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

D11.14
Fourth 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, Care and Trial Site Registry (CTSR)

Due date of deliverable: Month 48
Dissemination level: Public

Table of Contents

1. Introduction

This deliverable summarizes the additional facilities and features that were integrated during 2022 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 Virtual Platform, to suitable public repositories and resources.

Subtask 11.3.1 supports European and international resources and infrastructures that are highly relevant to the RD community, by improving and expanding their deposition capabilities and access mechanisms based on RD use-cases identified through community surveys in EJP-RD. Work done in this subtask include:

1. Develop and expand new user-friendly interfaces:
These interfaces allow to deposit data and metadata using the latest standards or ontologies (i.e. HPO, ORDO, OMIM), always based on FAIR (Findable, Accessible, Interoperable and Reusable) principles.

2. Assure data quality:
Quality of data is crucial to obtain reliable results and take informed decisions in clinical and scientific research. In EJP-RD, quality mechanisms are being introduced by implementing manual curation, automatic metrics generation or a mixture of both. In addition, systems like APIs (Application programming interfaces) and GUIs (Graphical user interfaces) are being developed to allow for better query functionality and data access.

3. Assure data security:
EJP-RD resources work with sensitive (potentially identifiable) human data, and therefore it is mandatory to follow all the legal requirements (i.e. EU GDPR or national legislations) and recommendations (i.e. the Global Alliance for Genomics and Health, GA4GH) to assure data security. One example is the use of the Data Security Toolkit, which provides a principled and practical framework for responsible sharing of genomic and health-related data.

4. Connect resources to the EJP-RD Virtual Platform Network:
EJP-RD, in an effort involving Overall Architecture, FAIRification, Query Builder or Metadata Work Foci, has developed the Virtual Platform Network (https://vp.ejprarediseases.org/discovery/), an ecosystem of data resources and auxiliary services to find, access, and use data for various purposes. Subtask 11.3.1 supports the EJP-RD resources in successfully connecting to this network in order to expand discoverability and accessibility of their data to the rare disease scientific community.

The objective of this Deliverable is to identify each of the current capabilities of the named resources and additional capabilities added during 2022 in terms of data deposition and access to data, and to identify areas for improvement to be implemented in 2023, which 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: Carles Garcia (EMBL-EBI), Mallory Freeberg (EMBL-EBI) Jordi Rambla (CRG), Thomas Keane (EMBL-EBI)

I. Resource data flow

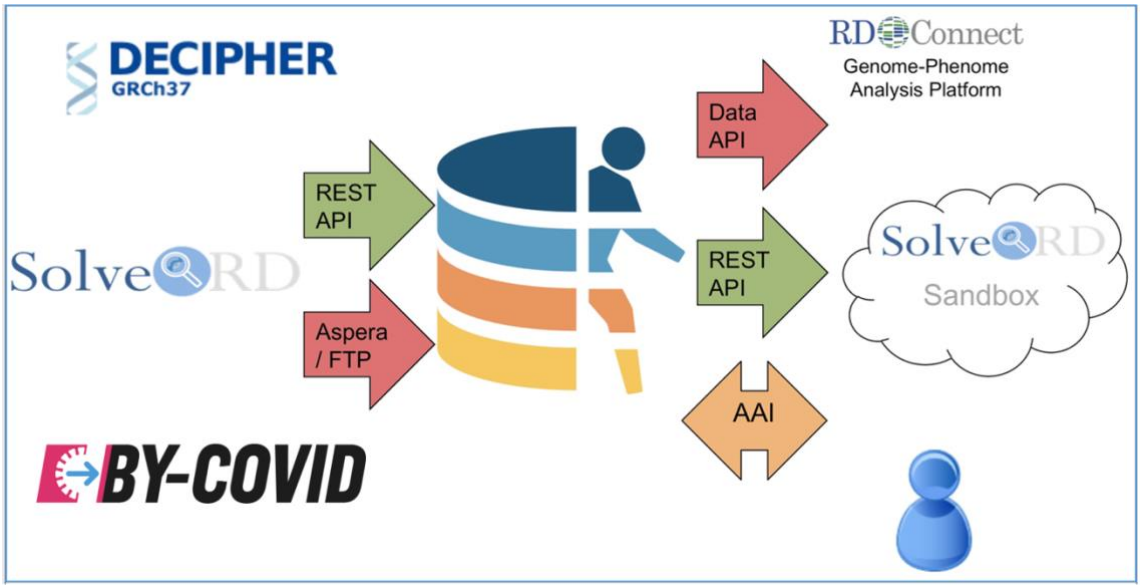

Figure 1. *Example data flows to and from EGA. Submitters, such as Solve-RD or DECIPHER, submit data to the EGA for archive and distribution. These data are distributed via the EGA Data API to authorised users. Use-cases include distributing data to the RD-Connect Genome-Phenome Analysis Platform, the 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.

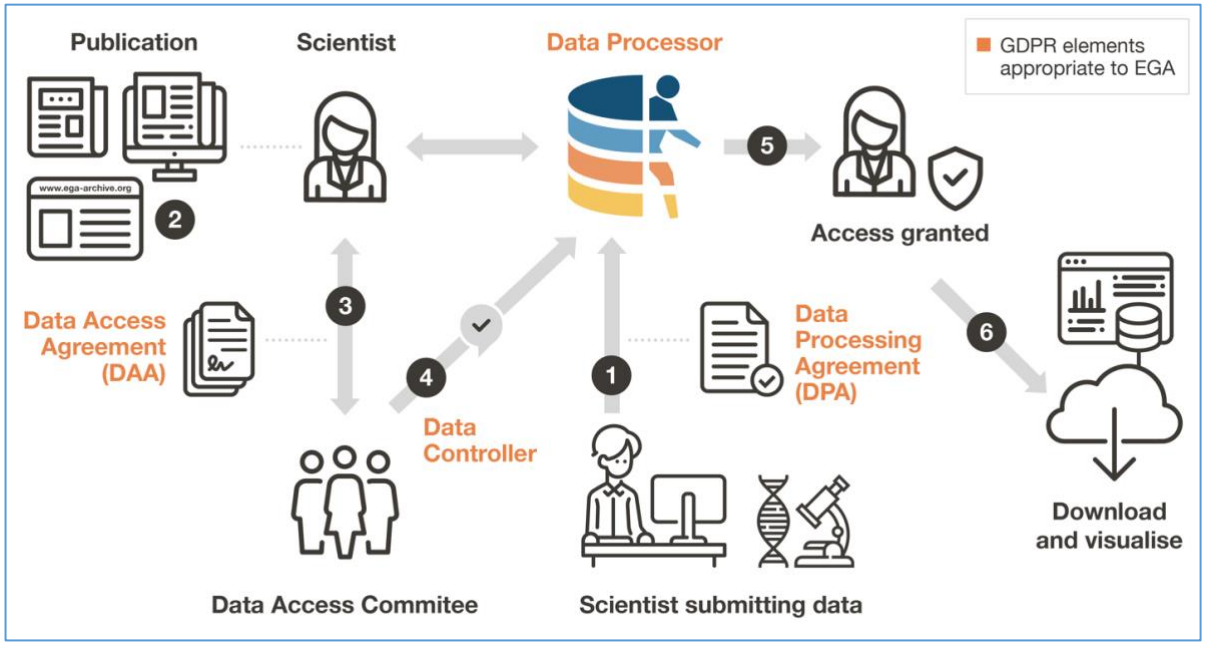

Figure 2. *EGA facilitates the submission, discovery, access and distribution of sensitive human data. A researcher submits controlled access human genetic, phenotypic and clinical data to EGA after signing a Data Processing Agreement (1). EGA processes, archives and releases the dataset to be findable. Another researcher discovers data of interest at 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 visualize the data (6). GDPR: General Data Protection Regulation.

