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

D11.15
Fifth 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 (RD Connect GPAP), DECIPHER, RD-Connect Registry and Biobank Finder, Care and Trial Site Registry (CTSR)

Due date of deliverable: Month 60
Dissemination level: Public

Table of Contents

1. Introduction *4*

2. Resources *5*

2.1 European Genome-phenome Archive (EGA) *5*

I. Resource data flow 5

II. Features or facilities added in 2023 6

III. Plans for improvement in 2024 7

2.2 RD-Connect GPAP *7*

I. Resource data flow 7

II. Features or functionalities added in 2023 8

III. Plans for improvement in 2024 10

2.3 DECIPHER *10*

I. Resource data flow 10

II. Features or facilities added in 2023 11

III. Plans for improvement in 2024 (complete only if you will adhere to the 2024 extension) 11

2.4 RD-Connect Registry and Biobank Finder *11*

I. Resource data flow 11

II. Features or facilities added in 2023 12

III. Plans for improvement in 2024 12

2.5 RD-Connect Sample Catalogue *12*

I. Resource data flow 12

II. Features or facilities added in 2023 14

III. Plans for improvement in 2024 14

2.6 BBMRI-ERIC Directory *14*

I. Resource data flow 14

II. Features or facilities added in 2023 15

III. Plans for improvement in 2024 15

2.7 Rare Disease Cohorts (RaDiCo) *16*

I. Resource data flow 16

II. Improvements made in 2023 19

III. Improvement planned for 2024 (complete only if you will adhere to the 2024 extension) 19
2.8 hPSCreg *19*
I. Resource data flow 19
II. Features or facilities added in 2023 20
III. Plans for improvement in 2024 21
2.9 Cellosaurus *21*
I. Resource data flow 21
II. Features or facilities added in 2023 22
III. Plans for improvement in 2024 22
2.10 INFRAFRONTIER *22*
I. Resource data flow 22
II. Features or facilities added in 2023 23
III. Plans for improvement in 2024 24
2.11 MetaboLights *24*
I. Resource data flow 24
II. Features or Facilities added in 2023 25
III. Plans for improvement in 2024 26
2.12 Care and Trial Site Registry (CTSR) *26*
I. Resource data flow 26
II. Features or Facilities added in 2023 26
III. Plans for improvement in 2024 26
3. Conclusions *27*

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, always following the FAIR (Findable, Accessible, Interoperable and Reusable) principles.

Work done in this subtask include data deposition, with the development of new user-friendly interfaces which improves users' rare disease data and metadata submission; data quality, by the addition of manual curation steps and automated metrics; data access, with the development of APIs to provide a better query functionality; and data security, by following legal requirements (i.e. EU GDPR or national legislations) and recommendations (i.e. the Global Alliance for Genomics and Health, GA4GH).

In addition, Subtask 11.3.1 has supported the EJP-RD resources in successfully connecting to the EJP-RD Virtual Platform (https://vp.ejprarediseases.org/discovery/), an ecosystem of data resources and auxiliary services to find, access, and use data for various purposes, which has involved the efforts of different Work Foci like Overall Architecture, FAIRification, Query Builder or Metadata.

This deliverable summarizes the additional capabilities added during 2023, and the work done to connect to the EJP-RD VP by the EJP-RD funded resources. In addition, it states the work some of the resources plan to do during 2024. EJP-RD has extended its project duration until August 2024, but not all resources will join this extension. Therefore, those resources not joining the project extension during 2024 have "Not applicable" in their respective "Plans for improvement in 2024" sections.

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

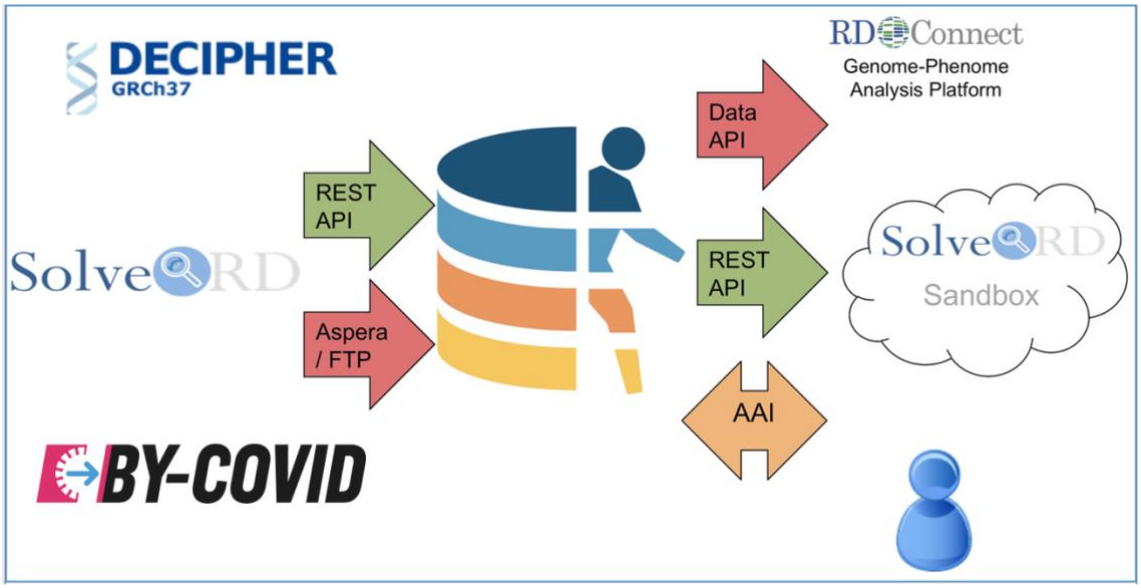

Figure 1. *Example data flows to and from EGA. Submitters, such as Solve-RD or DECIPHER, submit data to the EGA for archive and distribution. These data are distributed via the EGA Data API to authorised users. Use-cases include distributing data to the RD-Connect Genome-Phenome Analysis Platform, the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

