 version of the API designed with the Query Builder WF; (b) we started, in a joint effort with RD Nexus team, developing a second pilot that allows to select a list of biobanks to query for samples. The backend of the RD-Connect Registry and Biobank Finder was migrated to the Molgenis framework. Migration of the data model to the Molgenis EMX format to ease integrated search across biobanks, registries and biobanks. The new version RD-Connect Registry and Biobank Finder is a central part of the Federated Query Builder WF (resource level).

## **III. Plans for improvement in 2022**

Evaluate the extension of the BBMRI-ERIC data model with a star schema to characterise collections which would enable fine-grained count searches to find how many relevant samples exist within a collection. Benefit from the merge with sample catalogue and directory to improve user experience via uniform search, filter and display behaviour. Connect to JRC and ERN registry projects to update the contents (joint effort with contents update on Sample Catalogue). Implement JSON-LD based semantic layer to enable federated querying, i.e., to add the EJP-RD common data elements as semantic annotation.

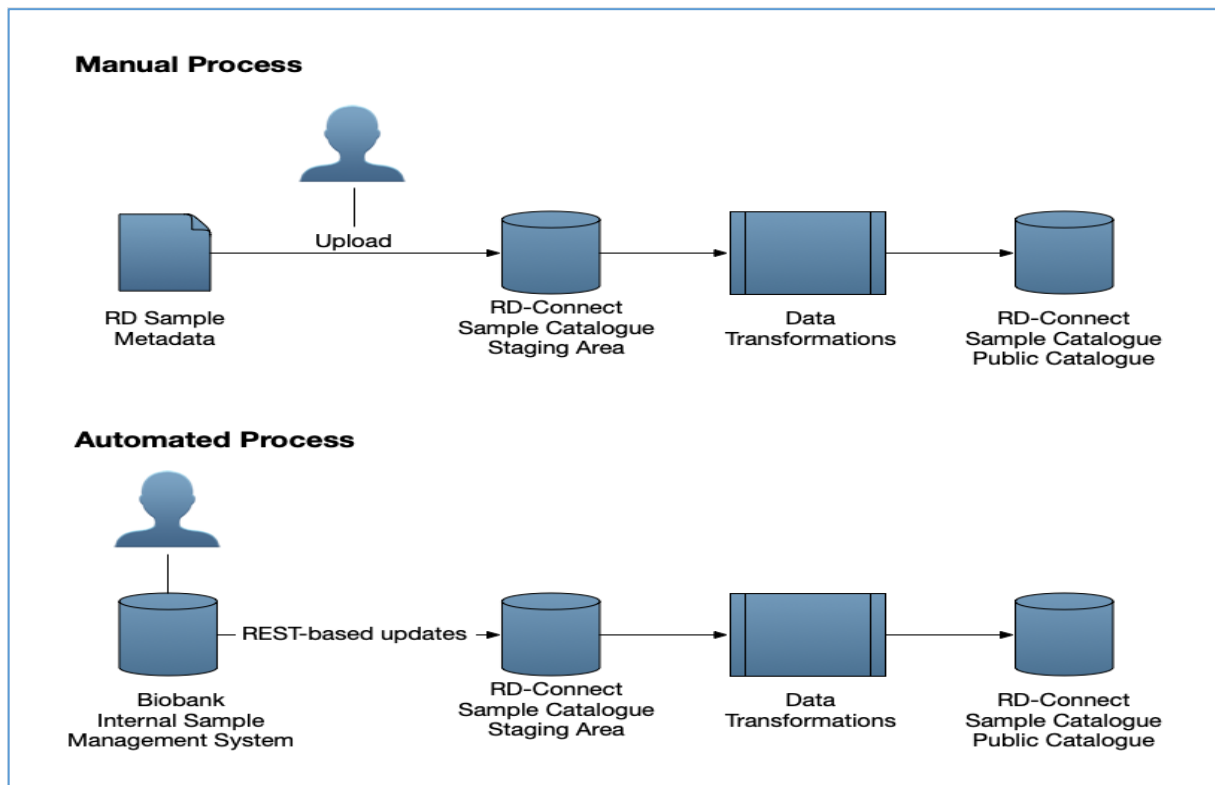
When collections or sub-collections (described with the star schema) are found, either locally or through a federated query in another catalogue, the user can start a negotiation with the owner of the resource through the BBMRI-ERIC negotiator. This important use case was already piloted in the Federated Query Builder WF. In 2022 this approach will be fully integrated, together with the BBMR-ERIC directory in the EJP-RD virtual platform, supporting the Life-Science AAI.

## **2.5 RD-Connect Sample Catalogue**

Contributors: Esther van Enckevort (UMCG), Mary Wang (FTELE)

### **I. Resource data flow**

The RD-Connect Sample Catalogue contains sample metadata for rare disease samples provided by the biobanks. There are two distinct workflows for the biobanks to add data to the catalogue. Most biobanks use the manual workflow where the biobank uploads an Excel file with sample metadata to the catalogue. The Italian TNGB network, however, has an automated workflow where the sample metadata is published into the catalogue automatically for each of the samples that have been released for publication in the sample catalogue.



**Figure 7.** RD Connect Sample Catalogue Data Flow

#### *Manual upload*

In the case of a manual upload the responsible person at the biobank extracts data from the internal sample management system into a Microsoft Excel or Comma Separated Values file to be uploaded into the Sample Catalogue. Together with the data managers from the UMCG that are responsible for the maintenance of the Sample Catalogue they describe the structure of this file to create a data model in the Sample Catalogue to support the upload as well as any data transformations needed to convert the data from the internal structure and encodings to the data model of the public sample catalogue. After this has been setup the file can be uploaded to a staging area in the catalogue and every night an automated job will run the transformations necessary to publish the sample data into the public catalogue.

