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

D11.11
Additional facilities integrated to resources regarding data deposition and access, including user guidelines and documentation for data deposition and access.

Organisation name of lead beneficiary for this deliverable:
Partner 1 – EMBL-EBI (EGA)

Collaborators: hPSCreg, Cellosaurus, MetaboLights, INFRAFRONTIER, BBMRI-ERIC, RD Connect Sample Catalogue, RaDiCo, RD Connect GPAP, DECIPHER, PRIDE, RD-Connect Registry and Biobank Finder & JRC-ERDRI

Due date of deliverable: Month 12

Dissemination level: Public

Table of contents

List of figures

1. Section 1

1.1. Introduction

This deliverable describes additional facilities that were integrated into the pre-existing resources described below and that regard data deposition and access. It includes user guidelines and documentation for data deposition and access to these resources.

EJP-RD aims to improve the integration, the efficacy, the production and the social impact of research on RD through the development, demonstration and promotion of Europe and world-wide sharing of research and clinical data, materials, processes, knowledge and know-how. To this end, Task 11.3 aims to serve the needs for depositing, integrating and storing quality controlled data and metadata produced by EJP-RD partners and the overall RD community 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 to appropriate public repositories and resources, making them discoverable through the platform.

Over the course of the whole project, the subtask 11.3.1 will support European and international resources and infrastructures highly relevant for the RD community by improving and expanding on their deposition capabilities and access mechanisms. The resources will deploy or expand user-friendly interfaces to deposit data and metadata using HPO, ORDO, OMIM and/or any other relevant ontology or standard, to ensure that data is FAIR[1] (Findable, Accessible, Interoperable and Reusable). In addition, mechanisms to ensure and/or assess the quality of the dataset, through manual curation, automatic generation of metrics or a mixture of both will be deployed. Query functionalities and data access should be possible through application programming interfaces (APIs) and graphical user interfaces (GUI). To build trust from the community the security of these resources will be evaluated and aligned with the Global Alliance for Genomics and Health[2] (GA4GH) recommendations taking into account the GDPR and other national legislations. Whenever relevant, the resources will implement transparency measures and means to follow up re-usability of the submitted dataset. The task also includes implementation and further development of the federated EGA infrastructure and its interoperability with RD-related databases.

Deliverable 11.3 seeks to identify each of the named resources current capabilities and additional capabilities added during 2019 in terms of data deposition and data access, identify areas for improvement which will benefit the RD community and therefore the community as a whole. This deliverable will be used to scope and plan the WF on "Resources for sharing experimental data and materials" work for the coming year.

1 https://www.go-fair.org/fair-principles/
2 https://www.ga4gh.org

2. Section 2 - Resources

2.1. European Genome-Phenome Archive (EGA)

Contributors: Giselle Kerry (EMBL-EBI), Dylan Spalding (EMBL-EBI), Jordi Rambla (CRG), Thomas Keane (EMBL-EBI)

2.1.1. Resource data flow

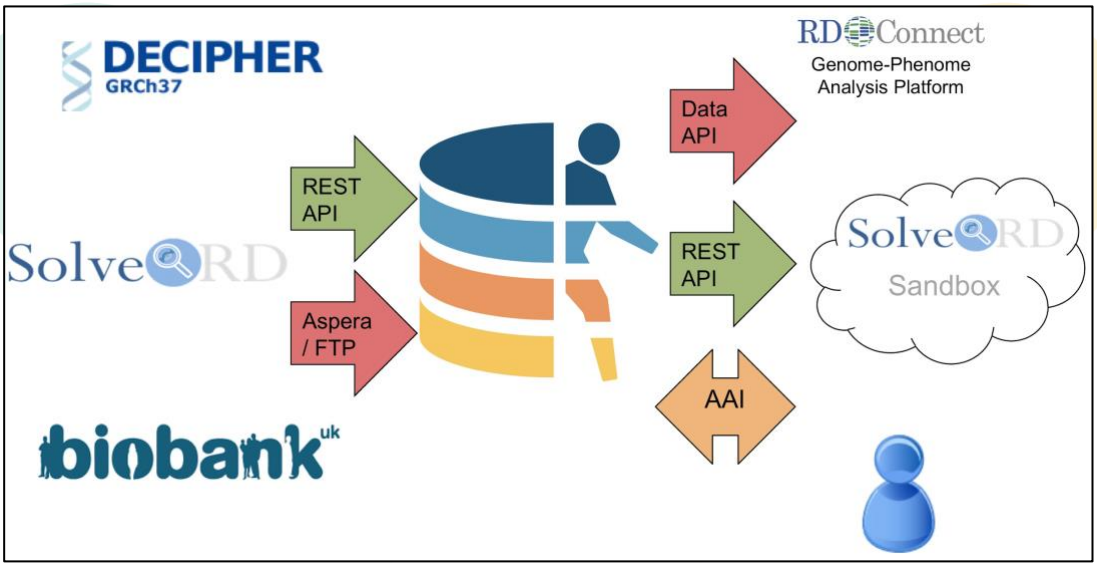

Figure 1 - Example data flows to and from EGA

Submitters, such as Solve-RD or DECIPHER, submit data to the EGA for archival and distribution. These data can then be 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 for further analysis, or individual users for local analysis. All uses 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[3] 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[4] or associated REST APIs[5]. 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

3 https://ega-archive.org/submission/quickguide
4 https://ega-archive.org/submission/tools/submitter-portal
5 https://ega-archive.org/submission/programmatic-submissions

access data is encrypted during transmission and at rest. Users request access to specific datasets via the Data Access Committee (DAC) (Figure 2), who decide if the users proposed research use is in accordance with the specific data use conditions for the dataset. If the user satisfies these conditions, then the user is granted access to the controlled access data via a user account. The user can then log into the EGA and download the data via the EGA Data API[6], FTP or Aspera.

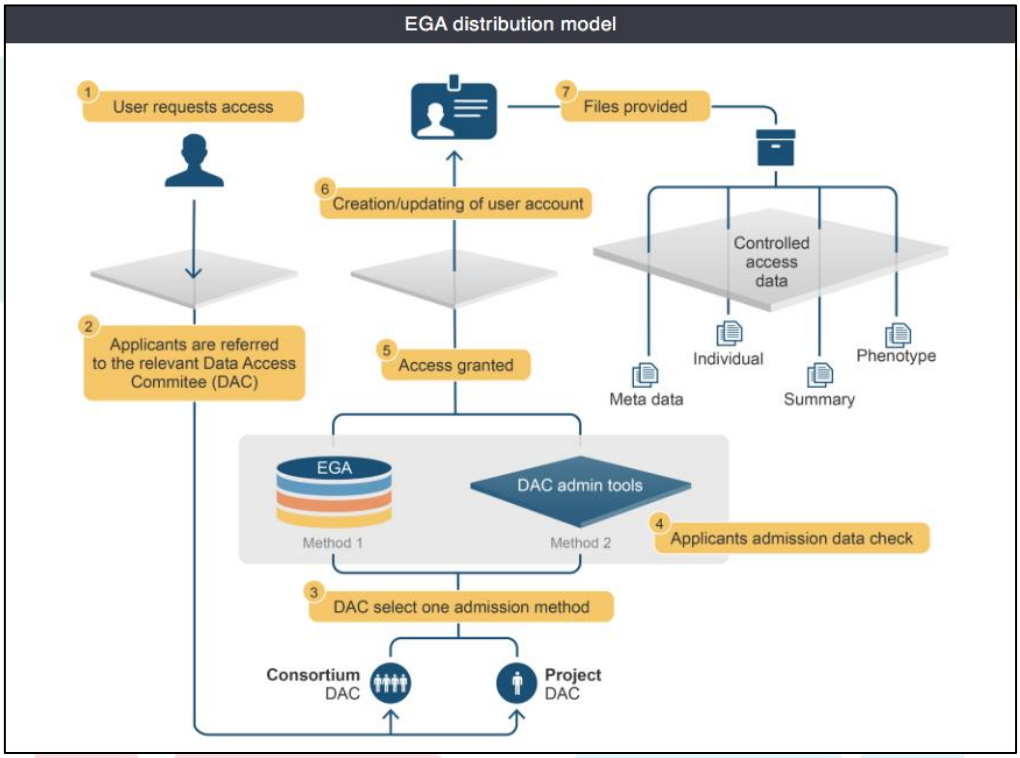

Figure 2 - Users who wish to access EGA data must first apply to the appropriate DAC (step 2), who will then inform EGA to grant access to the required data (step 5)