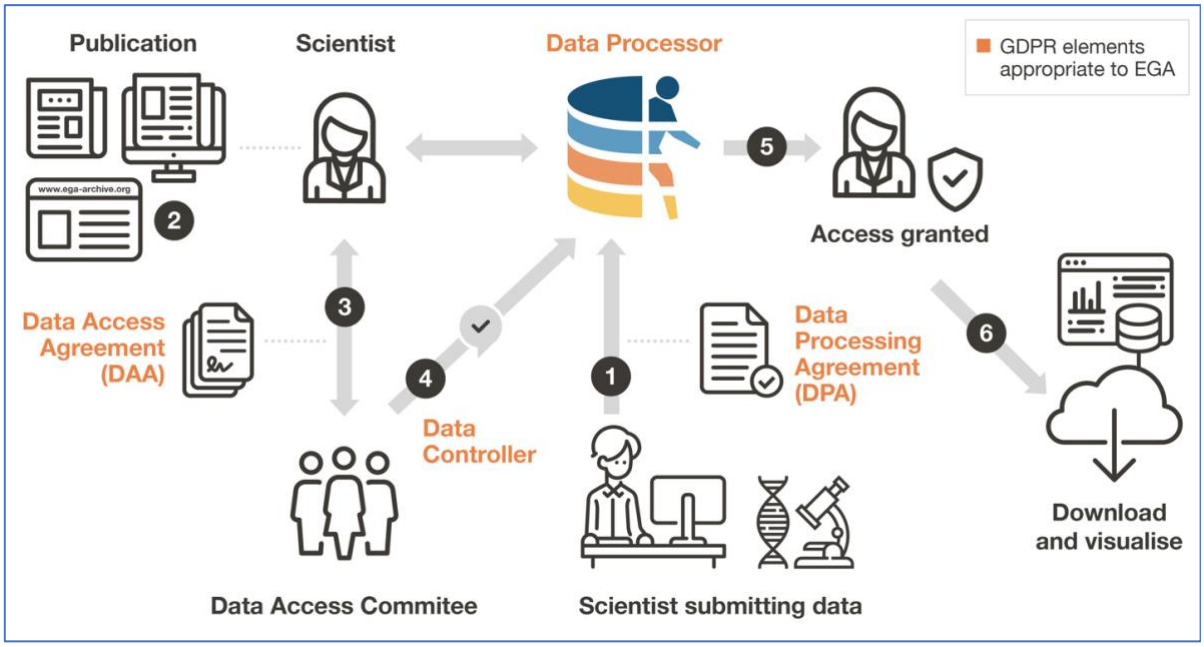

Figure 2. *EGA facilitates the submission, discovery, access and distribution of sensitive human data. A researcher submits controlled access human genetic, phenotypic and clinical data to EGA after signing a Data Processing Agreement (1). EGA processes, archives and releases the dataset to be findable. 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 2023

EGA successfully released a new version of the EGA Submitter Portal (https://submission.ega-archive.org/take-the-tour/) with a more structured and simplified user interface which eases the navigation and usability throughout the EGA submission process. Examples of usability improvements are, for instance, a clearer "Submitter request form" to obtain submitter credentials, or the possibility to add collaborators for the metadata submission. This release has also provided enhanced

features for users, like 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.

EGA has also released a new Data Access Committee (DAC) Portal (https://dac.ega-archive.org/), which allows DACs to manage access requests in an easier and more straightforward way. A whole new workspace for DACs has been created, offering a comprehensive interface for the oversight of registered DACs. This platform includes advanced tools for the efficient management of access requests, allowing for streamlined evaluation, acceptance, and rejection procedures. The new EGA DAC Portal represents a substantial improvement in the administration of access to sensitive data while maintaining the utmost standards of data protection. For instance, this new user-friendly interface will allow entities such as rare disease consortia, overseeing extensive datasets, to efficiently process multiple access requests with just a few clicks.

In addition, documentation and blogposts have been created for the new EGA Submitter Portal (https://ega-archive.org/submission/metadata/submission/sequencing-phenotype/submitter-portal/ and https://blog.ega-archive.org/new-submitter-portal), and for the EGA DAC Portal (https://ega-archive.org/access/data-access-committee/dac-portal/ and https://blog.ega-archive.org/new-dac-portal).

Finally, EGA followed the necessary steps to connect to the EJP-RD Virtual Platform. EGA created their own FDP at the CRG servers and has populated it with the necessary EGA metadata. Currently EGA is discoverable through the EJP-RD resources' hardcoded list, and once the VP Index is up and running, it will be discoverable through its own FDP.

III. Plans for improvement in 2024

Not applicable.

2.2 RD-Connect GPAP

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

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.

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 may be deposited at the EGA for long-term archive and controlled access.

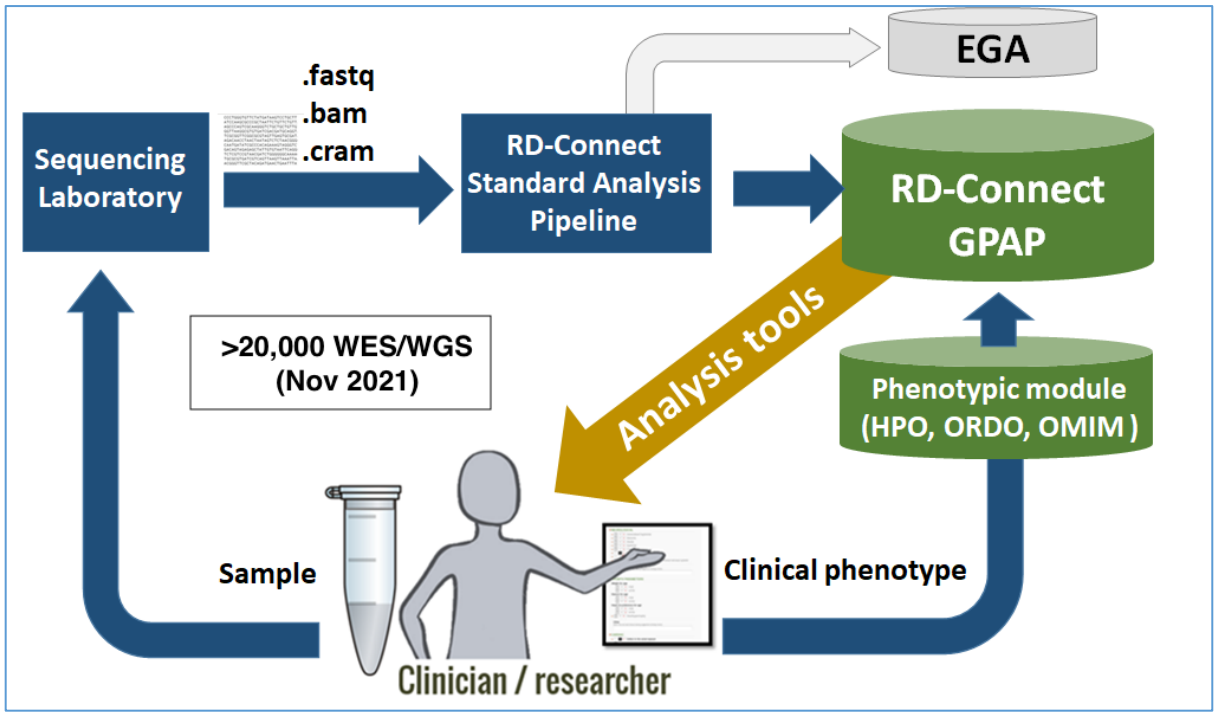

Figure 2. Data flow for RD-Connect GPAP

II. Features or functionalities added in 2023

In the RD-Connect GPAP PhenoStore module we have added a User profile view, where the user can easily get the specific numbers relating to their activities.