#### *Automated workflow*

In the case of an automated workflow the biobank's internal systems have implemented the MOLGENIS REST API to publish data into the Sample Catalogue at the moment that they are released for publication into the internal system. During the implementation of this connection the developer and the data managers from the UMCG have agreed on the data structure for the data that is pushed to the Sample Catalogue as well as any data transformations needed to convert the data from the internal structure and encodings to the data model of the public sample catalogue. Once this system is deployed any changes in the internal system will be automatically pushed to a staging area in the catalogue and every night an automated job will run the transformations necessary to publish the sample data into the public catalogue.

## II. Features or facilities added in 2021

We have updated the Sample Catalogue to MOLGENIS 9.0.6, which allowed us to link the system with the ELIXIR AAI for the authentication and authorisation of users and enable the link with the BBMRI-ERIC Negotiator. We made preparations to reach out the ERN associated biobanks and made contact with the ERICA project to help in coordination of this action. To support the outreach activity, we updated the documentation of the system and we provided training on the research as part of the Pillar 3, WP14 activity and as part of the outreach activities of Pillar 2. We worked with hPSCreg to include samples from the human pluripotent stem cell lines registry in the Catalogue. To support this, we have updated the ontologies in the system.

## III. Plans for improvement in 2022

Together with the ERICA project we will make inventory of the biobanks that are associated with the ERNs and invite them to publish their samples in the catalogue. At the same time, we will work with the biobanks that are already in the catalogue to update their information in the catalogue. We will continue to support the development of the Virtual Platform and make the necessary adjustments to the system to make it operate as part of the Virtual Platform.

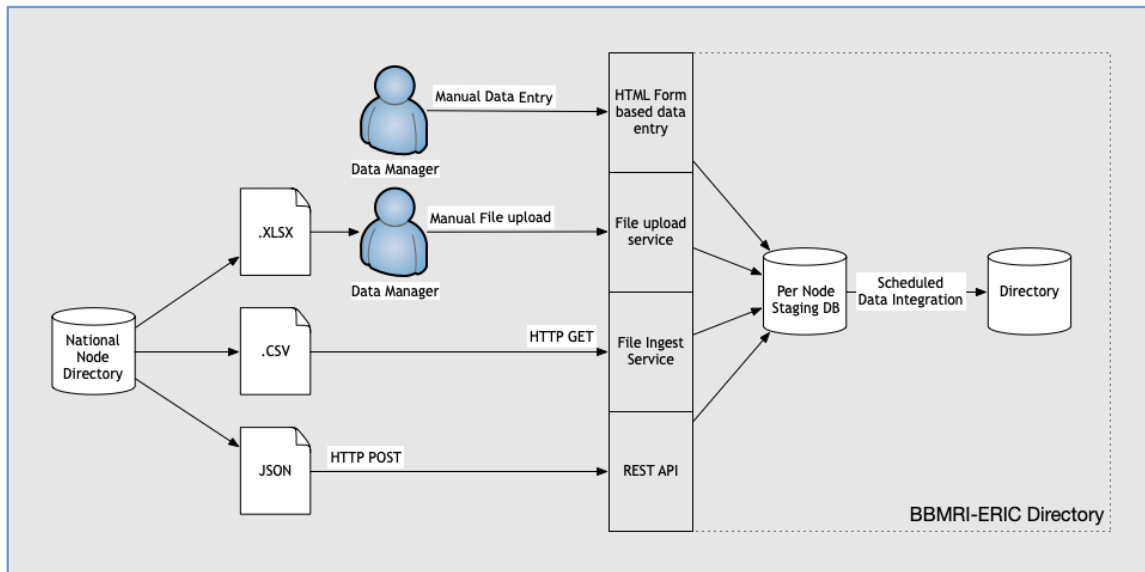
### 2.6 BBMRI-ERIC Directory

Contributors: Esther van Enckevort (UMCG), Heimo Müller (BBMRI-ERIC), Petr Holub (BBMRI-ERIC)

#### I. Resource data flow

The BBMRI-ERIC Directory has a federated process of updating the data, where each National Node is responsible for updating the data for the biobanks in the node. This is done in a staging area that gives the national node exclusive access to update the data. Data in the Directory can be managed in four different ways:

- Manual data entry if the National Node does not host a National Directory
- Manual upload of Excel or CSV files exported from the National Directory
- Scheduled file ingest of CSV files from the National Directory
- Programmatic updates initiated by the National Directory (using the Directory's RESTful API's)



**Figure 8.** BBMRI-ERIC Directory data flow

Regardless of the method used to update the staging area the data from the staging area is integrated into the Directory through a nightly scheduled job. This means that it takes one day before changes are visible to the outside world. In the meantime, the data manager of the National Node can access and verify the data in the National Node's staging area.

Next to the data that is provided by the National Node, the Directory displays quality marks that are based upon the self-assessment filled in by the biobanks. These parameters are managed by BBMRI-ERIC's quality management team and cannot be updated by the National Node. However, for a smooth process of application for the quality marks it is paramount that the biobank and the collections are registered in the Directory before the self-assessment is filled in.

The above description was taken from the BBMRI-ERIC Directory Data Manager Manual, DOI: 10.5281/zenodo.3452137

## II. Features or facilities added in 2021

The Directory was updated to 8.7.2 of the MOLGENIS platform and the Directory application was updated to 4.5.6. We have improved the search capabilities by allowing the user to change the behaviour of a search facet from Satisfy all to Satisfy any and we improved the selection mechanism in the Diagnosis available facet. The mechanism to publish data from the staging areas to the production database was improved to support deletion of stale records and to improve validation of the data in production. The new method also supports hooks for the addition of enrichment steps to make the data in the production database richer based on separately managed quality data and the support for persistent identifiers.

We added support for bioschemas and enabled the FAIR datapoint implementation to improve interoperability of the Directory with the Virtual Platform. We discussed together with the BBMRI-ERIC CS-IT WP4 (Interoperability) the development of a pilot for a data cube implementation that would support counting of samples.

An alpha version of the integration of the Biobank and Registry Finder with the Directory has been developed as well as a pilot version of the implementation of the Query Builder interface (Virtual Platform portal V0).



