EJP RD
European Joint Programme on Rare Diseases

H2020-SC1-2018-Single-Stage-RTD
SC1-BHC-04-2018
Rare Disease European Joint Programme Cofund

Grant agreement number 825575

Del 11.12
Second 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 76 – ELIXIR/EMBL-EBI (EGA)

Collaborators: CNAG-CRG, WTSI [DECIPHER] (Associated partner), EC-JRC, UMCG, FTELE, BBMRI-ERIC (including BBMRI-CRS4 & BBMRI-MUG), INSERM-RaDiCo, Charité, SIB [ELIXIR-CH], INFRAFRONTIER.

Due date of deliverable: month 24

Dissemination level:
Public

Table of Contents

Section 1

1.1. Introduction

This deliverable describes the additional facilities and features that were integrated in 2020 into the data deposition and access resources participating in EJP RD.
The objective 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 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 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 is guiding data producers to submit data, making them discoverable through the 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 WP2. To ensure that data is FAIR (Findable, Accessible, Interoperable and Reusable), the resources are deploying or expanding user-friendly interfaces to deposit data and metadata using HPO, ORDO, OMIM and/or any other relevant ontology or standard. In addition, mechanisms are being implemented to guarantee and/or evaluate the quality of the dataset through manual curation, automatic metric generation or a mixture of both. Application programming interfaces (APIs) and graphical user interfaces (GUIs) should allow for query functionality and data access. In order to build community trust, the security of these resources will be assessed and aligned with the recommendations of the Global Alliance for Genomics and Health (GA4GH)[1] by using such examples as the Data Security Toolkit[2] which provides a principled and practical framework for responsible sharing of genomic and health-related data, whilst considering the GDPR and other national legislation. Where relevant, transparency measures and means of monitoring the re-usability of the submitted dataset will be implemented by the resources. The task also includes the implementation and further development of and interoperability of the federated EGA infrastructure with RD-related databases.
Deliverable 11.3 seeks to identify each of the current capabilities of the named resources and additional capabilities added during 2020 in terms of data deposition and access to data, to identify areas for improvement that will benefit the RD community. This output will be used to scope and schedule the Work Focus' (WF) on "Resources for the sharing of experimental data and materials" for the coming year.