Figure 4: *User profile view in the RD-Connect GPAP PhenoStore*

The RD-Connect GPAP has launched its connection to the EJP-RD VP using real data instead of synthetic data from its Playground. The Beacon 2 API has been implemented and the RD-Connect GPAP is connected to the Virtual Platform on Level1 and is available for Level 2 queries.

On the bi-weekly EJP-RD Beacon 2 meetings, we have been working together with other resources that would like to connect to the Virtual Platform through the Beacon API.

We have evaluated the inclusion of additional variant types, and we have worked on the inclusion of CNVs. The platform now is ready to accept CNV data, although it still needs development to facilitate CNV interpretation.

The RD-Connect GPAP is ready to connect to the LifeScience AAI in the technical terms. It should be still discussed 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.

Together with the ERP-RD coordination team we have organised a webinar to show the recently released RD-Connect GPAP user interface and functionalities and to train on their use. The webinar was held on 28th June 2023 online, and the recording is available on the EJP-RD YouTube channel: https://www.youtube.com/watch?v=-do0uF-gilg.

We have held an "EJP-RD Virtual Platform data resources: assets and legacy" workshop on 21st July 2023, where the resources presented how is their resource serving

the RD community better since the EJP-RD project started, what were the key developments funded by the project, what do they aim to finish by the end of the project and what are their plans regarding the connection to the EJP-RD Virtual Platform.

We have been working on a use case description of a pilot to use a Privacy-Preserving Record Linkage (PPRL) service to link pseudonymised records from different contexts with different pseudonyms without disclosing the patient´s identity. The use case describes a pilot connecting the RD-Connect GPAP Playground instance and the SIOPEN Bioportal via EUPID services based on mock data. We are expecting to finish the testing of 6 cases between the two platform before the end of the project.

III. Plans for improvement in 2024

We will work further on the Level 2 connection between the RD-Connect GPAP and the Virtual Platform in line with the ongoing discussions and decisions about the onboarding process in the project.

We will continue working on the pilot to demonstrate the use of PPRL service, specifically on the use case that describes connecting the RD-Connect GPAP Playground instance and the SIOPEN Bioportal via EUPID services, and we will evaluate the testing results.

We will carry on with the deployment of the CNV inclusion to the RD-Connect GPAP to facilitate CNV interpretation.

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).

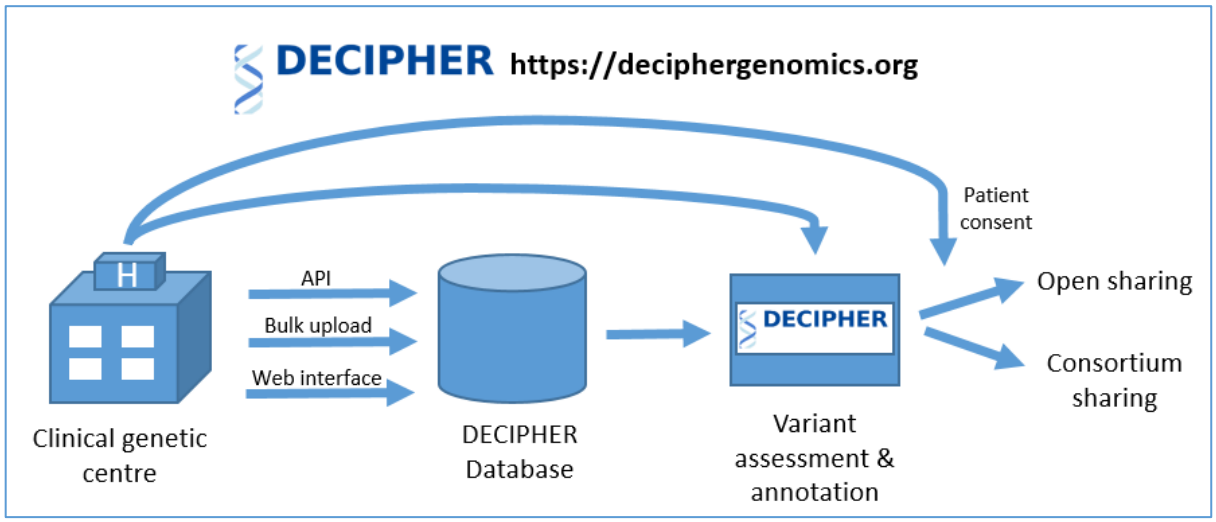

***Figure 3**. DECIPHER Data Flow*

II. Features or facilities added in 2023

No update was reported for Decipher.

III. Plans for improvement in 2024 (complete only if you will adhere to the 2024 extension)

No plans were reported for Decipher.

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 - https://doi.org/10.1038/s41431-017-0085-z).

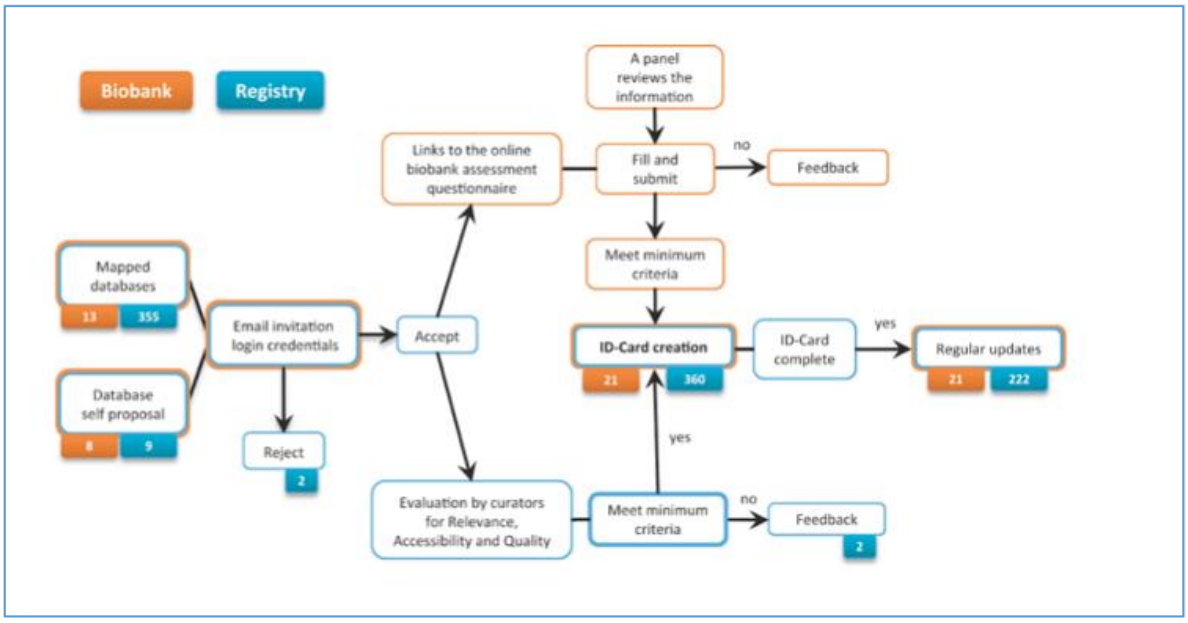

Figure 4. Inclusion of biobanks and registries in the Registry and Biobank Finder. The process of inclusion and evaluation of biobanks and registries in the Registry and Biobank Finder (mapped and self-proposed)