### III. Plans for improvement in 2022

The BBMRI-ERIC Directory development plan defines several topics to improve RD functionality with the following priorities for the Directory in 2022:

- Continue optimization of user experience for various browse and search scenarios in particular ability to count samples + filter on links to registry.
- Integrate the EJP-RD funded contents and functionalities in the production version of the Directory, in particular the Query Builder interfaces and the Biobank & Registry Finder data.
- Support the RD-Connect Sample Catalogue by providing information on RD Resources.
- Finalize the implementation of Persistent Identifiers Policy to ensure cross references between EJP systems do not break. This is also important to identify and link resources, which are present in different catalogues (BBMRI-ERIC, ERDRI.dor, Orphanet).
- Continue the pilot integration of EJP-RD resources in the BBMRI-ERIC negotiator workflow.

To implement these features at least one major new release of the system, and minor releases where applicable, will be delivered.

## 2.7 Rare Disease Cohorts (RaDiCo)

**Contributors:** Salomé Attia, Sonia Gueguen, Tarik Belgacem, Amine Moussaceb.

### I. RaDiCo introduction

RaDiCo is a national operational platform dedicated to the development, within a research framework, of many rare disease (RD) cohorts that meet strict criteria of excellence.

It is an infrastructure, which has been set up *ex-nihilo*: it pools all the resources needed for implementing within an industrialization framework a common RD database: Constructed on a "cloud computing" principle, it is oriented as an "Infrastructure as a Service"; Interoperable; Including the Exchange format and data security in compliance with the European directive on the General Data Protection Regulation (GDPR); Favouring the use of a secure, open-source, web application; Ensuring a continuous monitoring of data quality and consistency; RaDiCo will also contribute to collect data for French Health Data Hub.

It uses REDCap (Research Electronic Data Capture) which is a free and open-source Research Electronic Data Capture (EDC) suite created by Vanderbilt University (<https://www.project-redcap.org/>). REDCap is an internationally well-recognized EDC tool for research. It provides a large and complete toolset which allows for full management of all the steps within a clinical study, from study design to data analysis, going through to data collection and data monitoring. For more details, see the REDCap website.

It brings an eCRF service and it allows us to enter, host, and consult medical data from patients followed by RD centres from all over France and Europe. Medical data are stored in line with GDPR and informatic rules concerning sensitive data security.

RaDiCo also provides fine access rights management, using the following notions:

**Users:** internal users from RaDiCo (anonymous access to data) and /or users from the medical and healthcare field (full access to data)

**Cohort:** the cohort is dedicated to the study of patients with a defined rare disease. Each user has access to one or more defined study. By this way, the user has a delimited access to medical data.

**Study centre or inclusion group:** The medical group that takes care of a defined patient list. For example, users at Paris Trousseau Hospital only see medical data from patients followed at Paris Trousseau's hospital

**Users' role:** Investigator, Clinical Research Monitor, Data manager etc.

Each user's access to medical data is determined by **their role** in the **cohort** and by **their Inclusion group**.

The Patient Identification System Translator (PIST) was invented by the RaDiCo team. This tool allows allocating specific codes to every patient: a unique RD identifier, the national Rare Disease Identifier called IdMR, is generated by the BNDMR (French Banque Nationale de Données Maladies Rares) algorithm and, upon inclusion, a RaDiCo study code is added. It therefore allows: (1) to generate anonymized codes (2) to manage separately identification data and medical data.

It then allows to a pre-defined, authorised user of the RaDiCo study (i.e., medical staff in charge of the patient) to enter and search its eligible and/or included patients:

- either through the family name, first name, date of birth and/or gender,
- or, alternatively, by using the attributed codes. It also makes it possible to conciliate / clear ambiguities in patient identities (i.e., close identities, duplicates, possible changes of name, spelling mistake resolution, etc.).

## II. Resource data flow

### RaDiCo's resource organization:

In order to respect the segmented rights and accesses according to each role, resources are strictly separated in the system. Thus, resources are organized as following:

- **The Back Office:** a unique place where, for each cohort, RaDiCo organizes the user rights' delegation scheme. This component centralizes the delegation of user rights mirroring the organization and management of each cohort, as well as healthcare pathways/actors.
- **The CGM hosted part,** dedicated to medical and sensitive data and to patient identifying data management. It comprises of the following elements:
- **PIST:** Patient Identity System Translator: Patient identity database
- **EGCS:** The Electronic Data Capture Gateway Controller Service (EGCS) provides a table of correspondence between PIST and REDCap, as well as between BO and REDCap.
- **REDCap** The open-source Research Electronic Data Capture (EDC) which proposes four major services: 1. Form-building; 2. Patient Visit Planning; 3. Data Quality Management; 4. Export of the Capture.

Moreover, medical data entered in each REDCap can refer to several clinical metadata semantic standards:

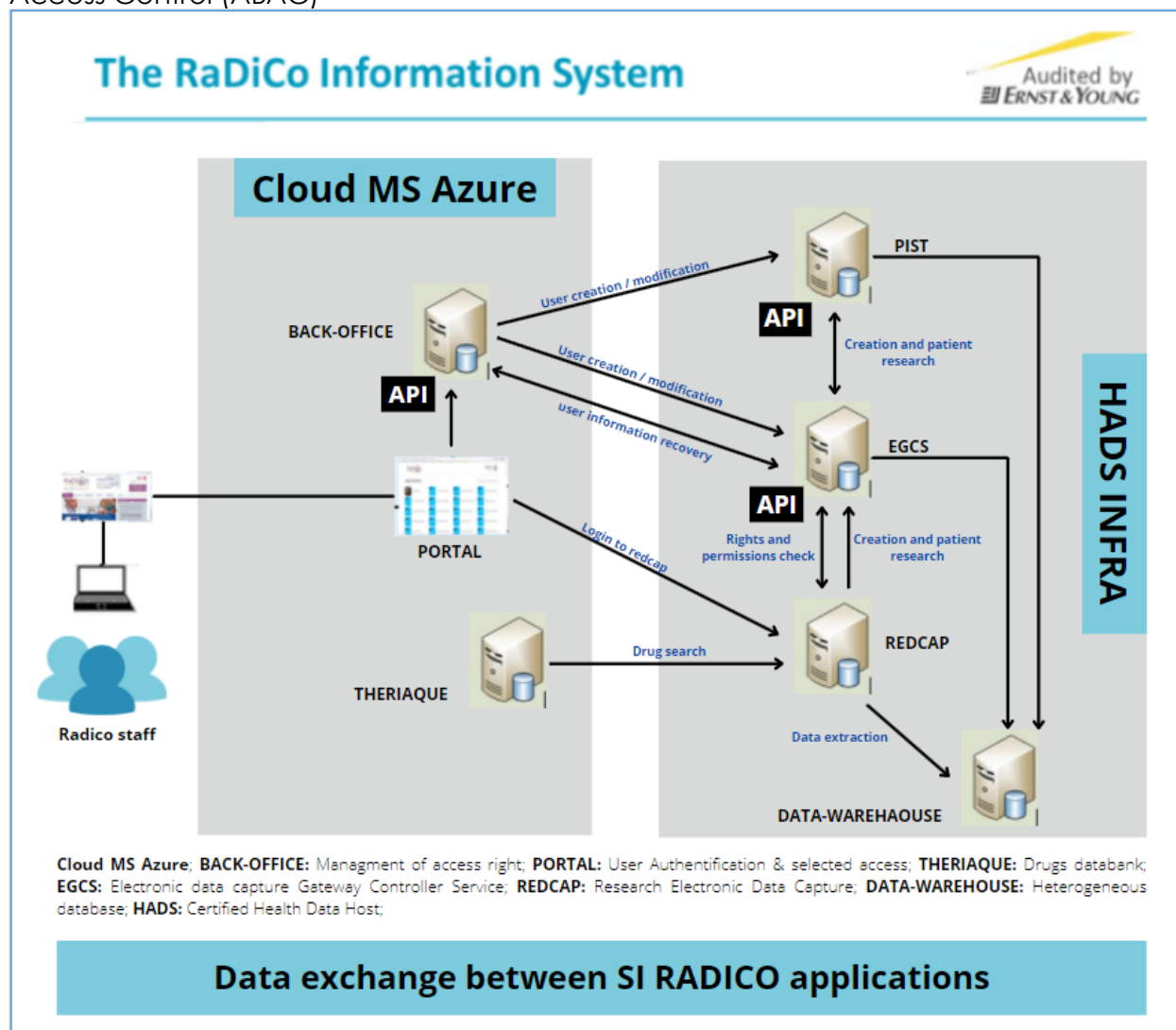
- **Human Phenotype Ontology:** The Human Phenotype Ontology (HPO) provides a standardized vocabulary of phenotypic abnormalities encountered in human disease. Each term in the HPO describes a phenotypic abnormality, such as Atrial septal defect.
- **MedDRA (Medical Dictionary for Regulatory Activities):** MedDRA is a highly specific standardised medical terminology that is used to facilitate the sharing of regulatory information internationally for medical products used by humans.

- **Ontologies from Bioportal:** Especially Orphanet and ORDO. However, the Bioportal gives access to more than 800 available ontologies.
- **Thériaque:** Thériaque is a database of all medicines available in France for health professionals.

### RaDiCo IS user's organization

RaDiCo IS user's organization reproduces the cohort's organization in the RaDiCo Information System (IS) through identified roles. Each user with access to RaDiCo cohorts online has a defined role. Each role has precisely defined rights and access to medical data.

More generally, RaDiCo IS user's organization follows the principles of Attribute Based Access Control (ABAC)



**Figure 9.** The RaDiCo Information System

### Clinical Research users:

- **Medical data entry:** Coordinating Investigators, Principal Investigators, Investigators, Clinical Research Technicians,
- **Medical data verification:** Data manager, Clinical Research Associate Monitor, Clinical Research Project Manager,

- **Medical data analysis:** Statisticians.

IT users:

- Informatic System Administrator
- Software developer

eHealth users:

- eHealth project managers

### III. Improvements made in 2021

- **Relocation of a patient** through the HMI managed by the clinical research project manager. Implementation of a "history of the patient monitoring centre". (Previous investigative centre ==> current centre). **Example 1:** in the case where a patient is taking care of by a given centre then by another centre. **Example 2:** Transfer of the patient from a paediatric centre to an adult centre.
- **Homonyms (duplicates) management:** This concerns the non-duplication of patients at creation. Implementation of a validation and creation process of homonyms in the database with notification (Homonym: name, first name and date of birth identical). The control of duplicates is now done at cohort level, while it was previously done at investigation centre level.
- **Family code management:** Identification of a relationship between patients of the same family by setting up a unique code for all the related.
- **ARM function** (Choice of arms on RedCap): Integration of the ARM function of RedCap from the creation of the patient. From a clinical standpoint, this function makes it possible to choose the ARM (Ex: Placebo) in which the patient is recruited. Technically, it directs towards an eCRF specifically dedicated to the patient follow-up path appropriate to the type of arm previously chosen.