II. Features or facilities added in 2022

EGA continues to work on implementing new functionalities to improve Rare Disease users' submission and access to the data and make progress towards Federated EGA.

It is important to ease the access of clinicians and researchers to the different life science infrastructures. For this purpose, the ELIXIR Authentication and Authorization Infrastructure (AAI) was created, which recently (April 2022) transitioned to the new Life Science Login service (https://elixir-europe.org/AAI-migration), which makes access to life science resources and services easier to users by using a single login, and allows service providers to control and manage the access rights and create different access levels to the data. To maintain compatibility and interoperability with this expanded service, EGA migrated from ELIXIR AAI to the Life Science Login, which provides users a service to map their EGA identity to their Life Science Login identity (https://ega.ebi.ac.uk:8443/ega-openid-connect-server/ega-login), and will allow them to federate their identity and expand access across the EGA and other rare disease life science research services and platforms like RD-Connect Genome Phenome Analysis Platform (GPAP) or the RD-Connect Sample Catalogue, which are also compatible with Life Science AAI.

In addition, during 2022 EGA has used the RD-Connect GPAP Synthetic Dataset (https://ega-archive.org/studies/EGAS00001005702) to successfully demonstrate the ingestion of Phenopackets with domain-specific ontologies (18 phenopackets from 18 participants, representing 5 different rare diseases were ingested), which will allow rare disease researchers to establish workflows that interact with phenopackets archived at the EGA without the need to apply for access to real data.

In December 2022, the "Onboarding and connection of resources to the EJP-RD Virtual Platform" hackathon took place in Barcelona, Spain. The EJP-RD Virtual Platform (VP) will provide a network of data resources via which researchers (and other stakeholders) can automatically find, access, and use data for various purposes, ranging from discovering data and samples to analysing data. There will be 4 levels of connection to the VP, and the aim of the Hackathon was to accelerate the Level 1 connection of resources, which allows for the resource to be discoverable through the network. In the Hackathon EGA implemented a proof-of-concept connection to the VP with a local installation of a "Beacon-in-a-box" solution using synthetic data. As a next step (Q1 2023) EGA will work on completing the Level 1 connection, moving from the proof-of-concept local installation to connecting to the EJP-RD VP, which will make EGA discoverable through the EJP-RD network, improving the visibility of EGA to the rare disease scientific community.

III. Plans for improvement in 2023

EGA will release the next version of the EGA Submitter Portal with enhanced features and capabilities to maintain the quality of the submitted data. One example is the possibility to add ORDO (Orphanet Rare Disease Ontology) terms while supplying phenotypic metadata of a rare disease sample, which will enrich the quality of the rare disease data submitted to the platform. In addition, documentation describing these new submission tools will be created and made accessible to the research community.

EGA also plans to release a new Portal for Data Access Committees (DACs) to ease management of data in the platform. This will be beneficial for the rare disease community as it will allow, for instance, a rare disease consortium to easily manage (accept or deny) access requests from the datasets they have submitted to the EGA, as the consortium DAC will be able to login to the Portal and manage multiple requests through a new user-friendly interface.

Moreover, in 2023 EGA will publish new documentation for programmatic submissions including features helpful for rare disease studies such as submission of Phenopackets and clinical research data.

Finally, during 2023 EGA will continue to work with the Query Builder and Metadata teams to evaluate the implementation of Level 2 connection to the Virtual Platform Network.

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.

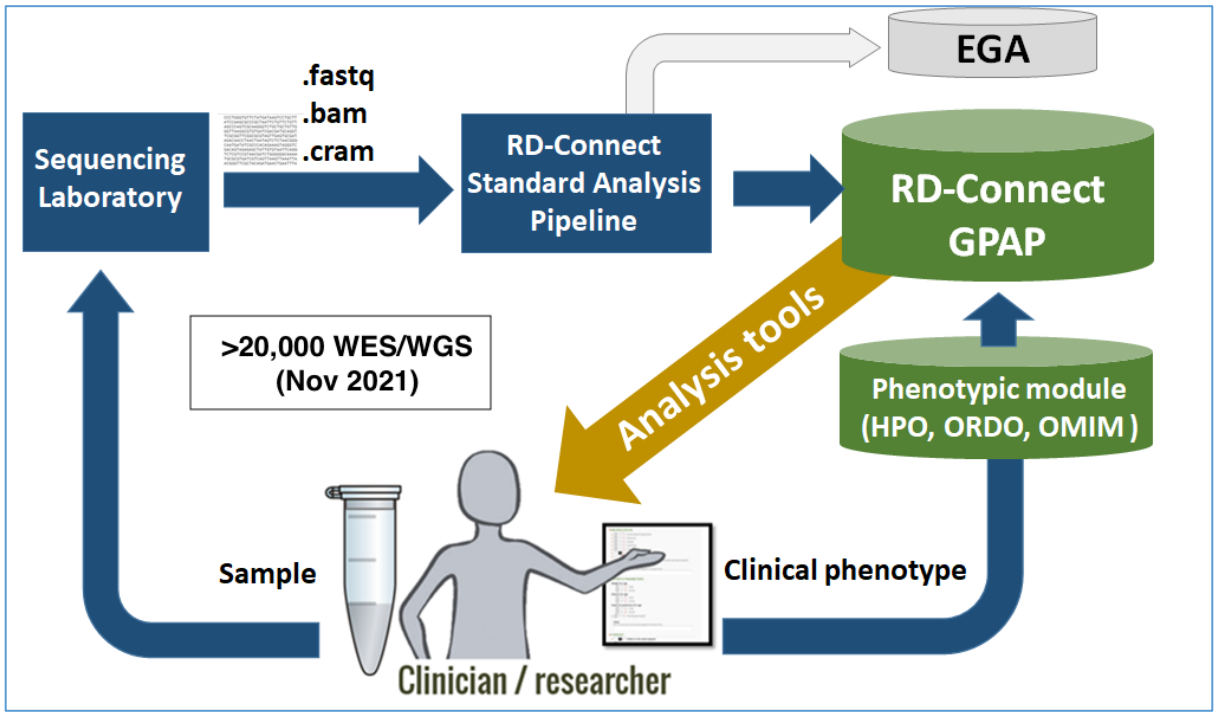

Figure 2. Data flow for RD-Connect GPAP

II. Features or functionalities added in 2022

The RD-Connect GPAP is ready to connect to the Lifescience AAI in the technical terms. We will continue further discussing how the process should look like from the user's perspective regarding authorisation of the user, given that the RD-Connect GPAP is a closed resource.

Improvements in the RD-Connect GPAP Data Management module regarding better guidance of the submission process and reducing time of submission:

- new features have been added: users are able to see the unfinished experiments where the metadata was already submitted but the corresponding sample file was not updated, and they are able to select the experiments and get a pre-filled table where they would like to submit file metadata for
- new dialog boxes were implemented in the RD-Connect GPAP when the user submits data
- A dashboard is added to the Data Management module with detailed information on user data (for example: number of submissions, files submitted)
- We have created a figure to explain the workflow of the data submission in the RD-Connect GPAP for the users to better understand the process.