II. Features or facilities added in 2023

The RD-Connect Registry & Biobank Finder has been fully integrated as a vital component within the BBMRI-ERIC directory. This integration represents a significant enhancement, streamlining the interface and access between these critical resources in the biobanking community. To ensure continuity and support for ongoing research, the original software platform, crafted with the robust MOLGENIS framework, has been preserved and is accessible for users who wish to reference legacy data and utilize familiar functionalities. This dual-availability approach ensures that the transition to the integrated system is smooth, and that historical data remains intact and accessible for comparative and longitudinal studies.

III. Plans for improvement in 2024

See 2.6 BBMRI-ERIC Directory.

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.

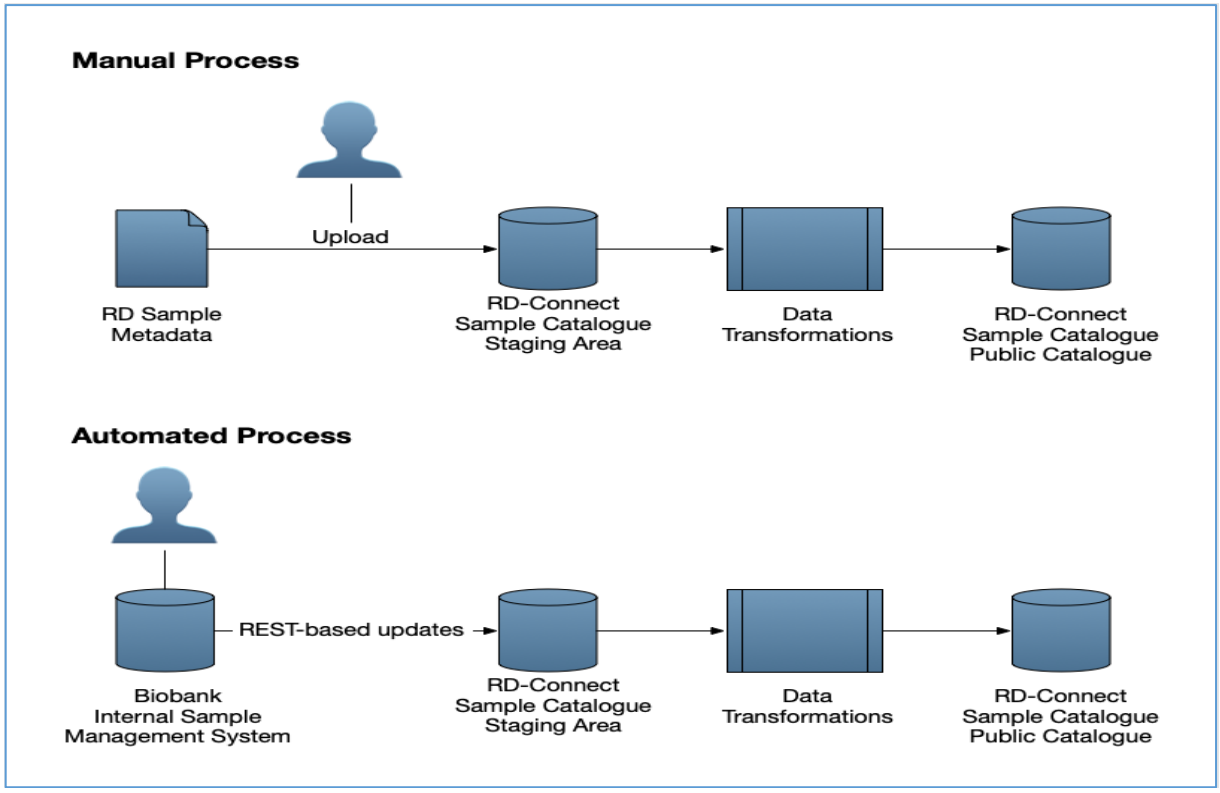

Figure 5. RD Connect Sample Catalogue Data Flow

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 2023

In 2023 we have been working on updating the underlying MOLGENIS platform, no new features were added to the RD-Connect installation, but we did build an integration with the BBMRI-ERIC Directory to show the biobanks as part of the Rare Disease network and populate the facts table with the sample counts from the RD-Connect Sample Catalogue.

III. Plans for improvement in 2024

Not applicable.

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 files ingest of CSV files from the National Directory
- Programmatic updates initiated by the National Directory (using the Directory's RESTful API's)