2.1.2. Features or facilities added in 2019

The EGA added support for submitting metadata via tab-separated-value (TSV) files along with the raw data files via Aspera or FTP. The advantage here is that the submitter can upload a file manifest, which includes the required metadata, along with the raw data files directly to their FTP/Aspera box, which for larger submitters means the submission is effectively done via FTP/Aspera, as opposed to FTP/Aspera and REST API/submission portal. The addition of this facility to EGA will save a considerable amount of time for large submitters as most metadata is usually collated in TSV/Excel format and will therefore negate the necessity to convert these into XML and submit using a different method to the raw file upload.
In addition to this, EGA has adopted the use of the Data Use Ontology (DUO[7]) - a Global Alliance for Genomics and Health (GA4GH) standard which tags datasets with

6 https://github.com/EGA-archive/ega-download-client
7 https://github.com/EBISPOT/DUO

the conditions applied to data use in a machine-readable format. It is anticipated that the adoption of this standard will reduce data access committee (DAC) burden by reducing the number of inappropriate applications, stream-lining the application process itself and will also improve data re-use as users will be able to identify more datasets that may be of use to their research by searching on DUO terms.
The EGA has also incorporated QC reports into the website[8] which means the users can now browse and review the associated statistics of files that have been uploaded to the EGA - including details such as base coverage, base quality, inferred assembly and read length.

2.1.3. Plans for improvement in 2020

During 2020, EGA will:

- Collaborate with both PRIDE (EBI) and MetaboLights (EBI) to enable the submission and linkage of non-DNA omics data requiring controlled access.
- Add CRAM support for the GA4GH standard htsget[9] streaming service, which will allow byte ranges or genomic regions to be specified for streaming, allowing on-the-fly interactive analysis of raw data supporting genetic variations
- Support GA4GH Passport[10] and Visas to allow interoperability with the ELIXIR AAI and Life Science AAI
- Continue with work to become a federated resource of interoperable services that will enable genomic and biomolecular data on a population scale to be available across international boundaries, enabling science and improving the health of people - thereby also supporting the RD community

2.2. RD-Connect GPAP

Contributors: Sergi Beltran (CNAG-CRG)

2.2.1. 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 and IRDiRC recognized resource hosted at the CNAG-CRG.

De-identified phenotypic data is collected using HPO, ORDO and OMIM ontologies through custom templates implemented in a dedicated PhenoTips instance. 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 prioritization and interpretation of genomic variants at RD-

8 https://ega-archive.org
9 http://samtools.github.io/hts-specs/htsget.html
10 https://github.com/ga4gh-duri/ga4gh-duri.github.io/blob/master/researcher_ids/ga4gh_passport_v1.md

Connect GPAP. Raw genomic data is deposited at the EGA for long-term archive and controlled access.

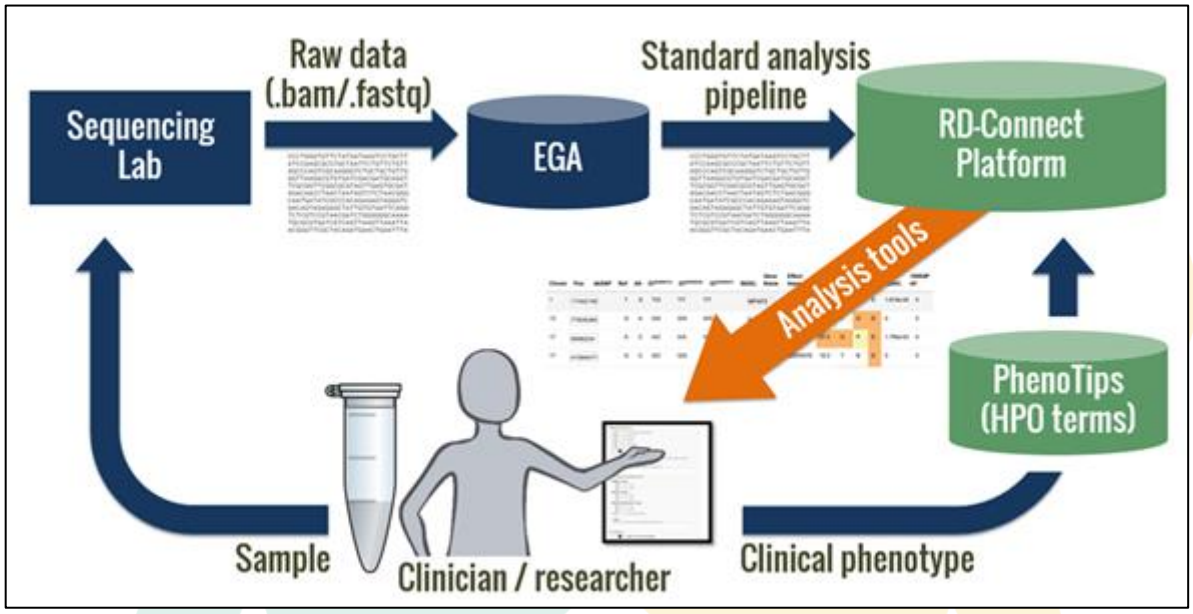

Figure 3 - Data flow for RD-Connect GPAP

2.2.2. Features or facilities added in 2019

The process of changing the User Management System from Central Authentication Service (CAS) to Keycloack has being started. This will facilitate changing the authentication protocol to OpenID Connect Protocol, and it is a step required to federate RD-Connect GPAP authentication service to the LifeScience Authentication and Authorization Infrastructure (AAI) in the future as devised in Task 11.3. In addition, the Keycloak user interface to adapt it to the RD-Connect GPAP Helpdesk requirements was modified in order to manage RD-Connect GPAP users' data in the near future. Prior to integrating Keycloak in the production system, it has been integrated in the RD-Connect GPAP Playground.

As part of the continuous improvements, the evaluation of the usage of Genomics England templates for phenotypic data entry, to which migration in the coming year might be considered, was performed. Furthermore, the development of a new phenotype data collation system was started to facilitate phenotypic data submission and integration with the RD-Connect GPAP.

A continuous work on RD-Connect GPAP maintenance and improvement was performed. Changes that improved metadata loading velocity during submission were implemented, and additional data visualization fields, that helped RD-Connect GPAP users to better follow their participant/experiment status in the platform, were introduced. In addition, the default submission options in different fields ("Project", "ERNs", "Tissue" or "kits") were updated following actual user necessities. Regarding genomic data upload, improvements to the Redlris service were implemented; the

generation and provision of Redlris credentials for users to use the Aspera protocol were automated as well as the verification and download procedure of the submitted sequencing files.