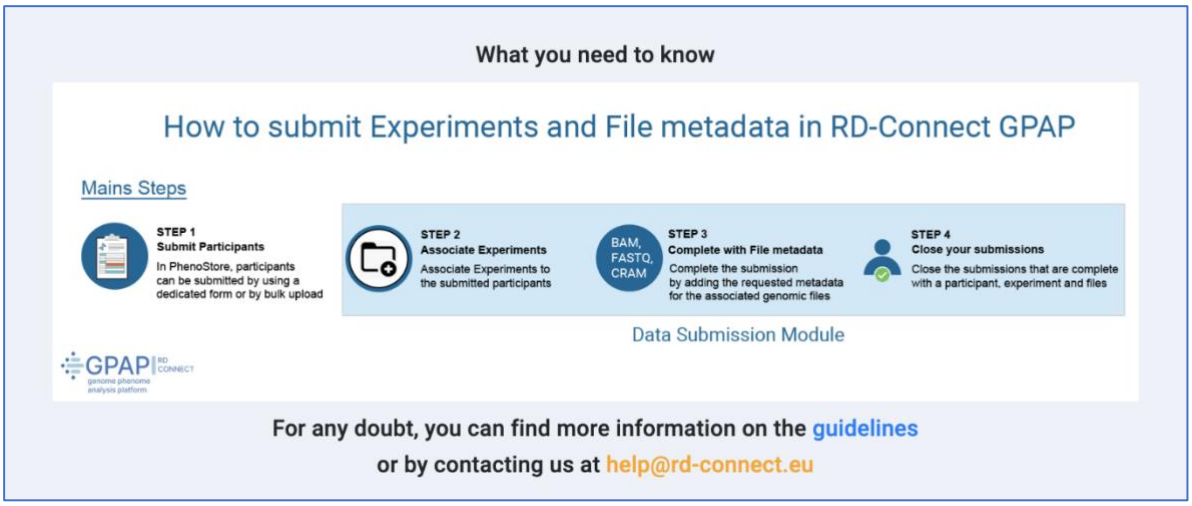

Figure 4: explanation of the workflow of data submission in the RD-Connect GPAP

New Guidelines have been created for the RD-Connect GPAP Cohort App module and they are available within the GPAP documentation on the following link: https://platform.rd-connect.eu/gpap_doc/.
We have created also new guidelines for the RD-Connect GPAP Genomics module: documents are available now to help the user better understand the functions of the Genomics module in the following topics:

- Use cases,
- Filtering,
- Variant dashboard,
- Variant row,
- Labels and tag variant,
- From one analysis/query to the other
- External links and external databases: save, load and share a study

CNAG-CRG has implemented the Beacon 2 standard (v2) and managed to connect to the Virtual Platform during the "Hackathon: Onboarding and connection of the resources to the EJP-RD Virtual Platform" event hosted by CNAG-CRG that was held in Barcelona on the 13-14th December 2022. At the moment the RD-Connect GPAP is connected with the Playground instance to the Virtual Platform.

We are working with hPSCreg to connect the two platforms through an API so that RD-Connect 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 have already implemented a link on the gene level but it currently returns too many genes to put the system into production. There are also plans to link at the variant level leveraging the Beacon v2 API.

The use case is being very similar to connecting RD-Connect GPAP and INFRAFRONTIER. Now that the RD-Connect GPAP is connected on the Level 1 and

Level 2 to the Virtual Platform, we will revisit the case and continue the discussion during 2023.

The discussion on the interoperability in between different pseudonymization solutions has been continued this year to find out which PPRL solution (EUPID, SPIDER) could be used. The EUPID Services were foreseen as the only PPRL solution to be applied within the EJP RD VP, in 2022, the EC's JRC introduced a new service supporting PPRL, i.e., SPIDER. However, SPIDER is currently only applicable to RD registries, while many other related resources (clinical trials, biobanks, etc.) are not supported so far. We have been working on setting up a pilot, and to further define which RD registry and/or biobank should be involved in the use case, that is already implemented within the PPRL Work Focus.

III. Plans for improvement in 2023

The RD-Connect GPAP will launch its connection to the EJP-RD VP using real data instead of synthetic data from its Playground.

RD-Connect GPAP will work together with other resources that would like to connect to the Virtual Platform through the Beacon 2 API.

We will evaluate which are the changes necessary to allow submissions of long read sequencing data such as that generated by Oxford Nanopore sequencers.

We will evaluate the inclusion of additional variant types.

We will continue the discussion of the use case connecting RD-Connect GPAP to INFRAFRONTIER and hPSCreg based on the experiences gained during the hackathon in Barcelona on 13-14th December 2022.

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 these tools. 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).

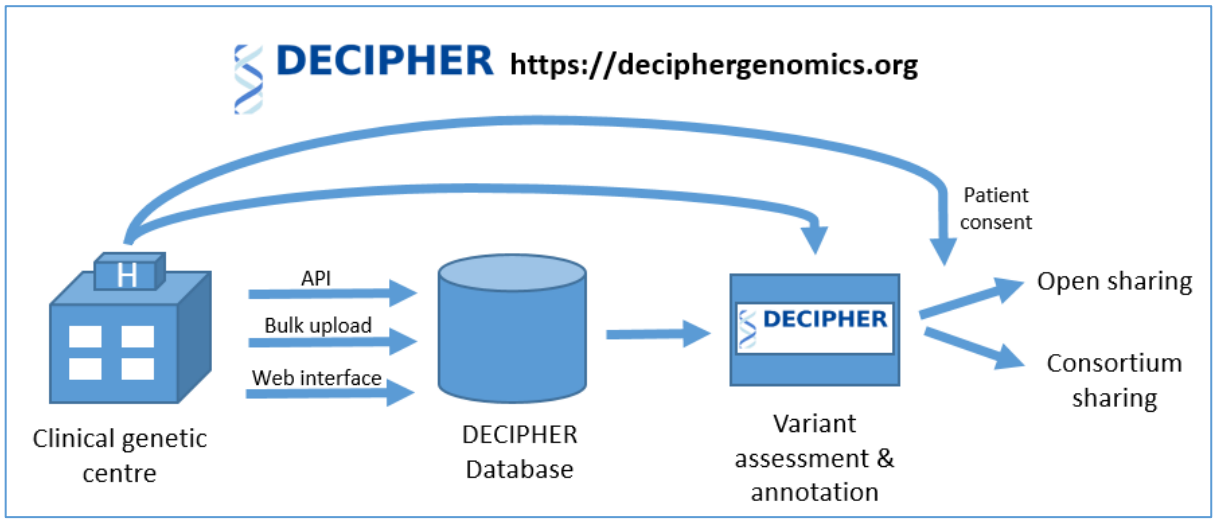

***Figure 3**. DECIPHER Data Flow*

II. Features or facilities added in 2022

The "de novo" inheritance option has now been replaced with "de novo (parentage confirmed)" and "de novo (unconfirmed parentage)". This allows depositors to record confirmed parental relationships versus assumed parental relationships status for de novo variants, which is important when determining the strength given to case level data during a variant pathogenicity assessment, specifically for ACMG/AMP sequence variant criteria PS2 and PM6.

III. Plans for improvement in 2023

We will continue to improve our data deposition interfaces and create clear documentation to ensure the high quality of data submissions we receive is maintained.

2.4 RD-Connect Registry and Biobank Finder

Contributors: Heimo Müller (BBMRI), Vittorio Meloni (BBMRI-CRS4), Alessandro Sulis (BBMRI-CRS4)

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).