[1] https://www.ga4gh.org/

[2] https://www.ga4gh.org/genomic-data-toolkit/data-security-toolkit/

2. Section 2 - Resources

2.1. European Genome-Phenome Archive (EGA)

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

2.1.1. Resource data flow

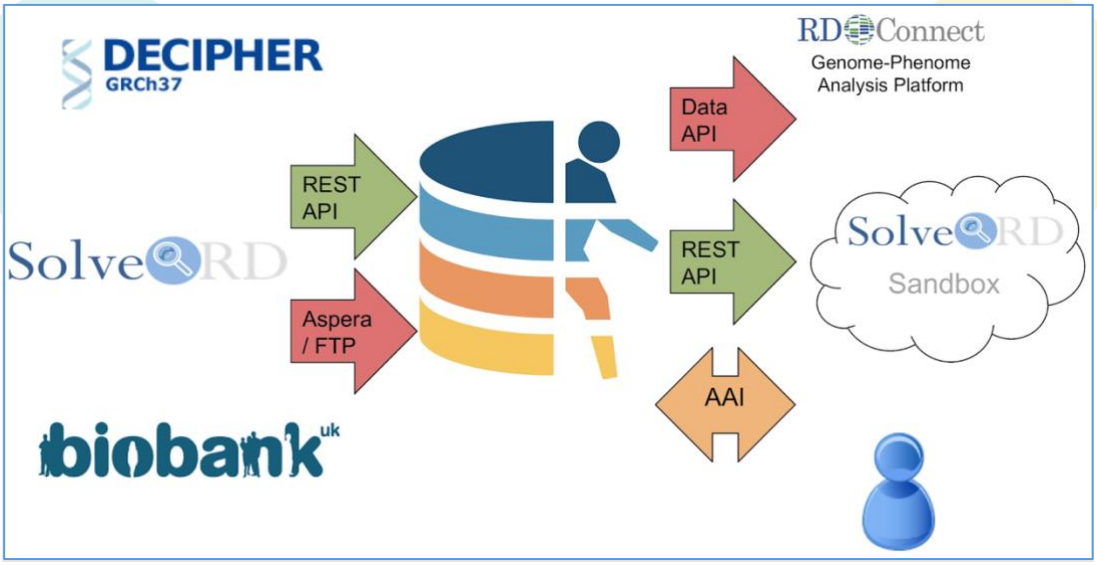


Figure 1. Example data flows to and from EGA. Submitters, such as Solve-RD or DECIPHER, submit data to the EGA for archival and distribution. These data can then be distributed via the EGA Data API to authorised users. Use-cases include distributing data to the RD-Connect Genome-Phenome Analysis Platform,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

[3] https://ega-archive.org/submission/quickguide
[4] https://ega-archive.org/submission/tools/submitter-portal

[5] https://ega-archive.org/submission/programmatic-submissions

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. 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.

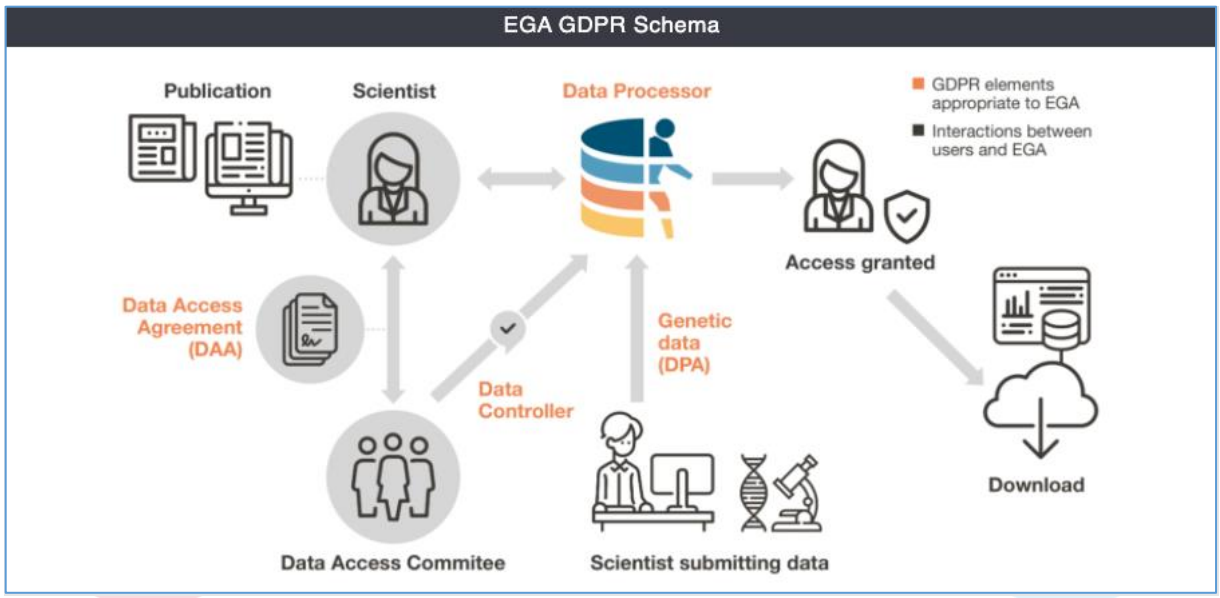


Figure 2. EGA in the context of GDPR. Users who wish to access EGA data must first apply to the appropriate Data Access Committee (Data Controller), who will then inform EGA (Data Processor) to grant access to the required data.

2.1.2. Features or facilities added in 2020

In 2020, EGA made progress on many fronts towards supporting improved data access.

CRAM support for the GA4GH htsget[6] streaming service had been added to the EGA Data API to enable specifying byte ranges or genomic regions of CRAM files for streaming. This functionality is currently available in an EGA test environment and will be released to production in 2021. In parallel, the EGA virtual filesystem (FUSE layer) was put into production to support on-the-fly data access, visualisation, and interactive analysis of raw data. The FUSE layer, which uses the EGA Data API, has been integrated with the RD-Connect Genome Phenome Analysis Platform (GPAP) to allow researchers to log in and visualise BAM files in EGA using the Integrative Genomics Viewer.