A new Use Cases (examples) in the RD-Connect GPAP Playground was introduced to allow users to test more features and better understand the platform.

Together with the European Genome-Phenome Archive (EGA), a new procedure to automate a secure transfer of sequencing files to the EGA is being created.

2.2.3. Plans for improvement in 2020

A Keycloak will be integrated as the User Management system in the main RD-Connect GPAP. With OpenID Connect supported, the authentication service will be federated to the LifeScience AAI if prioritized by the corresponding subtask in T11.3.

RD-Connect GPAP user submission workflow will be eased and improved by integrating the submission of Phenotypic and Genomic metadata, thanks to the integration of Keycloak and the release of a new system for collating phenotypic information.

In addition, changes in the RD-Connect GPAP front-end that will clarify and facilitate user's access and analysis of data will be introduced.

The platform's "Data Management" front-end framework will be changed from Django to React, which will allow the improvement of the website time-response at a user level, and the issue time-response at a developer level.

A new procedure to transfer, in an ease and secure way, sequencing files from RD-Connect to the EGA will be implemented.

New Guidelines and Documentation related to the changes that will be implemented in the RD-Connect GPAP platform will be generated.

2.3. DECIPHER

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

2.3.1. Resource data flow

DECIPHER[11] is a web platform that helps clinical and research teams to assess the pathogenicity of variants and to share rare disease patient records. Patient genotype and phenotype data is uploaded by academic clinical genetic centres worldwide, using the web interface, via bulk upload or through a deposition API. The DECIPHER web interface provides a suite of tools to assist users in assessing the pathogenicity of variants. Registered DECIPHER users at the depositing centre annotate the variants using the tools provided in DECIPHER. With explicit patient consent, the patient record is shared openly through the web portal. DECIPHER also supports the sharing of patient data between defined clinical genetic centres (consortium sharing).

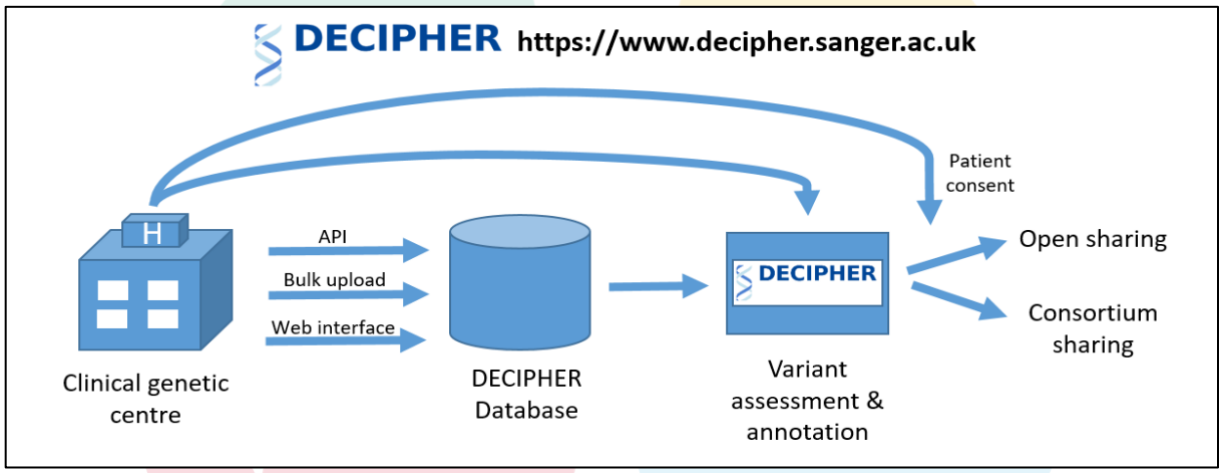

Figure 4 - Decipher Data Flow

2.3.2. Features or facilities added in 2019

During 2019 DECIPHER has continued to support the sharing of rare disease patient records worldwide. DECIPHER provides a module to support sequence variant classification according to ACMG pathogenicity evidence and has included additional information to support this framework from the ClinGen Sequence Variant Interpretation Group and the Association for Clinical Genomic Science. DECIPHER is a founding member of the Matchmaker Exchange and during 2019 has connected with RD-Connect. This year there has been a major restructuring of the DECIPHER database and codebase to ensure that the infrastructure fully facilitates the development of new functionality and features. The new version, version 10, will bring a plethora of new features in early 2020.

[11] https://decipher.sanger.ac.uk/

2.3.3. Plans for improvement in 2020

DECIPHER will release version 10 in early 2020. Online DECIPHER user guidelines and documentation for data deposition will be updated with this release. New features in version 10 include the broadening of the types of genetic variation shared using the DECIPHER platform from sequence variants and copy number variants, to all types of genetic variation (e.g. aneuploidy, inversions, uniparental disomy, insertions, short tandem repeats). The grouping of variants, such as compound heterozygous variants, will also be supported. During 2020, DECIPHER also plans to incorporate further predictive pathogenicity scores, improve splice annotation and incorporate additional information and modules to further support recommendations for ACMG pathogenicity evidence criteria.

2.4. JRC-ERDRI

Contributors: Simona Martin (EC-JRC)

The JRC's work and contribution to this thematic area are contained in JRC documents. They are communicated to the EJP-RD via presentations in EJP-RD meetings

2.5. RD-Connect Registry and Biobank Finder

Contributors: David van Enckevort (UMCG), Mary Wang (FTELE), Heimo Müller (BBMRI-MUG)

2.5.1. 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[12]. 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 Gainotti et. al., 2018).

12 http://www.orphadata.org/cgi-bin/index.php

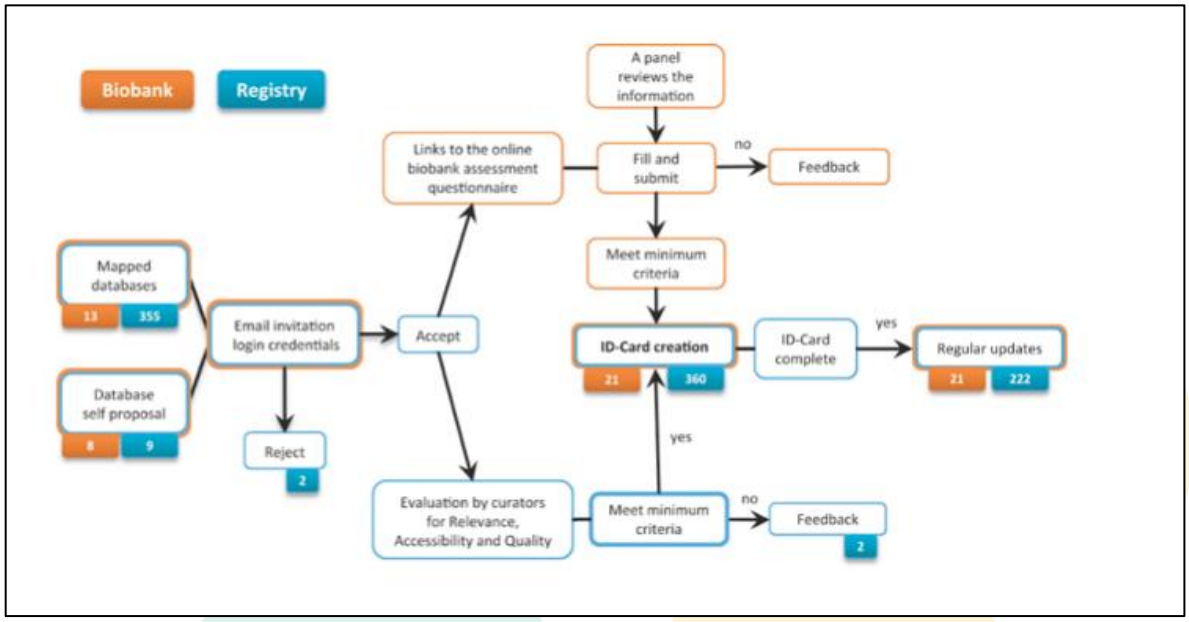

Figure 5 - Inclusion of biobanks and registries in the Registry and Biobank Finder. The process of inclusion and evaluation of biobanks and registries in the Registry and Biobank Finder (mapped and self-proposed)