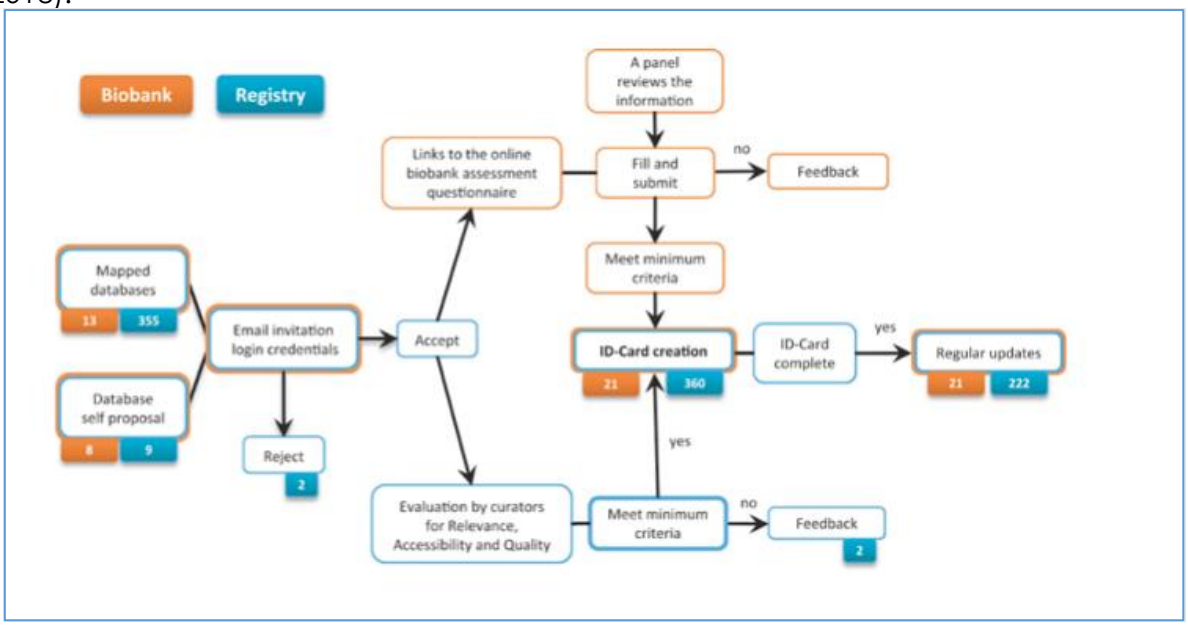

Figure 4. Inclusion of biobanks and registries in the Registry and Biobank Finder. The process of inclusion and evaluation of biobanks and registries in the Registry and Biobank Finder (mapped and self-proposed)

Gainotti, S., Torreri, P., Wang, C. M., Reihs, 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 2022

The transition of the RD-Connect Registry and Biobank Finder to the new MOLGENIS version (ID Card) was finished and the same graphical user interface as the Liferay version was implemented including the Disease Matrix. Data from the legacy system was imported to the new system. As a next step extension of the BBMRI-ERIC directory were developed as a POC, providing the functionality of the Disease Matrix by sub collections and a connection to the EJP-RD VP platform was implemented with the resource level query API. The interface was extended to show both internal search results as also biobanks found through the VP search (external results). In the list of research results items can be selected and send to the BBMRI-ERIC negotiator for further processing.

III. Plans for improvement in 2023

The RD-Connect Registry and Biobank Finder version following the BBMRI-ERIC directory approach will be further developed and the following activities will be carried on:

- move the technical solutions from the test instances to the production system (the BBRMI-ERIC directory with a special collection for RD biobanks)

- transfer biobank information from the RD-Connect Biobank and Registry Finder to the production system, this will be especially done for the EUROBIOANK network together with Telethon and onboard new biobanks.
- curate the biobank metadata and onboard the collection representatives to the BBMRI Negotiator (Life-Science AAI)
- connect the resource level (biobank, registry) information to the VP by exposing the FDP of the Finder. In particular, the FDP will expose:
    - metadata of the Catalogue (the Finder) including description of the QB API endpoint to query the Finder
    - metadata of the Biobanks
- implement the Beacon-like version of the QB API to query the Finder

2.5 RD-Connect Sample Catalogue

Contributors: Esther van Enckevort (UMCG)

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.

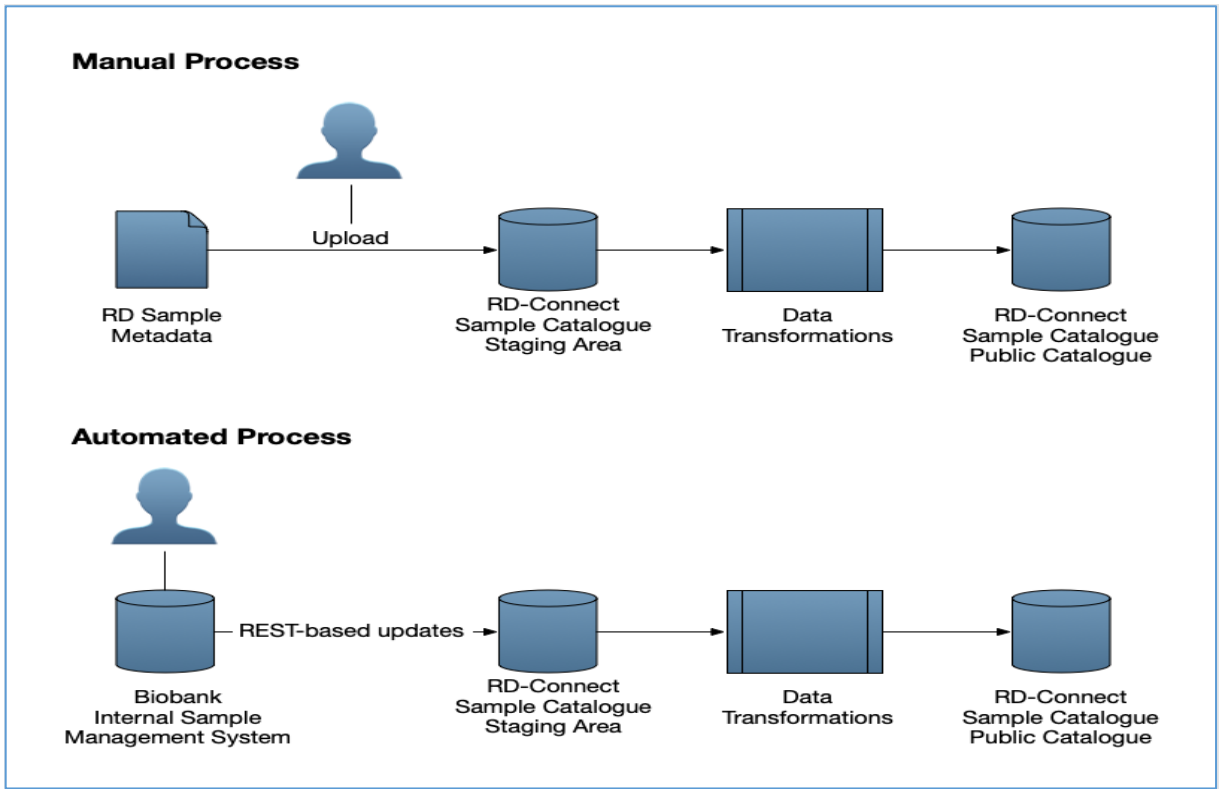

Figure 5. RD Connect Sample Catalogue Data Flow

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 2022

In 2022 we have updated the RD-Connect Sample catalogue to a new version of the MOLGENIS platform, which enabled users to login with their institutional accounts using ELIXIR / LS AAI. This reduces the burden of maintenance on the UMCG team.