EGA now supports GA4GH Passports[7] and Visas to allow interoperability with the ELIXIR AAI and Life Science AAI. This will ensure compatibility with the LifeScience AAI in future and will allow single sign-on between the GPAP and EGA when the GPAP supports the LifeScience AAI.

EGA collaborated with PRIDE and MetaboLights to provide technical and policy guidance for managing submission and access of human controlled proteomic and metabolomics data, respectively. This effort supports the increasing number of clinical proteomics and metabolomics datasets being submitted to these resources, and in particular the increasing number of rare disease studies being submitted to MetaboLights.

EGA has continued working to become a federated resource of interoperable services enabling genomic and biomolecular data on a population scale to be available across international boundaries. Much of the large cohort omics data is being generated by national or regional healthcare initiatives, and many European countries now have nascent personalised medicine programmes. Thus, the EGA is positioned to support rare disease studies directly or through coordination with national or regional initiatives where sensitive data cannot cross international borders.

2.1.3. Plans for improvement in 2021

In 2021, EGA will work to improve efficiency and quality of data deposition, while continuing efforts to improve data access.
EGA plans to develop and integrate metadata validation functionality with the existing submission service to assist users in providing FAIR data. For example, EGA can make it easier for submitters to map and validate free text against existing disease ontologies (e.g., Orphanet Rare Disease Ontology[8], Mondo Disease Ontology[9]) to ensure rare disease datasets are accurately tagged with standard terms and easily findable in EGA.
EGA is planning improvements to the underlying EGA data model to make it more straightforward for submitters to tag controlled access metadata separately from

[6] http://samtools.github.io/hts-specs/htsget.html
[7] https://github.com/ga4gh-duri/ga4gh-duri.github.io/blob/master/researcher_ids/ga4gh_passport_v1.md
[8] https://www.ebi.ac.uk/ols/ontologies/ordo
[9] https://www.ebi.ac.uk/ols/ontologies/mondo

public metadata. Allowing more fine-grained control of which metadata can be made publicly searchable is especially important for rare disease studies where less metadata is needed to potentially re-identify an individual. In addition, EGA plans to improve interoperability with GA4GH Phenopacket[10] standard format for representing phenotypic data. In 2020, Solve-RD piloted the use of Phenopackets for submission of phenotypic data to EGA, and it is planned to build upon this work, including expanding to other rare disease use cases, in 2021.
User-facing documentation and internal SOPs will be updated to support use of new metadata validation functionality and improvements to the EGA data model.
Building upon the success of integrating a FUSE-layer with the RD-Connect GPAP, EGA plans to improve the capabilities of this technology through iterative feedback with Solve-RD to support emerging use cases. Work will also be performed with RD-Connect to establish a single sign-on between the GPAP and EGA when the GPAP supports the LifeScience AAI. Taken together, these efforts will contribute to making genomics-phenomics rare disease data more easily accessible and interoperable for researchers within the rare disease research data ecosystem EJP RD is building.

2.2. RD-Connect GPAP

Contributors: Sergi Beltran (CNAG-CRG), Carles Garcia (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 an IRDiRC[11] recognized resource hosted at the CNAG-CRG.
De-identified phenotypic data is collected using HPO, ORDO and OMIM ontologies through custom templates implemented through the RD-Connect GPAP-Phenostore module. Pseudonymized experiment data (exomes and genomes) and metadata are collected in the RD-Connect GPAP and processed using a standardized analysis and annotation pipeline. Integrated genome-phenome results are made available to authorized users for prioritization and interpretation of genomic variants at RD-Connect GPAP. Raw genomic data is deposited at the EGA for long-term archive and controlled access.