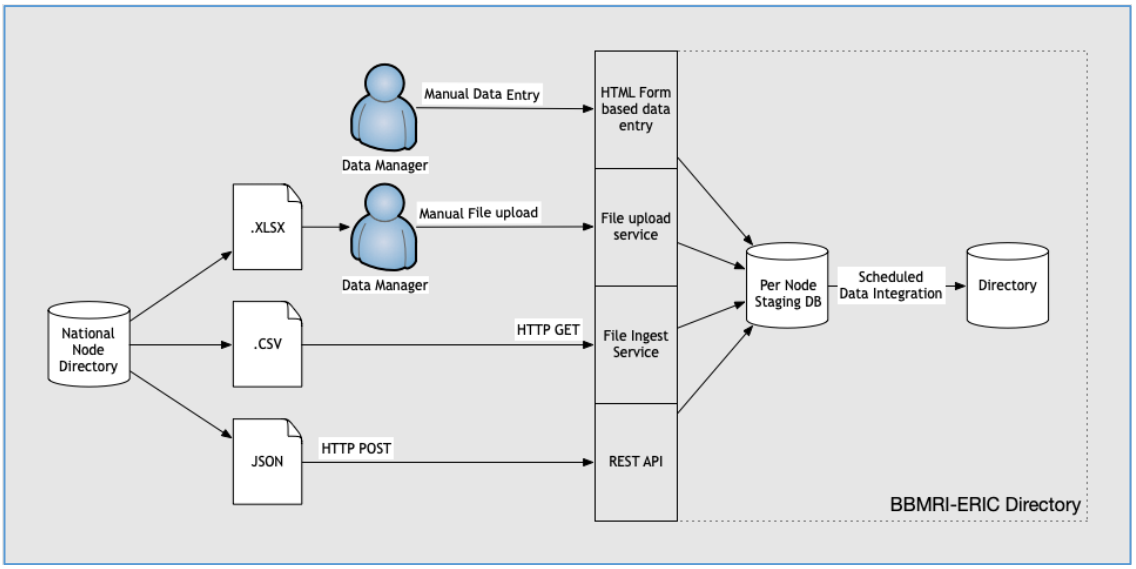

Figure 6. BBMRI-ERIC Directory data flow

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 2023

In 2023 we have continued to improve the usability of the Directory. We added several user experience improvements to make the Directory easier to use: the diagnosis available filter has been improved to show the structure of ICD-10 and split the search in an ICD-10 and Orphanet (ORPHAcodes) search; the filters adapt to remove options that have no results; we allow biobanks, collections and networks to provide links to other sites; filters now sort their options alphabetically; and commercial use is now a toggle instead of two mutually exclusive checkboxes;

We also created a proof of concept for a landing page for the directory that will give the users more guidance on how to use the Directory.
Also we have made the released a version of the Directory that includes the facts table (using a star schema implementation, https://en.wikipedia.org/wiki/Star_schema) to provide a breakdown of the samples by their main criteria: diagnosis, gender, age, sample type and we updated the data model of the Directory to remove some ambiguities in the data entry and have reviewed and updated the manuals for end users and data managers to reflect the latest status.

The FAIR Datapoint of the Directory has been updated to be compatible with the VIP specifications. We have integrated the sample information from the RD-Connect Sample Catalogue in the facts table and created a Rare Disease network and Rare Disease category to make Rare Disease biobanks more easily accessible.

III. Plans for improvement in 2024

Continuing the trajectory into 2024, our comprehensive plan will also include the following critical components:

- Bug Fixing: Implement a rigorous, ongoing bug tracking and fixing protocol. This will not only address current issues but also anticipate and mitigate future glitches through continuous monitoring and regular updates to the system.

- Data Stewardship: Execute a strategic plan for the transfer and synchronization of biobank and collection resources from the old RD-Connect Registry and Biobank Finder. This will be accomplished either through national nodes or directly via the BBMRI-ERIC HQ, ensuring a smooth migration to the new database while aligning with changes in the metadata schema.
- Metadata Schema Updates: Align the Directory with the latest changes in the metadata schema, which will involve updating the data entry protocols and ensuring that all data stewards are fully trained in the new system.
- Proof of Concept (POC) for DUC/CCEs: Develop and implement a proof of concept for Dynamic User Controls (DUC) and Complex Custom Events (CCEs). This will allow us to test and refine the ways in which we handle user permissions and event triggers within the Directory, paving the way for more sophisticated data management and user interaction capabilities.

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.

- **Thériaque:** Thériaque is a database of all medicines available in France for health professionals.

RaDiCo Information System user's organization

The organization of the user in the RaDiCo Information System (IS) reflects the structure of the cohort, with specific roles identified within the system. 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)

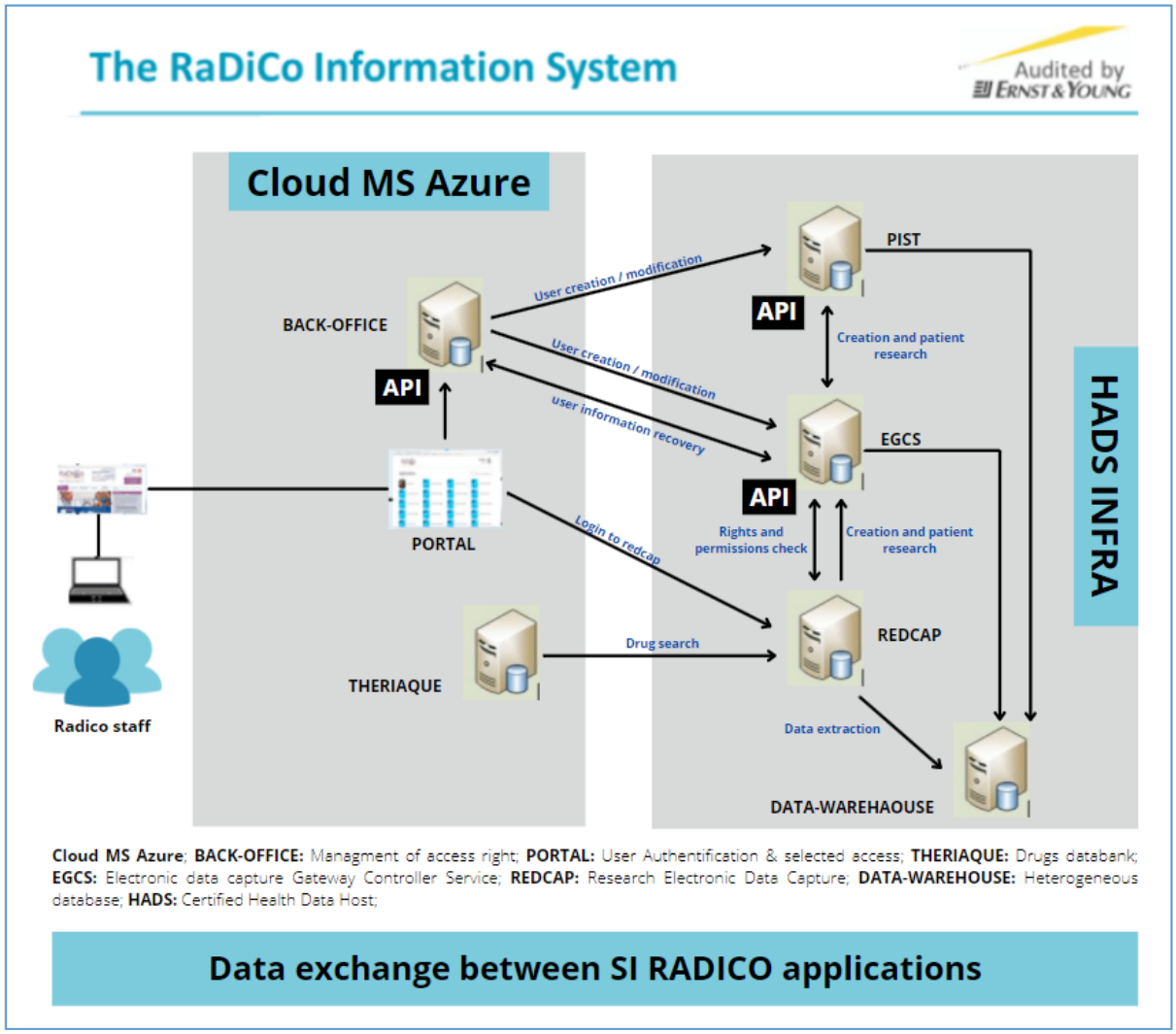

Figure 7. The RaDiCo Information System

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 2023

Preparations for the implementation of WP11 have been delayed due to the appointment of a new IT team, and also to waiting for final documentation from EJP-RD.
Today RaDiCo's priority is to establish the connection with the EJP-RD virtual platform for level 1.
To this end, workshops have been held to define the necessary IT infrastructure requirements, as well as a meeting with the EJP-RD team to discuss the possible choices concerning the connection and data transformation APIs on the FDP Fair Data Point model, in particular with MOLGENIS platform.
The RaDiCo endpoint is now published (https://index.vp.ejprarediseases.org/).