We also reviewed and updated the documentation for the biobanks and streamlined the submission workflow.

We worked with the hPSCreg team to include cell lines from their catalogue in the RD-Connect Sample Catalogue.

We started preparations for upgrading the RD-Connect sample catalogue to MOLGENIS EMX2, which will provide us services to integrate the RD-Connect Sample Catalogue with the Virtual Platform.

III. Plans for improvement in 2023

We will update the RD-Connect Sample Catalogue to MOLGENIS EMX2 and integrate with the Virtual Platform.

We will work together with the ERNs to identify relevant biobanks to include in the Sample Catalogue and we will work with the biobanks that have already published their data in the Sample Catalogue to keep the data up to date.

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)

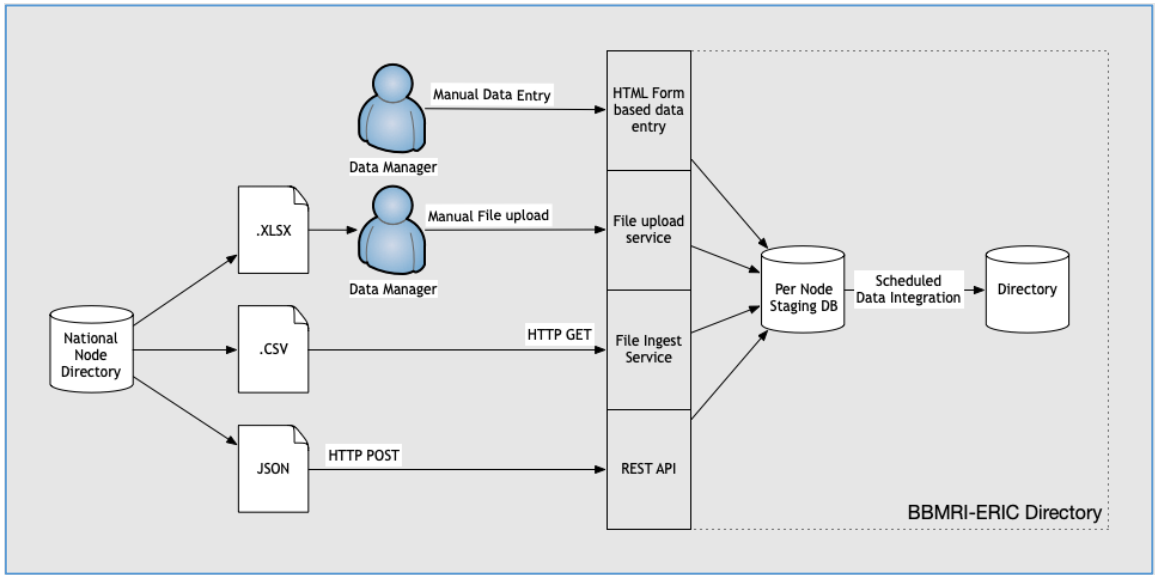

Figure 6. BBMRI-ERIC Directory data flow

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 2022

In 2022 we implemented significant improvements in the user interface of the Directory. We simplified the search filters and made them more responsive. Of particular interest is that we merged the network filters and the quality filters so that a user does not have to choose between searching at the collection or the biobank level, which was very confusing for the user.

We simplified the data model to make it easier for the biobanks to enter data and to make it more consistent. We improved the staging areas and the publication of data to production. We added persistent identifiers for the biobanks and implemented several data enrichment steps in the publication script that make it easier for users to search the Directory. Of particular interest is the work that we did to map ICD-10 codes to Orphanet codes and vice versa, which makes the Directory more useful for rare disease biobanks.

We worked with the MIABIS team to develop a data warehouse model for the Directory that will allow users to search at a more detailed level.

III. Plans for improvement in 2023

We will update the Directory to the MOLGENIS EMX2 platform which provides us with the ability to integrate into the Virtual Platform.

We will continue to make improvements to the user interface to make it easier to find biobanks and to provide the available information in a more consistent manner.
We will pilot the data warehouse model with several biobanks to validate the use of this model for the Directory, from the viewpoint of the biobanks as well as that of a researcher searching for samples. Functionality to implement aggregated descriptions of collections (e.g. the disease matrix) using the Starmodel approach and the CCE-DUC attributes will be implemented (technical feasibility tests and POC) and evaluated for RD biobanks.

2.7 Rare Disease Cohorts (RaDiCo)

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

I. Resource data flow

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.).

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)

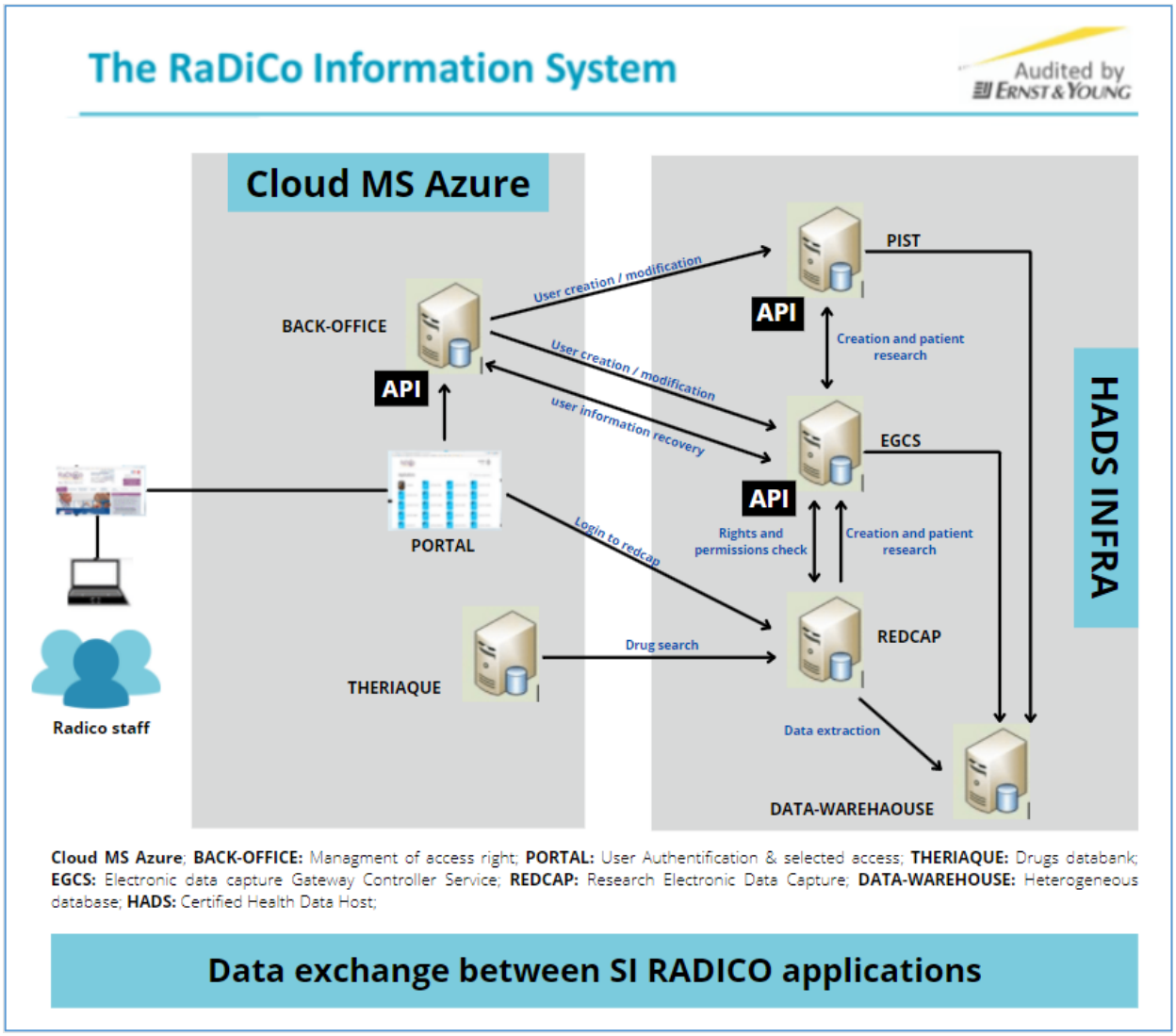

Figure 7. The RaDiCo Information System

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

II. Improvements made in 2022

RaDiCo is preparing the migration of its information system to the France Cohorts Information System and will report on its improvements in the next deliverable.