### IV. Improvement planned for 2022

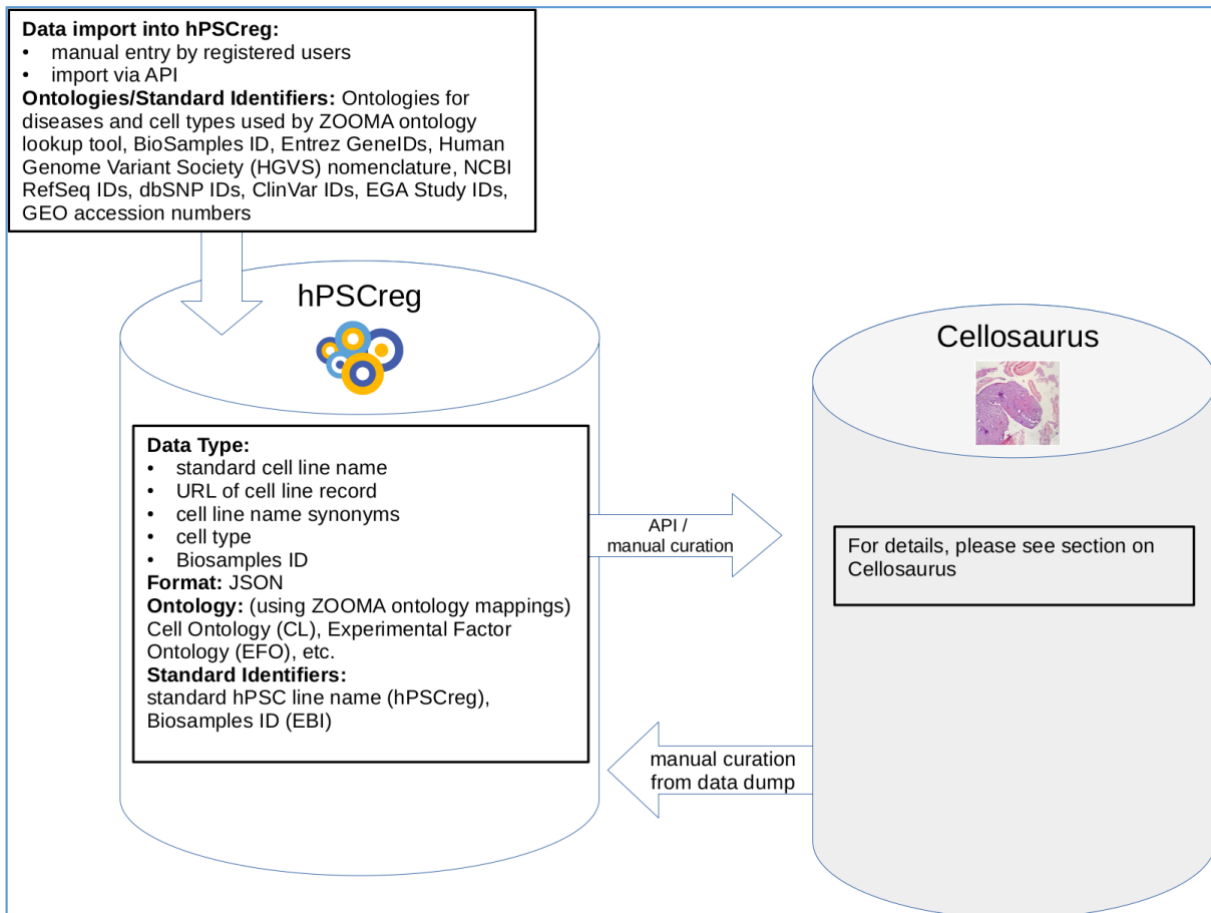
- **Migration France Cohorts:** Inserm has set up a private Cloud under the aegis of France Cohorts. Integration of RaDiCo IS into France Cohorts via a complete migration of the various components of the current RaDiCo IS.
- **Overhaul of the RaDiCo BI** with the France cohort platform.
  - Definition of the Data Warehouse data dictionary allowing an optimal analysis of health data and a better crossing over with other data sources.
  - Creation and enrichment of the DW on *Greenplum* Database
  - Supplying the DW by implementing interfaces with the ETL *Informatica*
  - Installation of controls / alerts on the consistency and quality of data
  - Creation of contractual KPIs with *Qlik Sense*.
- **EDM (Electronic Document Management):** The use of a document management platform established by France Cohorts to exchange documents between cohort members in a safe way.
- **Procedure and documentation:** Update of technical processes and documentations in relation to the various developments of the IS

## 2.8 hPSCreg

Contributors: Nancy Mah (FH-IBMT), Andreas Kurtz (FH-IBMT), Sabine Müller (FH-IBMT)

## I. Resource data flow

Cell line data on human pluripotent stem cell lines is entered by registered users, and subject to wilful submission by the user of the minimum dataset (required by hPSCreg), all data become publicly available on the hPSCreg website. Within other resources in the EJP-RD project, hPSCreg has been actively exchanging data with Cellosaurus via API and manual curation. An overview of the resource data flow is shown in the figure below.



**Figure 10.** Data flow for the human Pluripotent Stem Cell registry (hPSCreg)

## II. Features or facilities added in 2021

Connecting human pluripotent stem cell (hPSC) lines and their data from research applications to downstream uses such as hPSC-derived cell types for research or clinical translation gives comprehensive overview of hPSC research outputs for researchers, clinicians, patients, industry representatives and funders. As of November 2021, hPSCreg holds 3727 hPSC lines, of which approximately 25% of lines originated from donors with rare disease. The clinical study database contains 87 clinical studies involving the use of hPSC cells or their derivatives in interventional therapeutic applications. Of these, 35% were used to treat rare disease (by ICD-10 codes) listed in Orphanet. To complete the translational path between hPSC cells and their differentiated cell types for cell-based therapies, hPSCreg has implemented a prototype data structure to record iPSC-derived cell types, which could be developed further as a more comprehensive data resource for hPSC-derived cell types in the future.

To improve its visibility to the rare disease community, hPSCreg has been interacting with various EJP-RD partners: 1) The Registry has been represented in three EJP-RD online events in 2021. These include the Spring 2021 Biobank Training (May 2021), the Resources and Data Webinar (June 2021), and most recently, the Fall 2021 Biobank Training (Oct 2021), where hPSCreg (FH-IBMT) co-hosted the training with FTELE. 2) In conjunction with UMCG, hPSCreg cell line data are being deposited into the RD-Connect Sample Catalogue, which will greatly increase the findability of rare disease hPSC lines in the RD community. 3) hPSCreg started a Use Case with GPAP, in which genes of interest in GPAP could be linked to cell lines in hPSCreg (via gene modifications in the cell lines). Data exchange mechanisms between GPAP and hPSCreg were initiated with the help of both CNAG-CRG and the FAIRification team in Pillar 2. The hPSCreg cell line data provided as an OWL file to the EJP-RD FAIR team greatly facilitated the integration of hPSCreg data into the GPAP Use Case. 4) As the result of the work done in the GPAP Use Case, hPSCreg cell line data could be integrated into the alpha version of the Virtual Platform.