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

2.5.2. Features or facilities added in 2019

The BBMRI-ERIC Negotiator was extended to support multiple biobank networks and catalogues and added support to initiate a sample through the RD-Connect Registry and Biobank Finder in the BBMRI-ERIC Negotiator. The migration the RD-Connect Registry and Biobank Finder to the BBMRI-ERIC Directory in technical aspects (from Liferay to Molgenis), data transfer (through APIs) and support of curator features was planned. Furthermore, maintenance and minor bug fixes were done in 2019.

2.5.3. Plans for improvement in 2020

The RD-Connect Registry and Biobank Finder will be migrated to the BBMRI-ERIC Directory. With this migration a) the technical platform to Molgenis will be consolidated (the support of Liferay as basic framework is out of scope in EJP-RD) and b) a one-stop-shop and user experience for all stakeholders, as biobankers & data providers, curators and researchers, searching for samples, will be provided. With this migration step also a better integration of the "RD-Connect Registry and Biobank Finder" and the RD-Connect Sample Catalogue will be possible.

2.6. RD-Connect Sample Catalogue

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

2.6.1. 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.

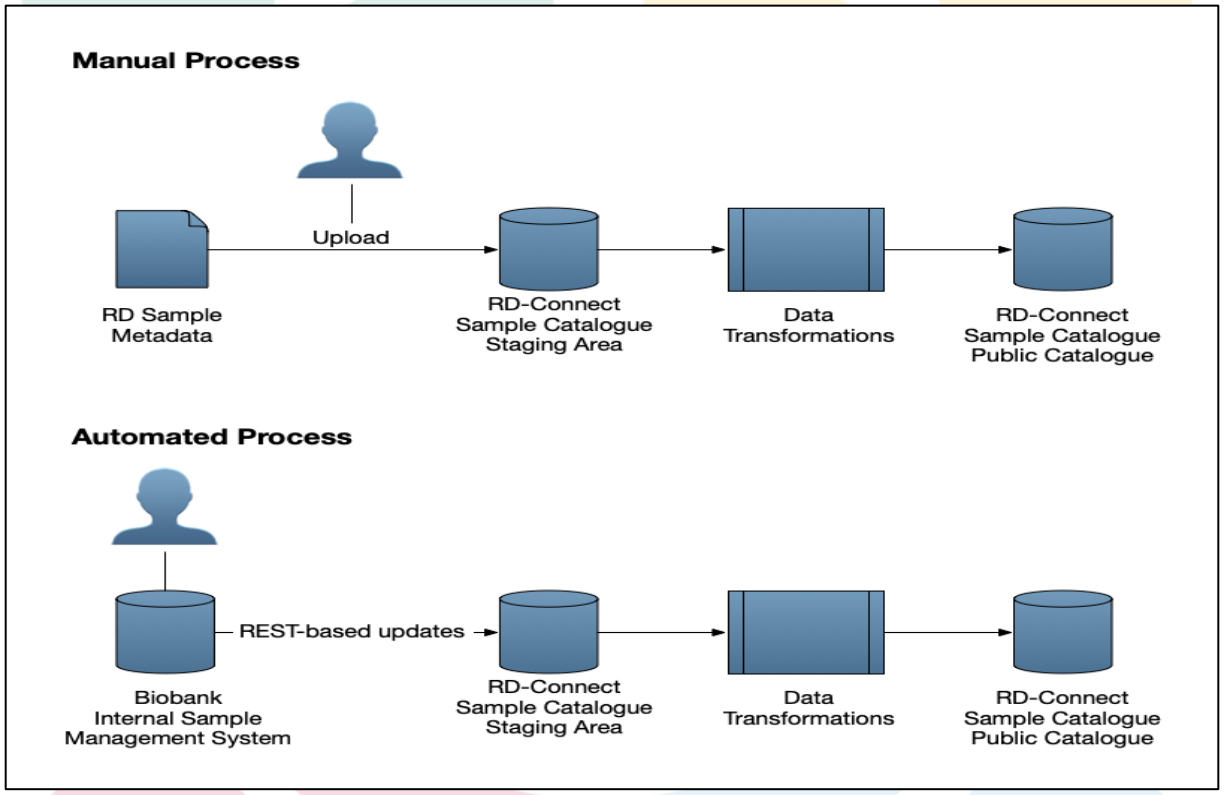

Figure 6 - RD Connect Sample Catalogue Data Flow

2.6.1.1. 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.

2.6.1.2. 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.

2.6.2. Features or facilities added in 2019

In 2019 we updated the Sample Catalogue to run on MOLGENIS 7.4.8[13] from MOLGENIS 5.2, which added to the following features compared to the previously deployed version:

- Navigator: Create, Edit, Move, Copy, Upload and Download resources.
- Functions to submit mapping and script jobs and return the execution href
- A new Captcha (recaptcha)
- OpenID Connect authentication
- Security Manager to manage roles and permissions of users
- FAIR Data Point
- Numerous stability improvements and bug fixes
- As part of the Hackathon in Paris in July 2019 a PoC Adaptor to the Virtual Platform[14] based on the Metadata schemas defined at that event was developed.

Based on feedback from the users, the user interface of the catalogue to improve the search experience of the users and to present the sample details in an easier to read table was fine tuned.

2.6.3. Plans for improvement in 2020

It is planned to upgrade the Sample Catalogue to MOLGENIS 8.x, which adds improved security to the system and introduces the MOLGENIS App Store as a mechanism to deploy a customized front end application to give users a better user experience and support for requesting samples through the BBMRI-ERIC Negotiator. Next to that the FAIR Data Point to the latest version will be updated and the system with the LS-AAI connected.