III. Improvement planned for 2023

The improvements related to the migration to France Cohort will be reported in the next deliverable.

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.

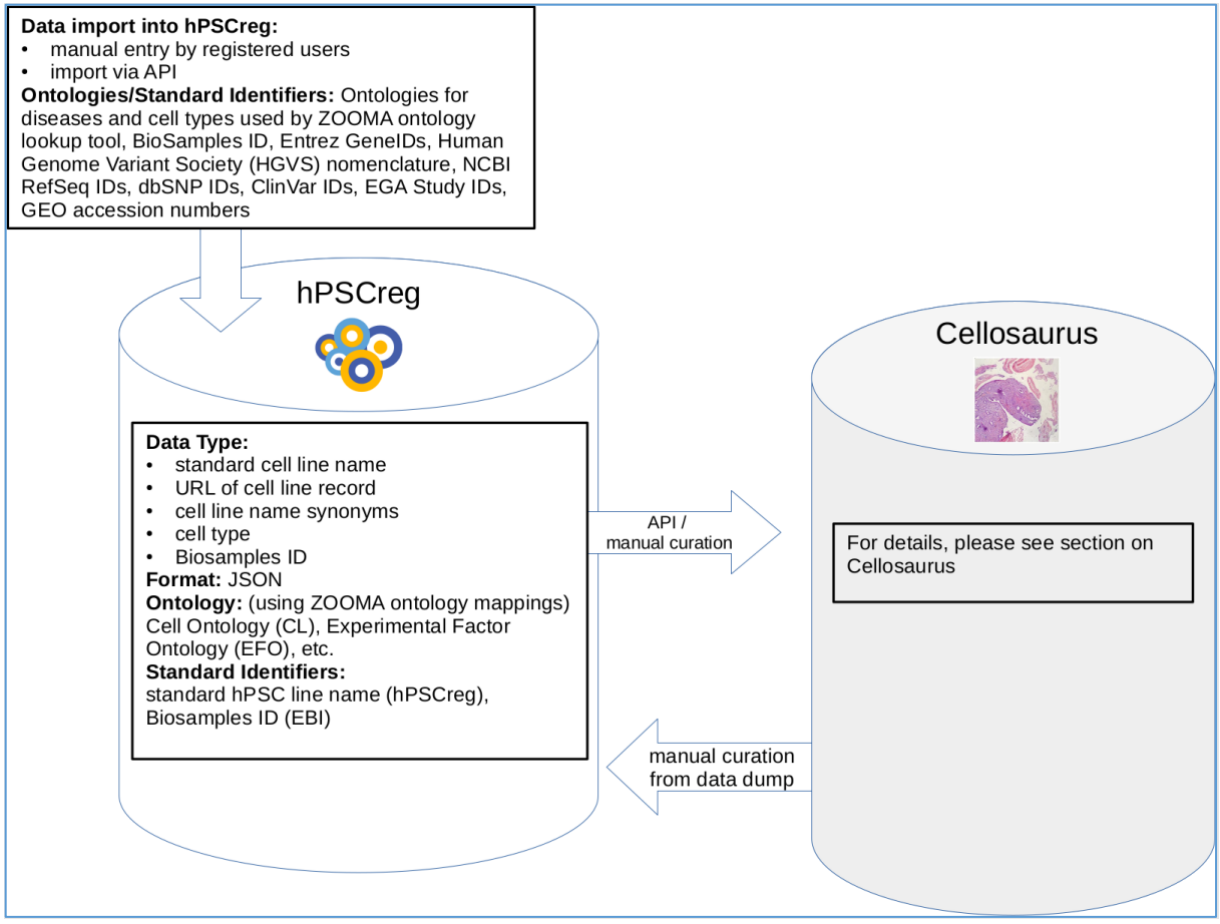

Figure 8. Data flow for the human Pluripotent Stem Cell registry (hPSCreg)

II. Features or facilities added in 2022

The import of approximately 2500 hPSCreg cell lines (Healthy, Rare Disease) into RD-Connect has been completed with the help of the RD-Connect / Molgenis Team, and there have been initial discussions for a sustainable, regular interaction between both resources. Bioschemas markup has been added to hPSCreg webpages, allowing for machine readability according the Schema.org specification. Finally, the interaction with Cellosaurus has been optimized to allow faster data imports.

III. Plans for improvement in 2023

Following the developer hackathon held in Barcelona Dec 13-14, 2022, hPSCreg will fulfil the Level 1 connection to the EJP-RD Virtual Platform (VP). The Level 2 connection to the *beta* version of the VP is functional; however, hPSCreg will have to re-create the Level 2 connection to the "live" version of the VP, and this will be done in consultation between hPSCreg and the EJP-RD VP Development Team. HPSCreg will continue to discuss with GPAP the bilateral use case, once both resources are connected to the "live" VP.

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 144'000 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 2022

As of August 2022, the Cellosaurus is available through an API (https://api.cellosaurus.org/). Thanks to this API it is possible to specifically query and retrieved any of the fields in the resource. The API response can be in JSON, XML, text or TSV.

We have retrofitted more than 80% of the cell line entries with information concerning the tissue/organ of origin and the cell type.

We added 2000 new cell line entries that are relevant to rare diseases. The number of rare diseases represented in the Cellosaurus went up from 1229 in the end of 2021 to 1264 at the end of 2022.

III. Plans for improvement in 2023

Anatomical information (tissue/organ) and cell type information will be linked to the UBERON and CL ontologies.

We plan to have a first version of a RDF/TTL version of the Cellosaurus as well as a Virtuoso endpoint to allow querying the resource using SPARQL.

Cellosaurus will continue to work with EJP RD Metadata and Query Builder teams to get Level 1 connection.

2.10 INFRAFRONTIER

Contributors: Sabine Fessele (INFRAFRONTIER), Montse 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 8400 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.