III. Improvement planned for 2024 (complete only if you will adhere to the 2024 extension)

Subject to obtaining the necessary budgets, these few months of extension will enable RaDiCo to extend the connection to the virtual platform to level 2 for the FDP data models and to other data resources such as Orphanet.

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.

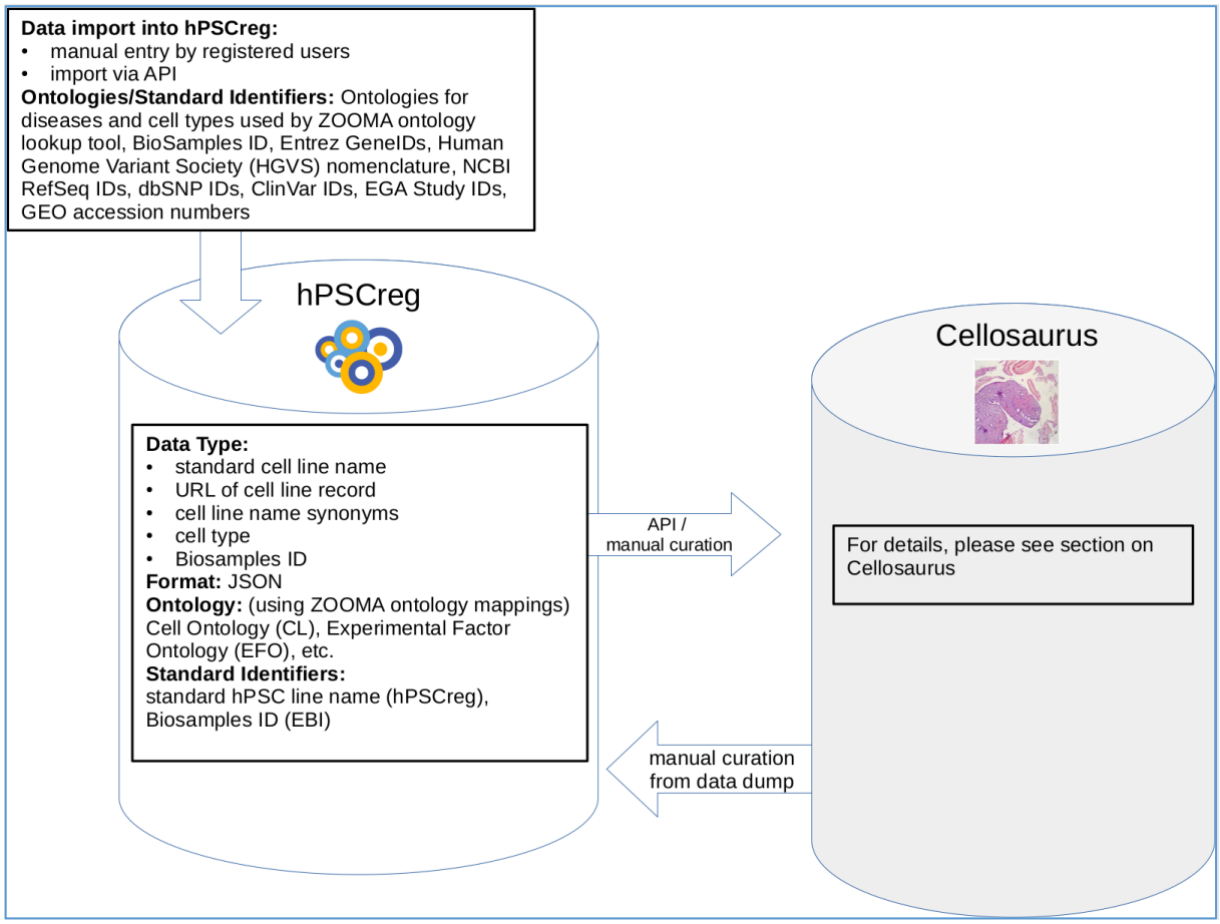

Figure 8. Data flow for the human Pluripotent Stem Cell registry (hPSCreg)

II. Features or facilities added in 2023

The connection between hPSCreg and the EJP-RD Virtual Platform will be completed at Level 1 (resource is discoverable in the VP). Further promotion of human pluripotent stem cell lines to the rare disease community includes the development of rare disease pages, showing which hPSC lines are associated with a particular rare disease (annotated by ORDO terms), and a direct link to the Orphanet record of that rare disease (Figure 11). Finally, hPSCreg will encourage the registration of new hPSC lines associated to rare disease by developing a guidance document for rare disease researchers.

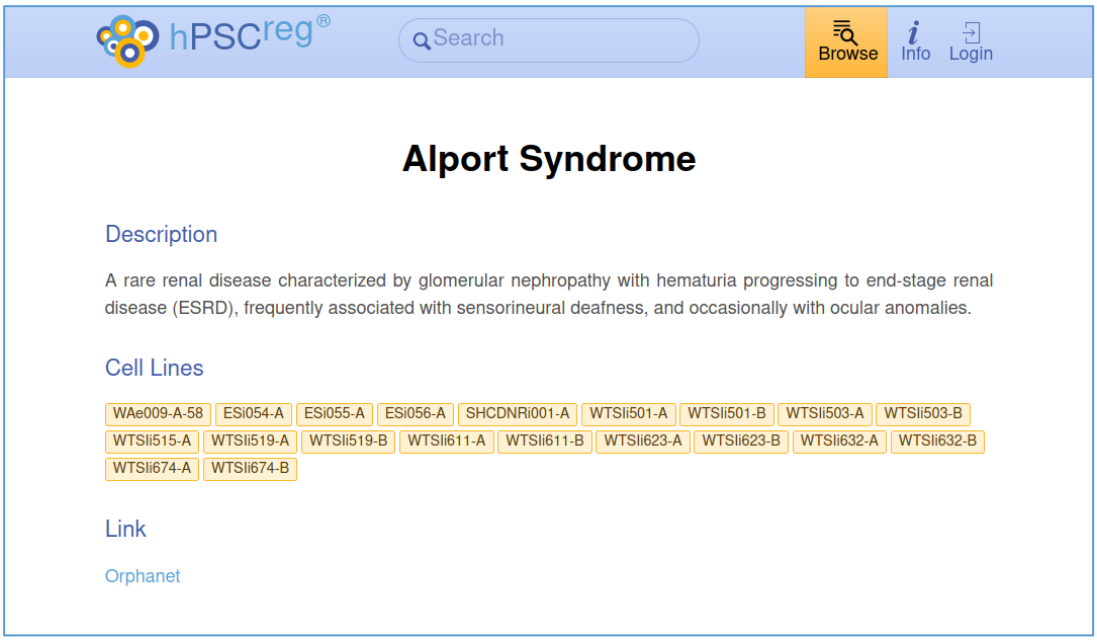

Figure 11. Mock-up of rare disease page for Alport Syndrome on hPSCreg.