13 https://github.com/molgenis/molgenis/releases
14 https://github.com/ejp-rd-vp/rd-connect-sample-catalogue-adaptor

2.7. BBMRI-ERIC Directory

Contributors: David van Enckevort (UMCG), Petr Holub (BBMRI-ERIC)

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

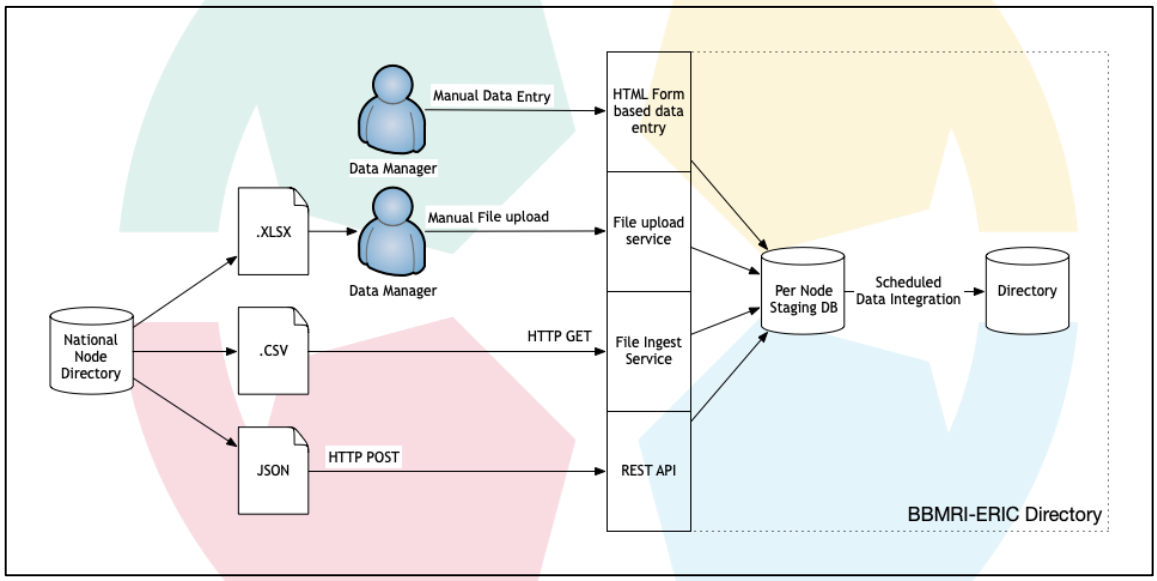

Figure 7 - BBMRI-ERIC Directory data flow

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

2.7.2. Features or facilities added in 2019

The version 5 of the Directory was released, based on MOLGENIS 7.2.14, which provided the following features:

- Navigator: Create, Edit, Move, Copy, Upload and Download resources.
- Functions to submit mapping and script jobs and return the execution href
- A new Captcha (recaptcha)
- OpenID Connect authentication
- Security Manager to manage roles and permissions of users
- FAIR Data Point
- Numerous stability improvements and bug fixes

Specific for the Directory further refinements to the user interface and search capabilities, such as an improved diagnosis search facet and the ability to search on the quality marks issued in the BBMRI Quality Program were implemented. To support the National Nodes, the mapping service and validation of the provided data was improved, and the Single Sign-On login based on the BBMRI-AAI was enabled.

2.7.3. Plans for improvement in 2020

The BBMRI-ERIC call for tender for the CS-IT defines the following priorities for the Directory in 2020:

- Optimization of user experience for various browse and search scenarios.
- Enriching data structures related to availability of data.
- Implementation of Persistent Identifiers Policy.
- Implementation of Bioschemas.
- Integration of the service into EOSC portfolio.
- Integration with upcoming LifeScience AAI.

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

2.8. Rare Disease Cohorts (RaDiCo)

Contributors: Daphne Jaoui (INSERM-RaDiCo)

2.8.1. RaDiCo introduction

RaDiCo is a platform, dedicated to the cohort building, follow up and study.

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 an open source software created by Vanderbilt University. 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[15] website.

It brings an eCRF service and it allows 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 from the Necker Hospital only see medical data from patients followed at the Necker hospital
- **users' role:** Investigator, Clinical Research Monitor, Data manager etc.

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

RaDiCo specifically developed the PIST (Patient Identification System Translator) (1) to generate anonymized codes and (2) to manage separately identifying data and medical data. Briefly, the PIST allows to give access to identifying data only to the authorized medical staff, whereas statisticians and data managers, for example, can

15 https://www.project-redcap.org/software/

see de-identified medical data only. Moreover, each medical centre only has access to the patients managed in this centre (and not to patients managed in other centres of the same cohort program).

2.8.2. Resource data flow

2.8.2.1. 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**, dedicated to the user management,
- **The CGM hosted part**, dedicated to medical and sensible data and to patient identifying data management. It comprises of the following elements:
- **PIST:** Patient Identity System Translator: link between medical data and patient identities,
- **EGCS:** EDC Gateway Controller Service: link between users (healthcare and Clinical research professionals), and patients,
- **REDCap:** Research Electronic Data Capture: application allowing clinical data entry for each patient from the cohort.

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

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

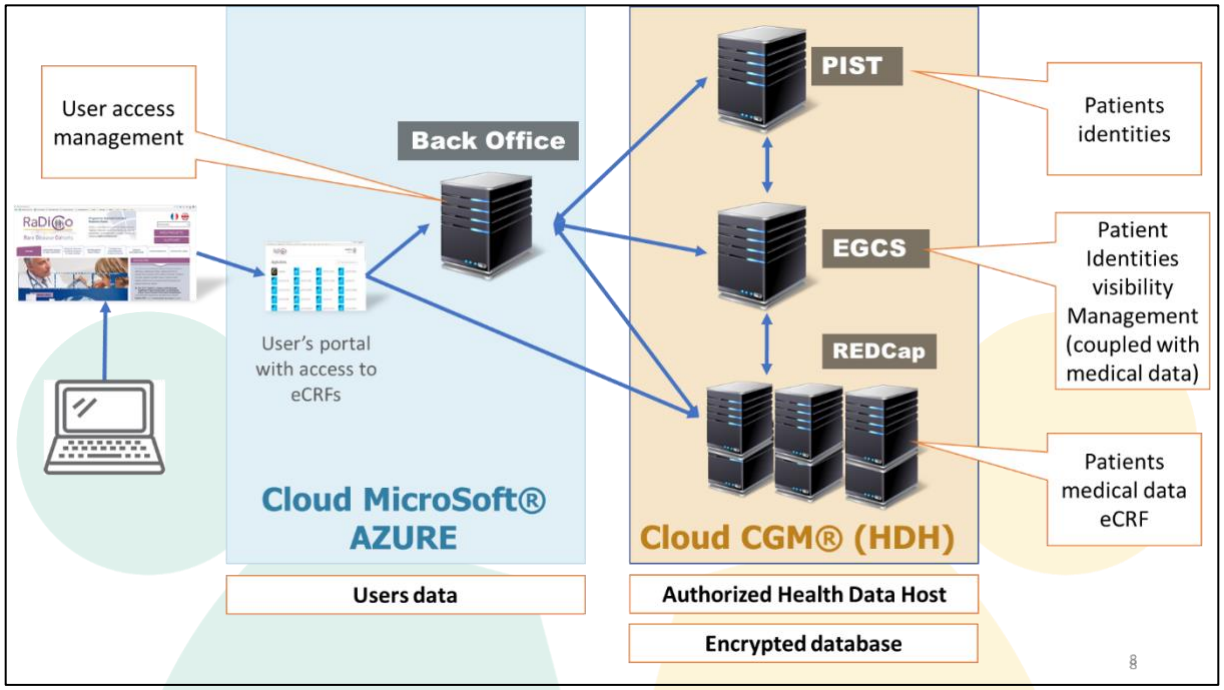

Figure 8 - RaDiCo General schematic description

2.8.2.2. 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 having access to online RaDiCo cohorts has a defined role. Each role has delimited rights and access to medical data.

More generally, RaDiCo IS user's organization respects Attribute Based Access Control (ABAC) principles.

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

2.8.3. Plans for improvement in 2020

For the year 2020, it is planned to add a Business Intelligence platform which will provide resources for statistical analysis, data management and Key Performance Indicators (KPI).

2.9. hPSCreg

Contributors: Nancy Mah (Charité), Andreas Kurtz (Charité)

2.9.1. 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[16]. 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.