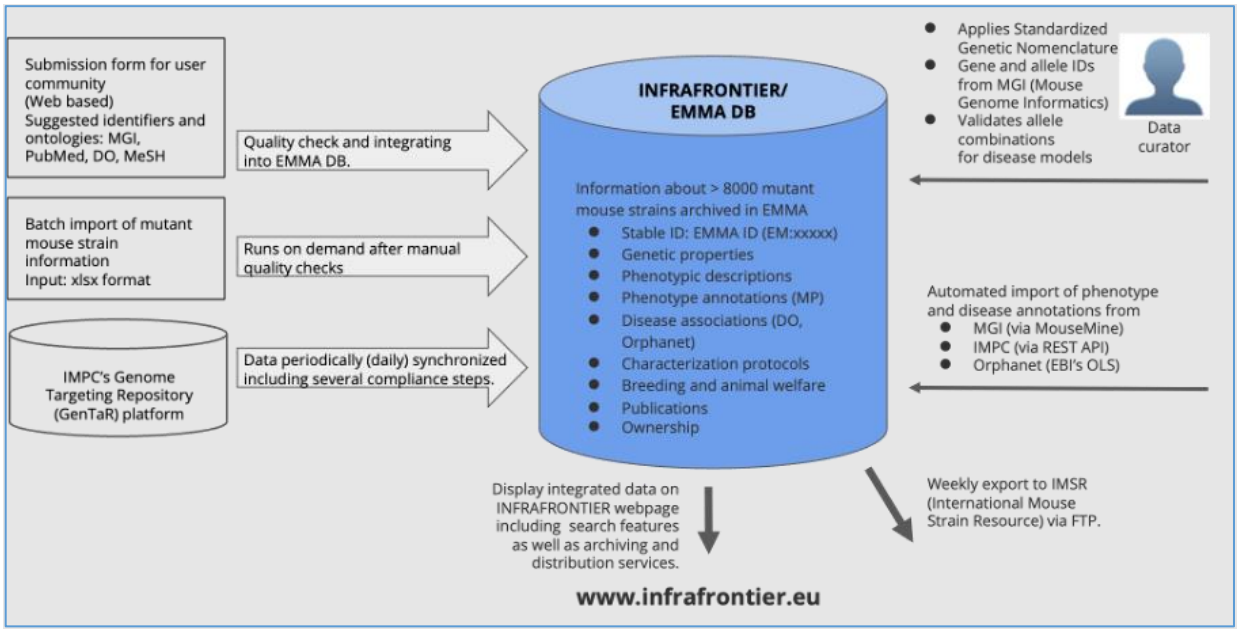

Figure 9. Data flow of EMMA

II. Features or facilities added in 2022

During 2022, 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 2248 mouse strains that carry mutations in 1323 genes that have been implicated to play a role in rare diseases (1670 different rare diseases). The strong growth of these numbers is mainly due to new mutant strains deposited at EMMA and to the effort of the INFRAFRONTIER data curators responsible for the appropriate annotation of these mouse strains.

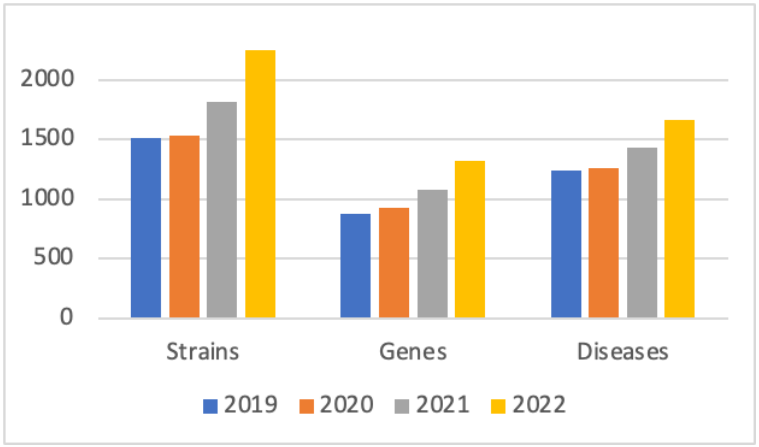

Figure 12. Development of the number of mouse mutant strains available from INFRAFRONTIER/EMMA that carry mutations in genes that are mapped to an Orphanet rare disease nomenclature. The mapping is done via the genes mutated in the different strains.

We also worked on further increasing the visibility of the Rare Disease Resource on the INFRAFRONTIER webpage. When shifting from the Drupal CMS based webpage to a new WordPress webpage, we made sure to preserve all search functionalities we described previously, but also better integrated rare disease related information in the new menu structure (https://www.infrafrontier.eu/services/disease-areas/infrafrontier-and-rare-diseases/).

Figure 13. Screenshot of the INFRAFRONTIER and Rare Diseases landing page on the recently revamped INFRAFRONTIER webpage.

INFRAFRONTIER participated in different meetings and workshops organized by EJP RD in order to start the development of the API for connecting to the EJP RD VP. In the Hackathon "Onboarding and connection of resources to the EJP RD Virtual Platform", which took place in Barcelona in December, we tested the FAIR-in-a-box technology for Level 1 connection to the Virtual Platform and started planning Level 2 connection as well. The decision on what technology to select will be based on the updated recommendations from the EJP RD expert groups.

In addition, we contributed to the communication about the EJP RD Innovation Management Toolbox (IMT), as well as different EJP RD initiatives (e.g. funding opportunities) via Twitter, see a few examples below:

- Innovation Management Toolbox (IMT): https://twitter.com/InfrafrontierEU/status/1550012708052926464
- ERN Research Training Workshops funding opportunity: https://twitter.com/InfrafrontierEU/status/1565308534891446272

III. Plans for improvement in 2023

From INFRAFRONTIER's side, we will complete Level 1 connection to the Virtual Platform and conclude the development of the API needed for Level 2 and 3 connection (initiated in 2022). This will allow INFRAFRONTIER's mutant mouse strains potentially interesting for rare disease researchers to be searchable via the EJP RD Virtual Platform.

Once INFRAFRONTIER will be one of the EJP RD Virtual Platform connected resources, we will focus on the potential connection to other EJP RD partner resources which had been postponed in the past due to the lack of personnel resources.

The collaboration with RD-Connect GPAP, which aims at improving the analysis capabilities in rare disease research, 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.

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: Claire O'Donovan (EMBL-EBI)

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 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.