III. Plans for improvement in 2024

Not applicable.

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 2023

No update was reported for Cellosaurus

III. Plans for improvement in 2024

No plans were reported for Cellosaurus.

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.

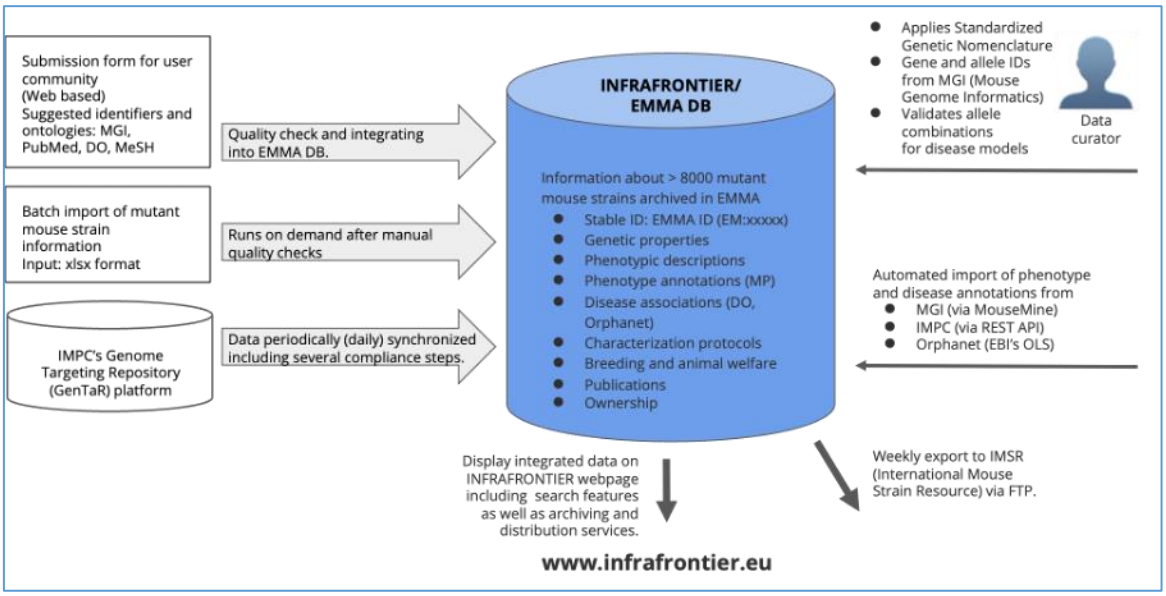

Figure 9. Data flow of EMMA

II. Features or facilities added in 2023

INFRAFRONTIER tasks during the first half of 2023 were severely impaired by a cyberattack suffered by our server provider in March 2023.
However, the list of EMMA strains that are potentially interesting for rare disease researchers on INFRAFRONTIER's rare disease landing page has further increased. Currently the EMMA repository holds 2279 mouse strains that carry mutations in 1340 genes that have been implicated to play a role in rare diseases (1692 different rare diseases). The numbers reported correspond to 15th March and could not be updated since then because of the above mentioned cyberattack. However, we expect a higher increase similar to previous years, when we can get the real numbers for 2023.

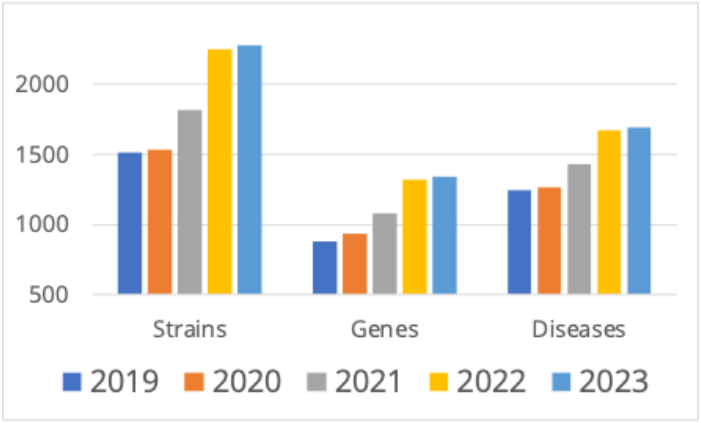

Figure 13. Development of the number of mouse mutant strains available from INFRAFRONTIER/EMMA that carry mutations in genes that are mapped to an Orphanet rare disease nomenclature. The mapping is done via the genes mutated in the different strains.

INFRAFRONTIER participated in different meetings and workshops organized by EJP RD in order to continue with the development of the FDP for connecting to the EJP RD VP.

Therefore, we utilized the FAIR-in-a-box solution to set up an FDP which will be used to enable Level 1 and Level 2 access to information about INFRAFRONTIER mouse mutant strains.

III. Plans for improvement in 2024

The 8 months extension in 2024 will allow us to complete any remaining tasks for connecting to level 2/3 to the EJP RD VP utilizing our FDP.
Additionally, we will connect to other EJP RD resources (RD-Connect GPAP and hPSCreg). Discussions with the two mentioned resources already started in the past.

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.