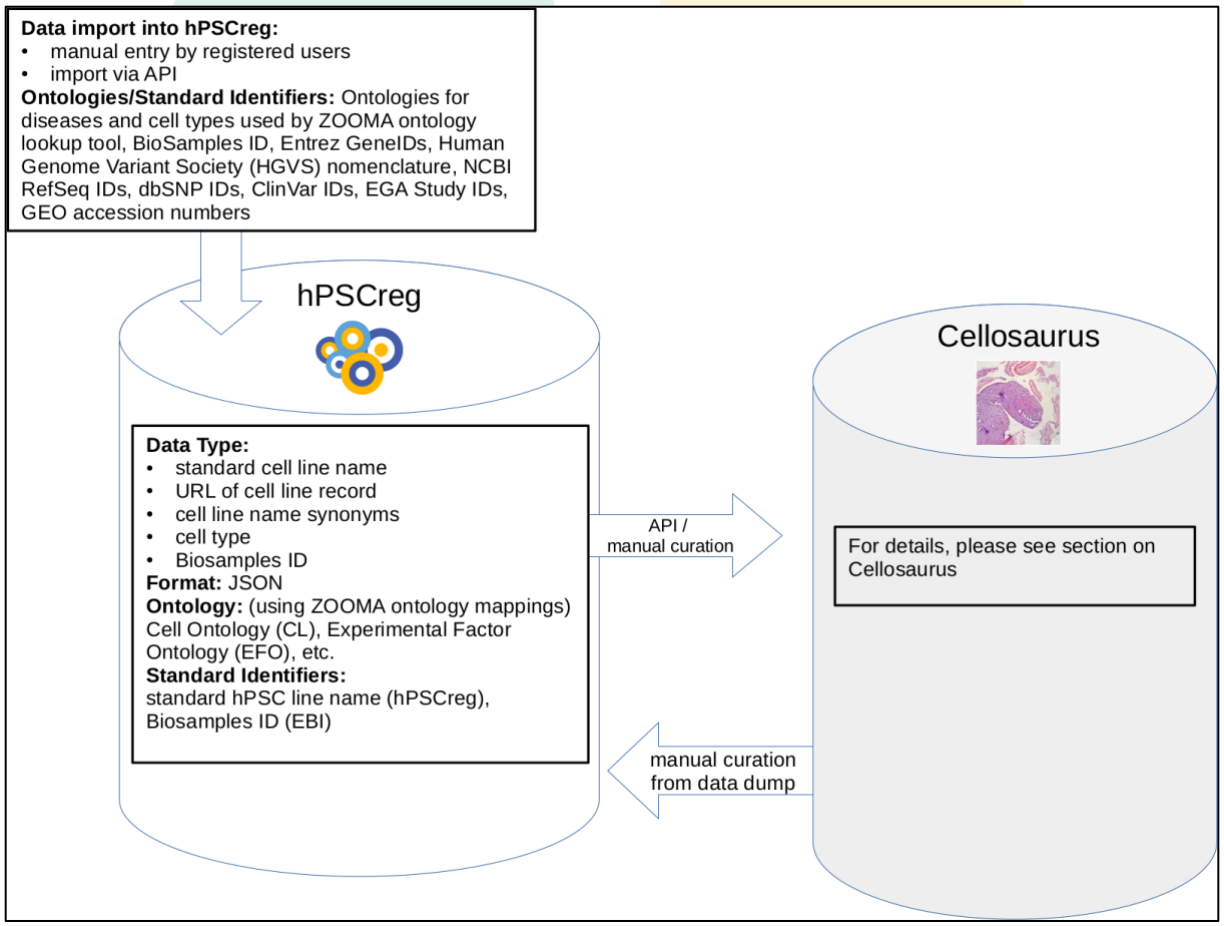

Figure 9 - Data flow for the human Pluripotent Stem Cell registry (hPSCreg).

16 https://hpscreg.eu/

2.9.2. Features or facilities added in 2019

With the initiation of phase 2 of the European Bank for induced Pluripotent Stem Cell lines (EBiSC)[17], the human pluripotent stem cell registry (hPSCreg); https://hpscreg.eu) will continue to store and disseminate cell line data from this EC-funded banking initiative (March 2019). Working together with Kyoto University, the "findability" of hPSCreg cell line data has been increased by becoming integrated into a searchable database holding major collections of hiPSC lines from around the world (http://icscb.stemcellinformatics.org/; May 2019). A clinical study database containing clinical studies that involve human pluripotent stem cells or their derived cell types (June 2019) was also established. Finally, A first version of an API for data import into hPSCreg (Dec 2019) was released.

2.9.3. Plans for improvement in 2020

hPSCreg will implement new data fields to define clinical grade hiPSC lines and will continue to pursue FAIRification on multiple levels to increase the findability of the cell lines to the wider scientific and translational community. The use of RD-specific terms in hPSCreg will be encouraged by up-ranking ORDO terms in the mapping of free text to ontology terms.

2.10. Cellosaurus

Contributors: Amos Bairoch (Cellosaurus)

2.10.1. Resource data flow

Cellosaurus is a manually curated resource. 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[18].

The ontologies used in Cellosaurus are numerous and examples include - for disease terms: 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; STR markers: ANSI/TCC ASN-0002-2011 + additional markers; other in house small "vocabularies": cell line categories, MHC genes, Ig isotypes, genders, etc.

17 https://ebisc.org/
18 https://web.expasy.org/cellosaurus/

2.10.2. Features or facilities added in 2019

- CLASTR: tool to search for similarity to STR profiles
- A new data field reporting the genome ancestry of a cell line was introduced
- The addition of STR markers for mouse cell lines started
- Cross-references were added to: Applied Biological Materials (ABM) cell line products, China Centre for Type Culture Collection (CCTCC), Cancer Dependency Map (DepMap), Foetal Calf Serum-Free Database (FCS-Free), IARC TP53 database, Iranian Biological Research Centre (IBRC) cell line collection, Cancer cell LInes GEne fusions portAl (LiGeA), PharmacoDB integrative pharmacogenomic database and Sanger Cell Model Passport resource.
- All entries with a field which provides information on when a Cellosaurus entry was created, when it was last updated and which version of the entry is currently available were retrofitted.

2.10.3. Plans for improvement in 2020

It is planned to start the process of providing ORDO terms for cell lines that are relevant to RDs during 2020.

2.11. INFRAFRONTIER

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

2.11.1. Resource data flow

The main data resource of INFRAFRONTIER is the EMMA (European Mouse Mutant Archive) database. It holds data more than 7000 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.