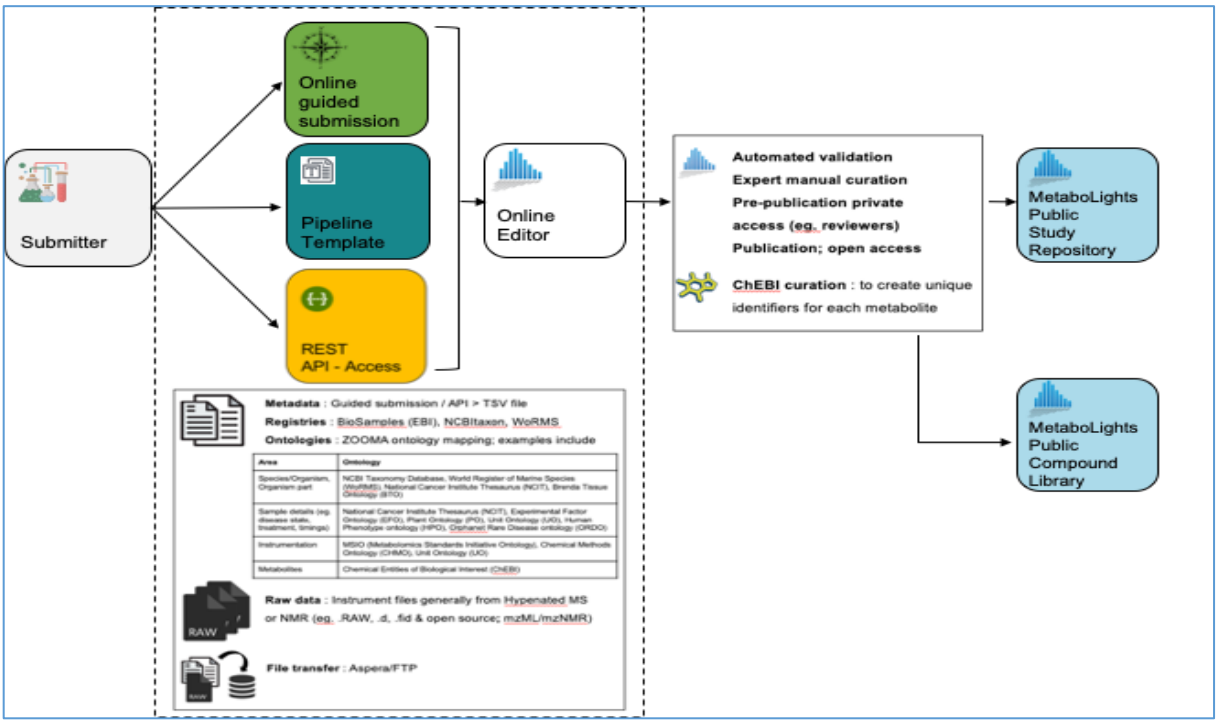

Figure 14. Data Flow for Metabolights

II. Features or Facilities added in 2022

The MetaboLights team continued their interaction with clinical researchers, expanding the ontologies and adding more reference examples. One particular area where we contributed was in the epidemiology workshop at the annual Metabolomics society meeting in Valencia, Spain in June 2022. This engagement with this community was fruitful and there will be another workshop in 2023. Callum Martin, a software engineer in the Metabolomics team, took part in the EJP RD Bio-Hackathon in Barcelona in December 2022 where he began the coding of the integration of MetaboLights with the EJP RD Virtual Platform. This had been preceded by a number of calls/meetings with EJPRD co-ordinators and expert advisors in the specifications needed.

III. Plans for improvement in 2023

We will complete the integration of MetaboLights with the virtual platform and continue to collect specifications/clinical data to enable the highlighting of gold standard examples for the community to use going forward to ensure FAIR principles and the maximum use of research done in this area. We will promote EJP RD once again at the annual Metabolomics society meeting in Niagara Falls, Canada. This will happen at two workshops: ***Multidisciplinary Metabolomic Epidemiology: The Pathway to Clinical Translation*** and ***Data Standardization and Reuse through Public Repositories.*** This conference is the largest and most globally attended one for the metabolomics community and has a significant attendance by clinicians.

2.12 Care and Trial Site Registry (CTSR)

Contributors: Hanns Lochmüller, Adrian Tassoni, Sunil Rodger, Dagmar Jäger (UKL-FR)

I. Resource data flow

The Care and Trial Site Registry is an online self-report database of clinical care and trial sites that see patients with neuromuscular and neurodegenerative conditions. It is an IRDiRC-recognised resource containing information about site experience in the provision of clinical care and trials, their facilities, equipment, and personnel, and aggregate data about their patient population.

The data held by the CTSR permits the identification of sites within and across particular geographies that match specific criteria. This aids in the planning and conduct of feasibility studies, clinical trials and other research projects by industry, academic researchers, and patient organisations. By enabling the reporting of site-level data, the CTSR also supports networks including the EURO-NMD ERN for Rare Neuromuscular Diseases and NMD4C, the Neuromuscular Network for Canada.

II. Features or Facilities added in 2022

During 2022, we have undertaken several significant steps in the progressive FAIRification of the CTSR and its integration with the EJP Virtual Platform. These have been in close collaboration with colleagues from across Pillar 2 Work Foci, including those in the Metadata, Resources Sharing Experimental Data & Materials, FAIRification, Use Case, and CDE Semantic Model WFs, and have included:

- The integration of the CTSR as a discoverable resource in the EJP Virtual Platform ('VP Ready').
- The identification of data held by the CTSR that is the highest priority for record-level FAIRification ('VP Queryable'), based in the first instance on the common use-case of being able to identify sites by RD conditions they see and their geography.
- The development of a DCAT-based model that captures the 'site-level' nature of the data held by the CSTR, and the mapping of data prioritised for FAIRification to appropriate ontologies for this use-case.
- The exploration and evaluation of potential technical approaches to FAIRification, enabling the choice of a data flow based on .csv, R2RML, and a GraphDB triple-store.
- The development of a functional proof-of-concept of this data flow, with expansion to cover additional data fields in progress as of December 2022.
- The deployment of a FAIR Data Point for the CTSR (https://w3id.org/ctsr-fdp)

In addition, the underlying architecture and the user interface of the CTSR is currently being upgraded in order to permit extensibility, expansion of data held to meet new and emerging needs, and additional features for collaboration and support of partner activities. Together, these steps have laid the groundwork for the continued, stepwise FAIRification of the CTSR during 2023.

III. Plans for improvement in 2023

The overall focus during 2023 will be the delivery of the CTSR as a FAIRified VP Queryable resource. This will involve moving from the current proof-of-concept workflow to one that is production ready, and the progressive integration of additional data fields within this FAIR data flow to enable more complex querying. In turn this will allow for additional use cases to be explored, providing the opportunity to explore how different scales and scopes of data relating to RD might be integrated (e.g. the site-level data held by the CTSR with that held by patient registries) in a FAIR manner.

3. Conclusions

During 2022 improvements regarding rare disease data deposition and access have been implemented by the resources involved in subtask 11.3.1.

Easier access to rare disease data will be achieved by the transition from ELIXIR AAI to Life Science (LS) Login by multiple resources. LS Login makes access to life science resources and services easier to users by providing a single Login to access different platforms, and allows service providers to control and manage the access rights and create different access levels to the data. Resources like EGA, RD-Connect Sample Catalogue or RD-Connect GPAP (which is currently evaluating its implementation in production).

In addition, multiple resource developments aim to improve data access and management workflows. Some examples are: RD-Connect GPAP, which has improved their Data Management Module, which allows for a better guidance of the submission process and reduces its duration; DECIPHER, which allows now to discern between confirmed and assumed parental relationships for de novo variants; BBMRI-ERIC Directory, which improved the user-interface of the platform and simplified the submission data model; hPSCreg, which has added Bioschemas markup to their website, allowing machine readability according the Schema.org specification; Cellosaurus, which has implemented an API that allows queries and retrieval of any data field in their resource; INFRAFRONTIER, by modifying their website so that rare disease related information is better integrated in the new structure. These are some examples among 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. Moreover, these new developments are supported by updating and/or creating new guidelines and documentation, which are accessible from the resources' websites.

Progress is being made on a variety of RD use cases to increase interoperability between the resources. For instance, RD-Connect GPAP is working with hPSCreg to connect the two platforms through an API, which will allow RD-Connect GPAP users to discover relevant cell lines interesting for their studies in hPSCreg. hPSCreg is also working with the RD-Connect Sample Catalogue to link their cell lines to the catalogue. And INFRAFRONTIER is collaborating with RD-Connect GPAP to allow gene queries from GPAP to the mouse mutant database INFRAFRONTIER.

Finally, during 2022 resources have been working to be compatible with the EJP-RD Virtual Platform (VP) Network. Meetings between the Query Builder and Metadata teams with the resources to work on their connection to the VP have taken place throughout the whole year, and on 13-14th December 2022 CNAG-CRG organized the "Hackathon: Onboarding and connection of the resources to the EJP-RD Virtual Platform" in Barcelona. The aim of this Hackathon was to accelerate resources' Level 1 and Level 2 connections to the VP Network. This Hackathon contributed to clarify the different connection options, and at the present time most resources are working on finalizing their Level 1 connection. Work will continue during 2023 for resources to evaluate and implement new levels of connection to the VP Network.