### III. Plans for improvement in 2022

The Use Case between GPAP and hPSCreg is expected to be finalized, and a new Use Case will be started to additionally link human cell lines to mouse resources in INFRAFRONTIER.

## 2.9 Cellosaurus

Contributors: Amos Bairoch (Cellosaurus)

### I. Resource data flow

Cellosaurus is a manually curated resource concerning cell lines. It provides a wealth of information on more than 131'000 different cell lines. About 25% of the cell lines are relevant to rare diseases (either genetic diseases or rare cancers) and are therefore used by the RD community at large. Providing a maximum of information on these cell lines benefit the RD research efforts.

In-flow of data is from the curation of literature, parsing of data sent by submitters (e.g., individual emails, excel files from companies or cell line collections or other resources), use of API from collaborating resources (e.g., hPSCreg) and scraping of web resources. Output from the Cellosaurus resource is available in 3 formats by FTP: text, OBO and XML and the web site.

The ontologies used in Cellosaurus are numerous and examples include - for disease terms: Orphanet ORDO and NCI thesaurus, for organisms: NCBI taxonomy; chemicals: ChEBI; DrugBank; genes: human: HGNC, mouse: MGI; rat: RGD, Drosophila: FlyBase, vertebrates: VGNC; for proteins: UniProtKB; sequence variations: HGVS nomenclature and NIH ClinVar; STR markers: ANSI/TCC ASN-0002-2011 + additional markers; other in house small "vocabularies": cell line categories, MHC genes, Ig isotypes, genders, etc. In 2021 the Cellosaurus became an ELIXIR Core Data Resource and an IRDiRC recognized resource.

### II. Features or facilities added in 2021

- We have completely restructured the sequence variation information in the Cellosaurus, each sequence variation is now linked to its corresponding gene entry in HGNC, described using HGVS nomenclature and cross-referenced (when a corresponding entry exists) to ClinVar. The zygosity of the mutation is

also captured. The format if both the text and XML versions of the resource have been updated to accommodate these changes.

- Mapping of rare diseases to ORDO was completed (although mapping could not be done to some rare cancers as the corresponding ORDO entries have not yet been created).
- We continued retrofitting the tissue/cell type origin of a cell line.
- Cellosaurus is now discoverable through the Virtual Platform Portal V0.

### III. Plans for improvement in 2022

We will start to map the tissue/cell type origin of cell lines to UBERON+CL. We plan to implement a beta version of an API that will allow users to extract part of all of the information corresponding to one or more cell lines in JSON and XML.

## 2.10 INFRAFRONTIER

Contributors: Sabine Fessele (INFRAFRONTIER), Montserrat Gustems (INFRAFRONTIER), Philipp Gormanns (INFRAFRONTIER), Andrea Furlani (INFRAFRONTIER).

### I. Resource data flow

The main data resource of INFRAFRONTIER is the EMMA (European Mouse Mutant Archive) database. It holds data about more than 8000 mutant mouse strains. There are three routes of data flow into the EMMA database, depending on the origin of the mutant mice. Deposition of data about mouse strains usually runs in parallel with submission, evaluation and import of the mouse material at a national node, where the strain will be frozen down and made available for distribution to other scientists. To add further value to the mouse strains archived in the material repository, both manual and automated processes are in place to standardize, QC and enrich the basic mutant mouse strain data.

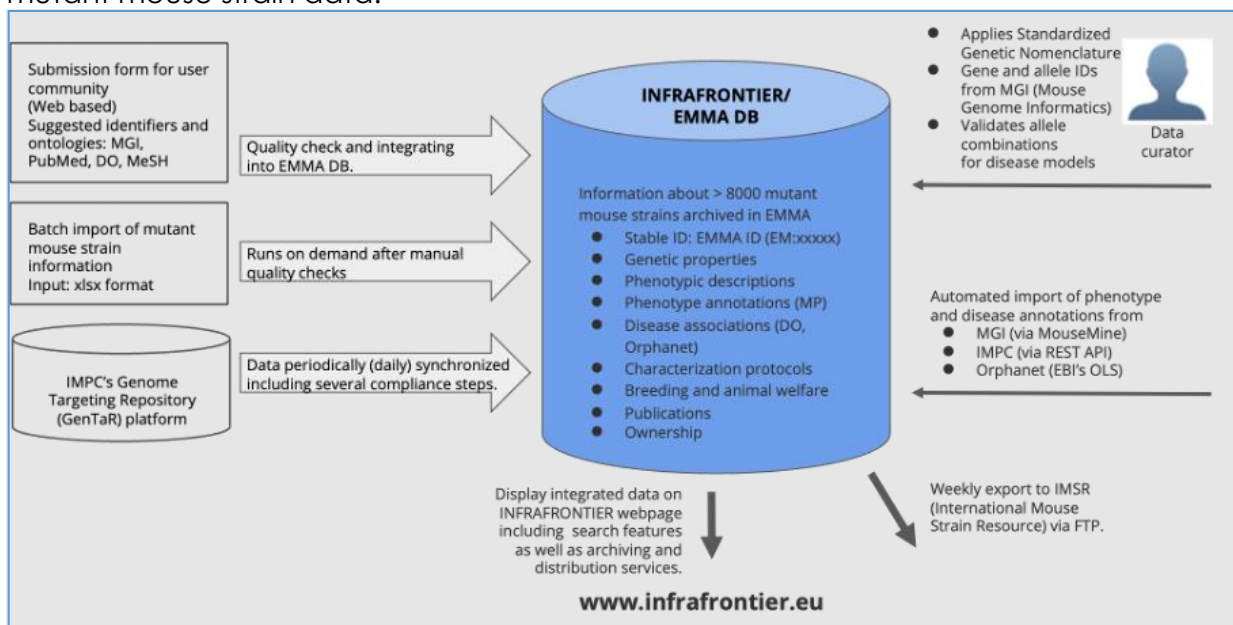
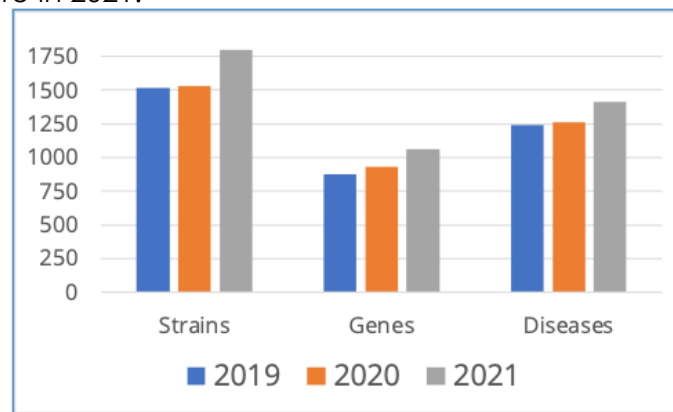