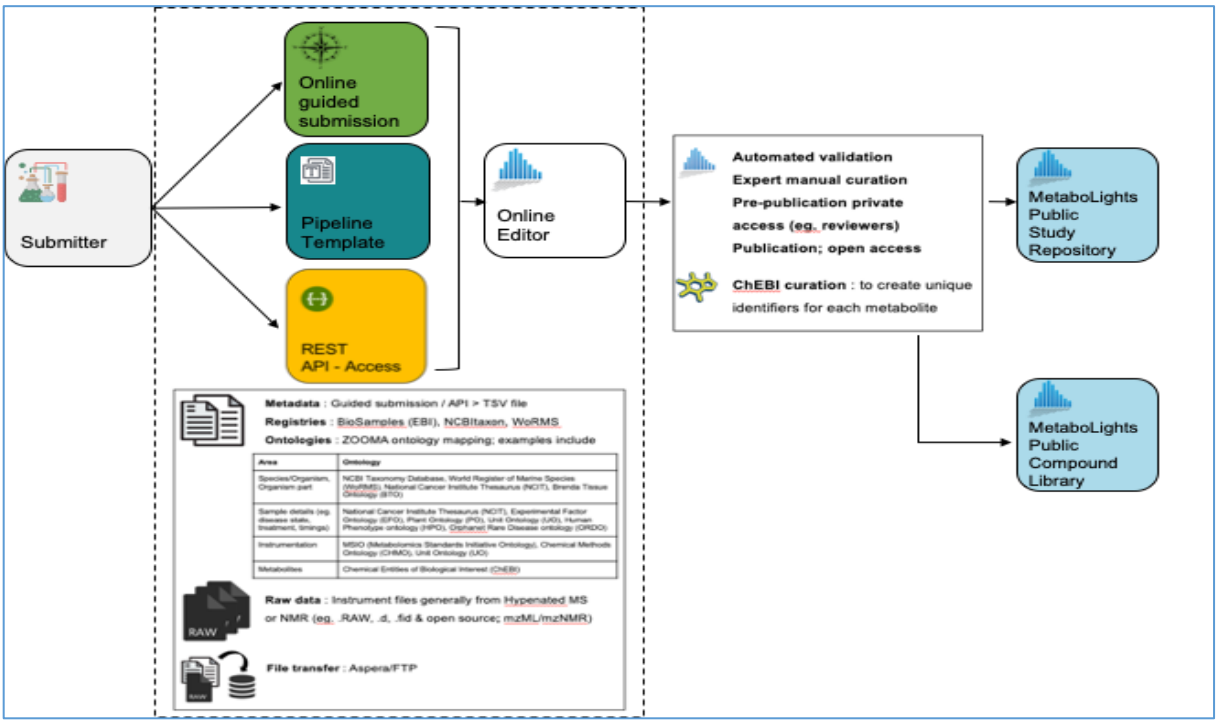

Figure 14. Data Flow for Metabolights

II. Features or Facilities added in 2023

Metabolights created a production FAIR Data Point, which can be found at https://w3id.org/mtbls-ejprd/index.html, using Fair in a Box (https://github.com/markwilkinson/FAIR-in-a-box) and configured for production using the steps outlined by EJP-RD documentation here (https://github.com/ejp-rd-vp/FDP-Configuration).

In addition, Metabolights implemented a translation layer which takes the resource's study metadata and maps it to the resource metadata schema outlined here (https://github.com/ejp-rd-vp/resource-metadata-schema/tree/master), and adds it to the MetaboLights study catalogue. This translation layer is run on a periodic basis to add new study information in the MetaboLights FDP study catalogue.
Moreover, in 2023 we have also been working on improving the breadth of our study indexing, as previously only Isa-tab metadata was available. We now have an internal functionality to index the contents of assay, sample and metabolite annotation information, as well as the free text of the study abstract and protocols. This data can be queried and viewed in dashboard format using kibana / elastic stack tooling. This brings us a step closer to being able facilitate level 2 queries.

Finally, we have had ~2200 studies submitted so far in 2023 (out of 8900 total in the repositories lifetime) for an average of ~7 new studies per day, with each study requiring a dedicated human curatorial effort to be released publicly.

III. Plans for improvement in 2024

Not applicable.

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 2023

During 2023 we have focused on integrating the CTSR as a queryable resource into the Virtual Platform. Further testing of the proof-of-concept data flow developed during 2022 for the CTSR (generating FAIR data based on a CSV export, R2RML, and a GraphDB triple-store linked to the reference implementation of the FAIR Data Point) found that this multi-stage approach was not well-suited to production use for the type of data held by the CTSR (site-level, rather than patient-level data, which does not conform to the EJP CDEs).

Given these challenges, and in collaboration with EJP colleagues, we have developed an alternative approach which serves FAIR data from the CTSR via a low threshold 'text based' FDP. To support this, the CTSR has been modified to directly create RDF data, which is served it via a web interface as its own FAIR Data Point. As such, this entirely avoids the use of the FDP reference implementation, and is indexed in the FDP index via a separate shell script. This approach is in final testing and should support the deployment of the CTSR as a FAIR VP queryable resource before the end of 2023.

III. Plans for improvement in 2024

Not applicable.

3. Conclusions

During 2023 resources involved in Subtask 11.3.1 have continued to implement rare disease data deposition and access improvements. Some examples are; RD-Connect GPAP by creating a new user profile interface in Phenostore, which eases the management of submitted rare disease data by users, and the platform now allows the submission of CNV data; EGA by releasing two new user interfaces, the Submitter Portal and the DAC Portal, which improves the workflow to submit data in the platform and to manage access requests; RD-Connect Registry and Biobank Finder and RD-Connect Sample Catalogue by being integrated into the BBMRI-ERIC Directory, which unifies the capacity to search for key players in the biobanking field in one location, and, in addition, BBMRI-ERIC by implementing new filters, better menus and the capacity for biobanks, collections and networks to provide links to other sites, which improves the platform's user experience; hPSCreg by developing new rare disease pages with links to Orphanet sites, which provides enhanced information for users interested in studying rare diseases; and INFRAFRONTIER, which suffered an important cyberattack, but nevertheless increased the amount of EMMA strains with potential interest to rare diseases.

In addition, these resources have updated/created guidelines and documentation to inform users of these updates, changes and new functionalities. These improvements exemplify significant advances from resources to enhance and facilitate user data deposition and access to further advance rare disease studies.

Resources have also been focused on implementing the necessary steps to connect to the EJP-RD VP. The table below summarizes the status of resources' connections at the end of 2023. Most resources are already discoverable through the EJP-RD VP at Level 1. In addition, RD-Connect GPAP and BBMRI-ERIC are queryable through Level 2 connection, and resources like hPSCreg, Metabolights and CTSR have almost completed their Level 2 connection by the end of 2023.

Resource	Level 1 connection	Level 2 connection
EGA	Completed	Ongoing
RD-Connect GPAP	Completed	Completed
Decipher		
BBMRI-ERIC*	Completed	Completed
RaDiCo	Ongoing	-
hPSCreg	Completed	Almost completed
Cellosaurus	Completed	-
INFRAFRONTIER	Ongoing	-
Metabolights	Completed	Almost completed
CTSR	Completed	Almost completed

* RD-Connect Registry and Biobank Finder and RD-Connect Sample Catalogue are integrated in BBMRI-ERIC Directory.

This work, together with the efforts the resources joining the EJP-RD project extension will continue to do during 2024, will enrich the amount of rare disease data discoverable through the EJP-RD Virtual Platform.