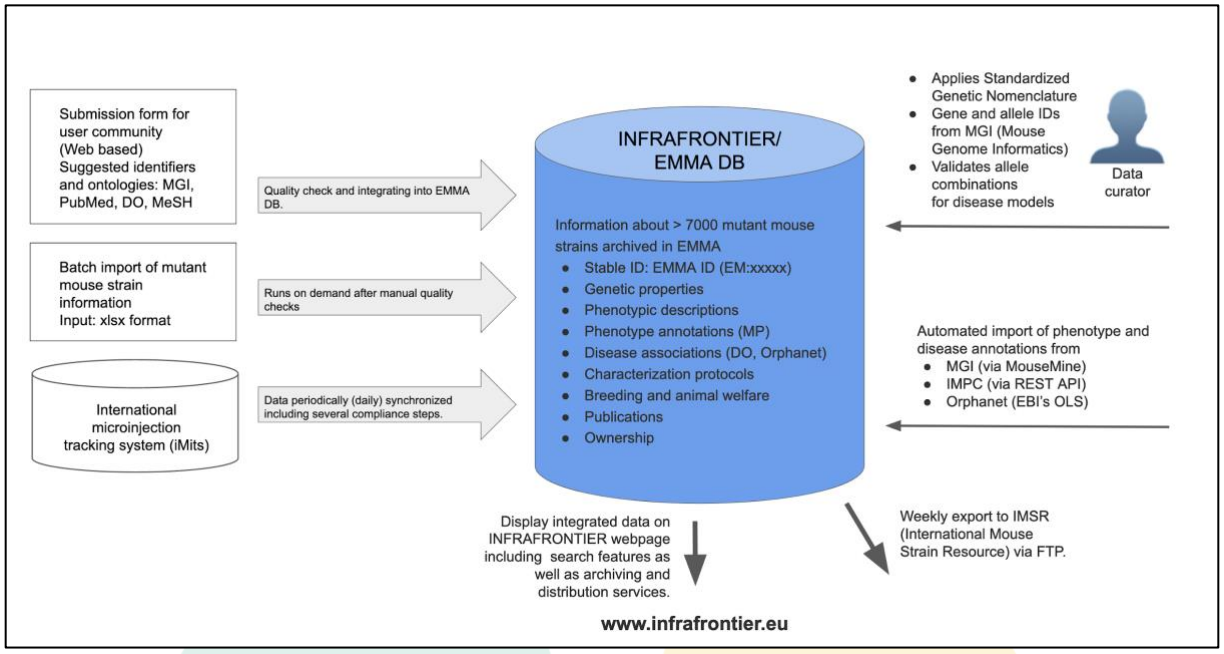

Figure 10 - Data flow of EMMA

2.11.2. Features or facilities added in 2019

To consolidate INFRAFRONTIER's resources and services related to rare diseases in one easily accessible location, a rare disease landing page[19] was set up. It is part of a new menu item "Therapeutic Area" in the knowledgebase on the INFRAFRONTIER portal. On this landing page the global research community can find information about EMMA strains that are related to rare diseases, publications from rare disease related EMMA strains and rare disease conferences involving INRAFRONTIER.

The new section "EMMA Strains and Rare Diseases" shows a list of EMMA strains that are potentially interesting for rare disease researchers. To identify the strains which could be related to a rare disease, the existence of a human ortholog to the mouse gene (Ensembl) was first checked. Then, EBI lookup service for rare diseases which were linked to this gene was searched. Currently the EMMA repository holds > 1516 mouse strains that carry mutations in 879 genes that have been implicated to play a role in rare diseases (1244 different rare diseases).

In addition to the data and web page work, a workshop at ASHG (October 2019, Houston) was organised to better align phenotypic information between human and mouse.

19 https://www.infrafrontier.eu/infrafrontier-and-rare-diseases

2.11.3. Plans for improvement in 2020

While the new dedicated rare disease related pages described above provide a focused view for users with a special interest in rare diseases, INFRAFRONTIER's central mouse strain search page[20] is the one with the second most visits (after the home page). This central search currently only shows Disease Ontology (DO) annotations and has a limited DO based browsing feature. To further increase visibility of rare disease aspects at INFRAFRONTIER, it is planned to integrate rare disease links (via Orphanet IDs) directly on this main EMMA search and to allow searching by rare disease names and IDs there as well.

Towards integration of INFRAFRONTIER data in the EJP RD virtual platform the requirements for an API will be defined.

2.12. PRIDE

Contributors: Juan Antonio Vizcaino (EMBL-EBI)

2.12.1. Resource data flow

The PRIDE database can store all types of mass spectrometry (MS) proteomics datasets, although DDA (shot-gun data dependent acquisition) approaches are better supported. Each dataset must contain at least the raw data (MS data coming out from the mass spectrometers), processed results (identification and optionally quantification) and the required metadata. Other types of files are optional, e.g. search database, spectral libraries, among others). PRIDE is leading and complies with the submission guidelines established by the members of the ProteomeXchange Consortium[21] . Supported file transfer protocols are FTP and Aspera. PRIDE uses different controlled vocabularies for annotation purposes, including the PSI (Proteomics Standard Initiative) MS, BRENDA, Cell Type ontology and NCBI Taxonomy, among others.

20 https://www.infrafrontier.eu/search
21 http://www.proteomexchange.org/

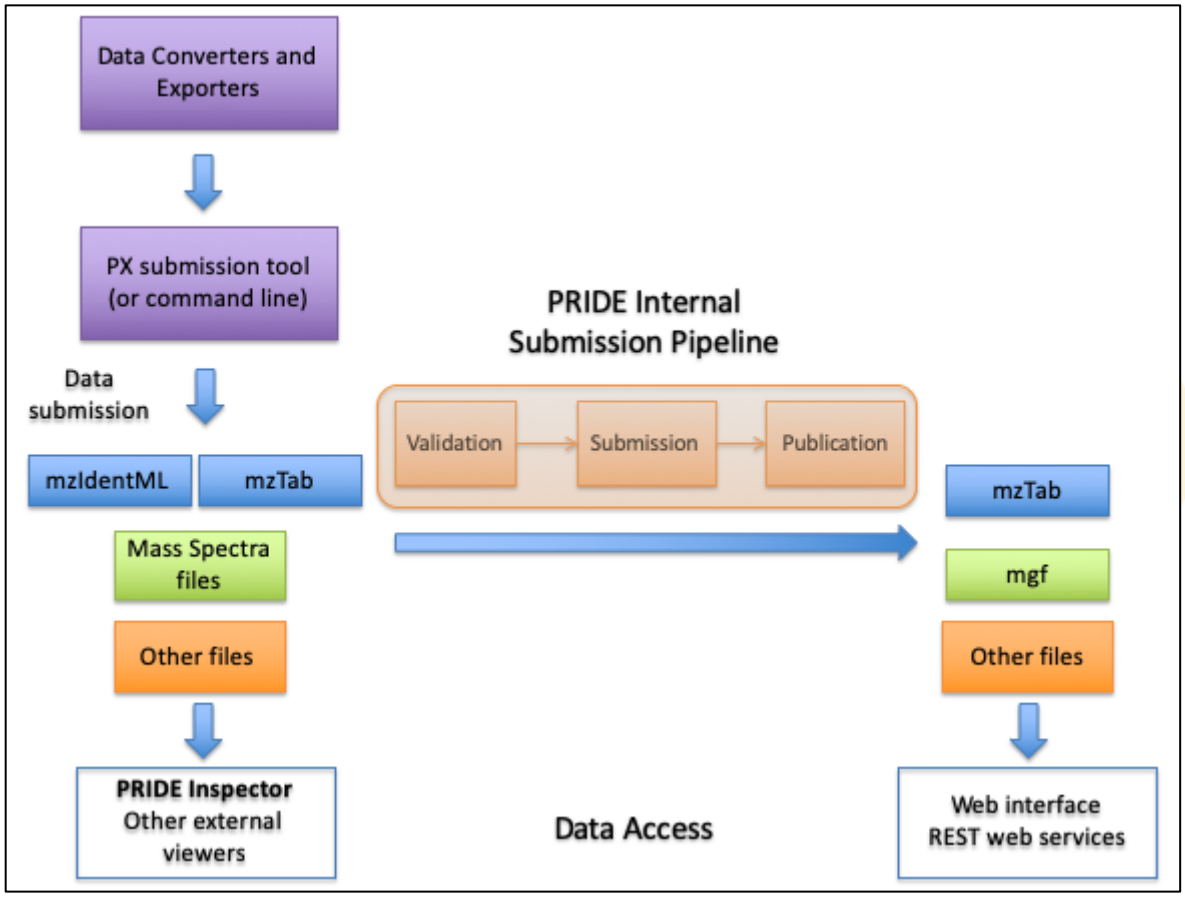

Figure 11 - Data flow for PRIDE

2.12.2. Features or facilities added in 2019

PRIDE infrastructure has been completely redeveloped from scratch, including a new database backend, application programming interface (API) and web interfaces, with new visualisation capabilities and functionality. The release of the new PRIDE system is planned in the coming weeks (at the moment, it is still in beta, http://wwwdev.ebi.ac.uk/pride).