Figure 11. Data flow of EMMA

## II. Features or facilities added in 2021

The list of EMMA strains that are potentially interesting for rare disease researchers on INFRAFRONTIER's rare disease landing page set up in 2019 has further increased. Currently the EMMA repository holds 1799 mouse strains that carry mutations in 1065 genes that have been implicated to play a role in rare diseases (1413 different rare diseases). Growth of these numbers has been so strong, since we have been able to conclude license agreements and implemented processes that now allow us to distribute mouse models generated by use of the CRISPR/Cas9-technology via the EMMA platform, which has not been possible before. Also, this year, the International Mouse Phenotyping Consortium (IMPC), for which INFRAFRONTIER/EMMA takes the part of distributing the mouse strains generated by the European partners to the biomedical research community, has switched to its new GenTaR (Genome Targeting Repository) platform. The latter records information about the mutation created and tracks the state of phenotyping performed on the mutant mouse line. GenTaR has been created to capture more information about CRISPR alleles and INFRAFRONTIER has fully adjusted its importer suite to take all relevant information from GenTaR. In line with this, the INFRAFRONTIER curators have put a special focus on reviewing CRISPR allele nomenclature in 2021.



**Figure 12.** Development of the number of strains that carry mutations in genes that are mapped to an Orphanet rare disease via the genes mutated in these strains. While the initial data integration process was established in 2019, the focus in 2021 was on adding data about mouse models generated by use of the CRISPR/Cas9 technology

In addition to increasing the number and diversity of mutant mouse strains that will be made searchable via the EJP-RD Virtual Platform later on, we contributed to the communication about the EJP RD Resource Finder via:

- Twitter: <https://twitter.com/InfrafrontierEU/status/1403224101221801986>
- A news item on our webpage: <https://www.infrafrontier.eu/news/ejp-rd-resource-finder-“one-stop-shop”-rare-disease-researchers>
- Included the Resource finder in our dedicated rare disease subpage (in the last expandable section): <https://www.infrafrontier.eu/infrafrontier-and-rare-diseases>

## III. Plans for improvement in 2022

From INFRAFRONTIER's side, the activities already started in the previous years will be continued in year 4. The focus will be on developing the API for the integration of INFRAFRONTIER in the EJP-RD virtual platform.

The collaboration with RD-Connect GPAP, which aims at improving the analysis capabilities in rare diseases and that had to be postponed due to the lack of



personnel resources, will be resumed in year 4. This collaboration will allow users from RD-Connect GPAP with a candidate gene to query INFRAFRONTIER for our database of mouse models; this query will return the available mouse models with alterations in the corresponding mouse gene of interest, which can help initiate new lines of research.

In the course of regular virtual work focus meetings, it has also turned out that a similar collaboration between hPSCreg and INFRAFRONTIER could provide an added value for RD researchers and, if resources on both sides allow, we would also kick this off in 2022.

In both collaborations we will try to expand queries using ontologies like HPOs or ORDOs, which would allow for genome-phenome comparisons between both species. This could provide valuable information to, for instance, define human candidate genes from a mouse model phenotype. Finally, we would like to study the possibility to expand these queries from RD-Connect GPAP and hPSCreg to mutant mouse strains stored at INFRAFRONTIER.

Furthermore, a collaboration with the University of Leicester on PaR-RaDiGM is still considered. PaR-RaDiGM has information about rare disease researchers (some of them already using model organisms) and INFRAFRONTIER could provide (upon user agreement) information of model organism researchers (not necessarily involved in rare diseases before). Researchers working on the same gene/allele could be the starting point to bring rare disease researchers and model organism researchers more closely together, which is both interesting for the researchers and the patients.

## 2.11 MetaboLights

Contributors: Keeva Cochrane (EMBL-EBI), Claire O'Donovan (EMBL-EBI)

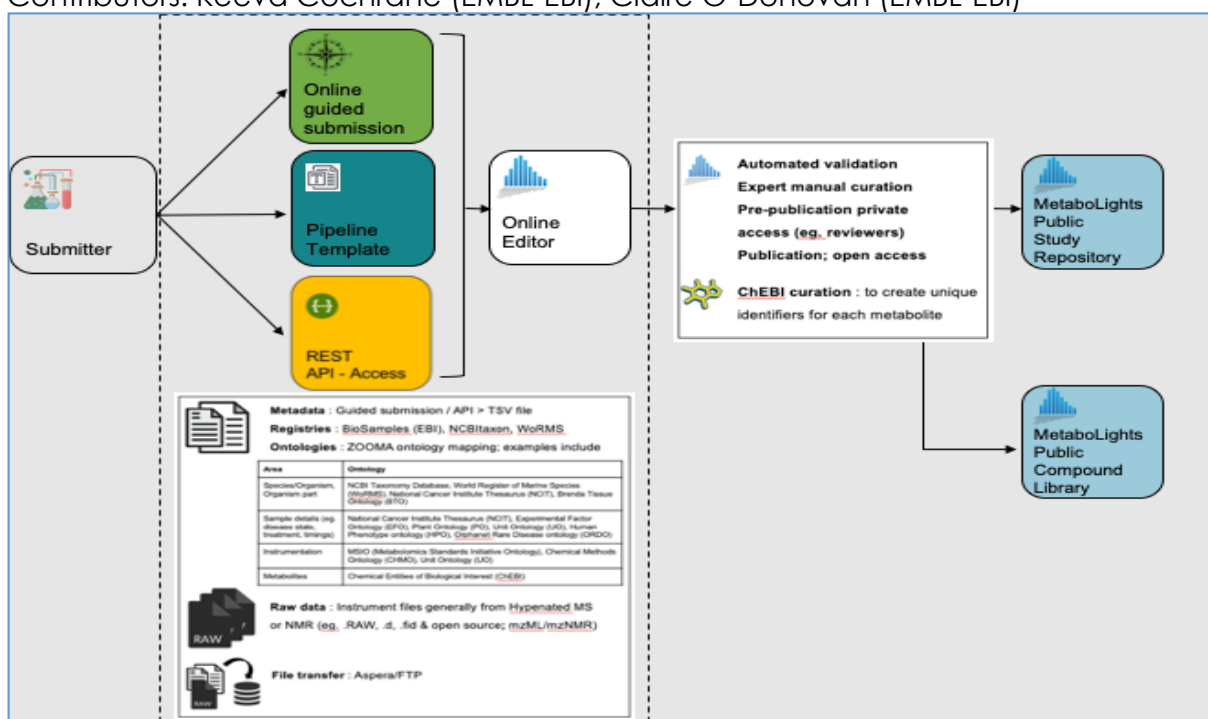


Figure 13. Data Flow for MetaboLights

## I. Resource data flow

MetaboLights is a data repository for metabolomics data. Each new study is assigned a unique and persistent identifier. Submitters can choose to use the online guided submission, a pre-populated template or API to deposit a study. The primary requirement for a MetaboLights study is the raw data (or open source converted format of raw) for which users have the option of Aspera or FTP transfer methods. In each case submitters are asked to provide the relevant metadata as instructed including sample information, experimental protocols, and a derived table of metabolite identifications, all of which is under pinned with ontology references. Metabolites identified in studies are curated into the ChEBI ontology if a record does not exist. Each study is automatically validated with a series of checks and once passed, submitters can change the study status to request curation. Following successful curation, a study is held in private mode and a link is available to share with e.g., journal reviewers until the requested publication date is reached and the study is made publicly available. MetaboLights also supports a compound library which essentially provides a synopsis of the chemical features (based on ChEBI ontology integration) together with biological references including all study identifiers & associated relevant metadata (e.g., species, disease) per metabolite identified within the repository.

## II. Features or Facilities added in 2021

MetaboLights has continued to grow exponentially and now contains 2749 studies; more than half of which are from Homo sapiens and Mus musculus reflecting the use of metabolomics in clinical research. This has enabled us to continue our work in standards and ontology development and implementation for the benefit of EJP RD. In addition, Metabolights presented a webinar in the series "Resources Webinars" organized by EJP-RD, which allowed the platform to provide an overview and explain its functionalities to the rare disease community.

## III. Plans for improvement in 2022

The development of a controlled access version of MetaboLights had to be delayed due to circumstances related to the pandemic. However, we plan on beginning again during 2022. In addition, future EJP-RD use cases will be explored for efficient processing of external datasets.

### 3. Conclusion

During 2021 the resources involved in Subtask 11.3.1 have continued on improving their data deposition and access interfaces and tools for rare disease use-cases.

Multiple improvements have been done to facilitate user access and improve platform usability in the different resources. For instance, GPAP, by rewriting their Data Management module, which is now aligned with other GPAP modules, or by providing a new improved experiment view. EGA, by improving the Programmatic Submission, or developing the Star2xml tool. DECIPHER, by releasing a new version of its platform, which has improved visualization and access of the data. RaDiCo by releasing Management improvements involving patient registry relocations or duplications. Cellosaurus releasing new mappings of rare diseases to the ORDO ontologies. And other important developments by other resources, all with the aim to facilitate and improve the user interaction with the resources to improve rare disease studies.

In addition, guidelines and documentation are continuously being updated for current processes, or released for new implementations, like those created by GPAP, EGA or the RD-Connect Sample Catalogue, among others. Moreover, during 2021 the series of "EJP-RD Webinars" has continued, with one resource presenting each month their platform to the Rare Disease community, helping in the dissemination of the available platforms and new tools released.

There are important developments that are being adopted by the resources to increase data accessibility, such as the GA4GH Authentication and Authorization infrastructure, which has been implemented by platforms like EGA, GPAP, RD-Connect Sample Catalogue or the BBMRI-ERIC Directory, and will allow resources to adopt the Lifescience AAI in the future.

One of the main objectives of Subtask 11.3.1 is to increase interoperability between the different resources. Steps are being done to link datasets between GPAP and EGA, and there is currently a use-case to connect data information between GPAP with hPSCreg. This use-case will be used by other resources as a base to create new connections, for instance between GPAP and INFRAFRONTIER. Other examples are the support of the BBMRI-ERIC Directory to the RD-Connect Sample Catalogue by providing information on RD Resources, or the work of hPSCreg to add their cell line data into the RD-Connect Sample Catalogue. These new connections will drastically improve the user capabilities when performing rare disease studies, as it will ease and facilitate the gathering of relevant information for their research.

Finally, work is being done to make resources compatible with the future Virtual Platform. For instance, BBMRI-ERIC Directory has added support for Bioschemas and enabled a FAIR datapoint. In addition, work done in the GPAP-hPSCreg Use Case can facilitate the hPSCreg as a resource to be integrated into the alpha version of the Virtual Platform, as well as Cellosaurus. In 2022 this Subtask will work closely with the different resources to follow the necessary steps to integrate them into the Virtual Platform network.