2.12.3. Plans for improvement in 2020

Current plans involve focusing first in improving experimental metadata for PRIDE datasets. Second, QC (Quality Control) charts for datasets will be provided. Third, PRIDE's data dissemination pipelines into other EMBL-EBI resources, mainly with UniProt (PTM data), Expression Atlas (quantitative proteomics data) and Ensembl (proteogenomics data) will be improved.

2.13. MetaboLights

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

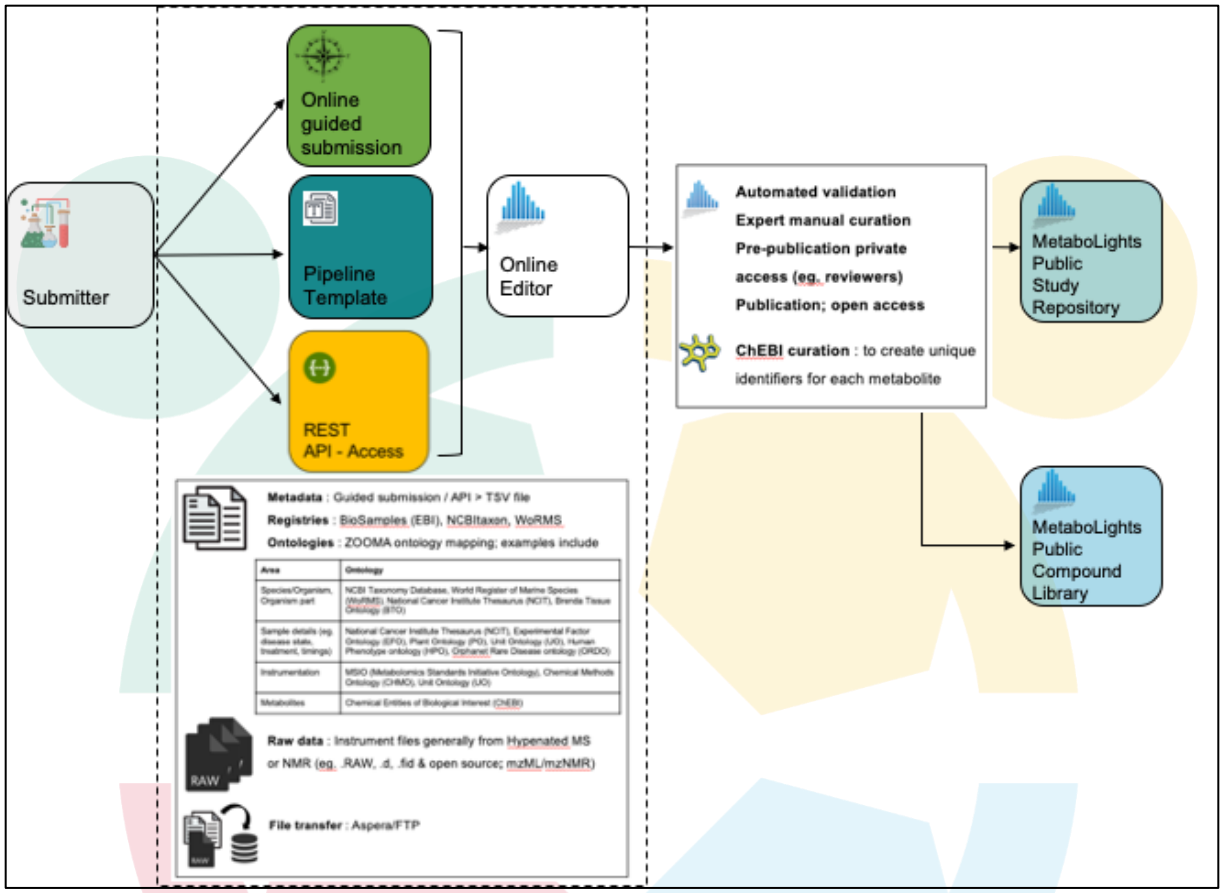

Figure 12 - Data Flow for MetaboLights

2.13.1. Resource data flow

MetaboLights is a data repository for metabolomics data. Each new study submitted creates a unique and persistent identifier. Submitters can choose to use the online guided submission, 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 complete, submitters can change a study status to request curation. Following successful curation, a study is held in private mode and a link is available to share with e.g. journal reviewers until the requested publication date is reached and the study is made publicly available.

MetaboLights also supports a compound library which essentially provides a synopsis of the chemical features (based on ChEBI ontology integration) together with biological references including all study identifiers & associated relevant metadata (e.g. species, disease) per metabolite identified within the repository.

2.13.2. Features or facilities added in 2019

In 2019 the study submission process was overhauled moving from an offline tool to a new streamlined online step by step process with further ontology support. A template version of studies for different centres which are pre- populated with their standard protocols, and have developed MetaboLights API access, was provided as further components aiding submission. In the last quarter a controlled access project which aims to provide further options for e.g. clinical data storage was initiated. This is currently at the proof of concept stage.

2.13.3. Plans for improvement in 2020

The aim in 2020 is to develop the controlled access instance of MetaboLights & make it available to submitters.

Also, as a result of the overhaul of the submission process and the automatic validations, the database content is now more granular and MetaboLights search capabilities will be redeveloped to enable users to query the data more effectively.

For EJ PRD specifically, advice from the other working groups and the use cases work focus will be sought to determine if there are further resources that would be beneficial to incorporate for the rare disease community (e.g. ontologies), and if additional changes to metadata structure (e.g. required fields, instrumentation options) are required.

3. Section 3

3.1. Conclusion

A review of the documentation provided by the resources above, quickly reveals that the submission of, and access to, data is multifaceted using a plethora of different tools - some of which are "in-house" whilst others use common applications. It is clear that moving forwards requires the adoption of existing community standards wherever possible in order to unify the processes by which data is submitted and accessed.

Over the coming year (2020), the information provided by each of the resources above will be used to develop a map of data infrastructures for data deposition and sharing of rare disease data across Europe. This map will go on to form the basis of the documentation and outreach in subtask 11.3.4.

A significant focus in 2020 for task 11.3 will be the upgrade of existing open data deposition resources, such as PRIDE for proteomics, and MetaboLights for metabolomics, to accept controlled access submissions. In parallel, EGA will increase the efficiency of both data deposition and access via developing GA4GH standards, such as htsget. This also includes improving the submission pipelines from rare disease resources such as RD-Connect GPAP.

Connectivity between different resources will also be enhanced via the improvement of interfaces. For example, it is envisaged that queries formulated by different tools with different contexts will be exchanged and answered such as the queries created by EU RD-Platform (ERDRI) will be transferred to the BBMRI-ERIC Negotiator tool to also request access to the biobanks, registries and other resources available in BBMRI-ERIC RI and vice versa.

Over the course of 2020, the Work Focus on "Resources for sharing experimental data and materials" will continue to meet in order to update and discuss the progress of the resources based on this deliverable and continue to identify improvements for data submission and access processes for the RD community.