[10] http://phenopackets.org/

[11] https://irdirc.org/

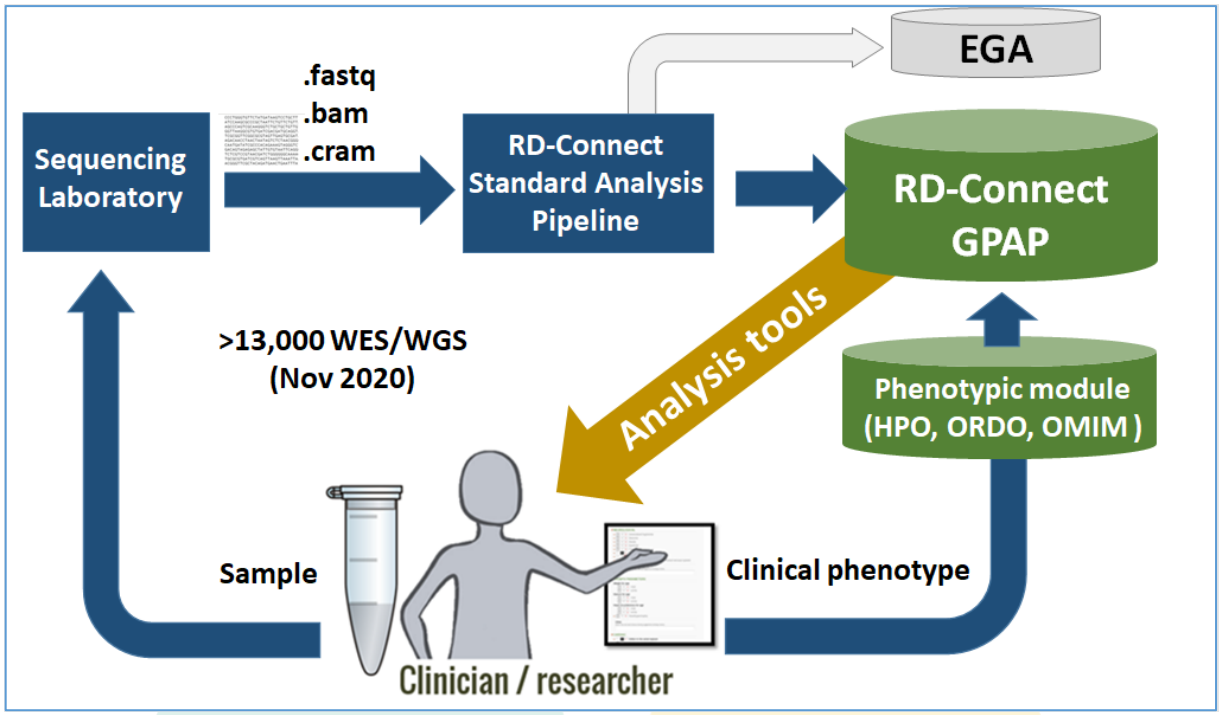


Figure 3. Data flow for RD-Connect GPAP.

2.2.2. Features or facilities added in 2020

The implementation of Keycloak as the main User Management System for all RD-Connect GPAP modules has been finalised. It will allow in the future to federate RD-Connect GPAP authentication service to the Lifescience Authentication and Authorization Infrastructure (AAI).

The new RD-Connect GPAP-Phenostore module was released (accessible through the Playground at *https://playground.rd-connect.eu/*); it is used to collate standardized phenotypic data. This development will improve usability, it will clarify data workflows for users, it will enhance integration of the phenotypic data with the genomic data in the platform, and it will allow us to add ad-hoc functionalities that will be better integrated in the RD-Connect GPAP ecosystem. In addition, it will reduce the technology stack as the provision of a similar interface design and user-experience across the different modules of the GPAP is aimed. During 2021, the introduction of new improvements to the RD-Connect GPAP-Phenostore module will be performed to adapt it to the new developments and requirements necessary for the rare disease community.

New User-guidelines and Video-tutorials have been generated for the RD-Connect GPAP-Phenostore module (which are accessible through the Homepage *https://platform.rd-connect.eu/*).

New disease-specific clinical submission forms were introduced (see figure 4 below with an example of 3 disease-specific forms) to facilitate and improve the collation of phenotypic data during submission. These forms are aligned with Genomics England data models (https://www.genomicsengland.co.uk/?wpdmdl=5500) and provide a list of specific signs and symptoms to accurately describe the participant's diseases.

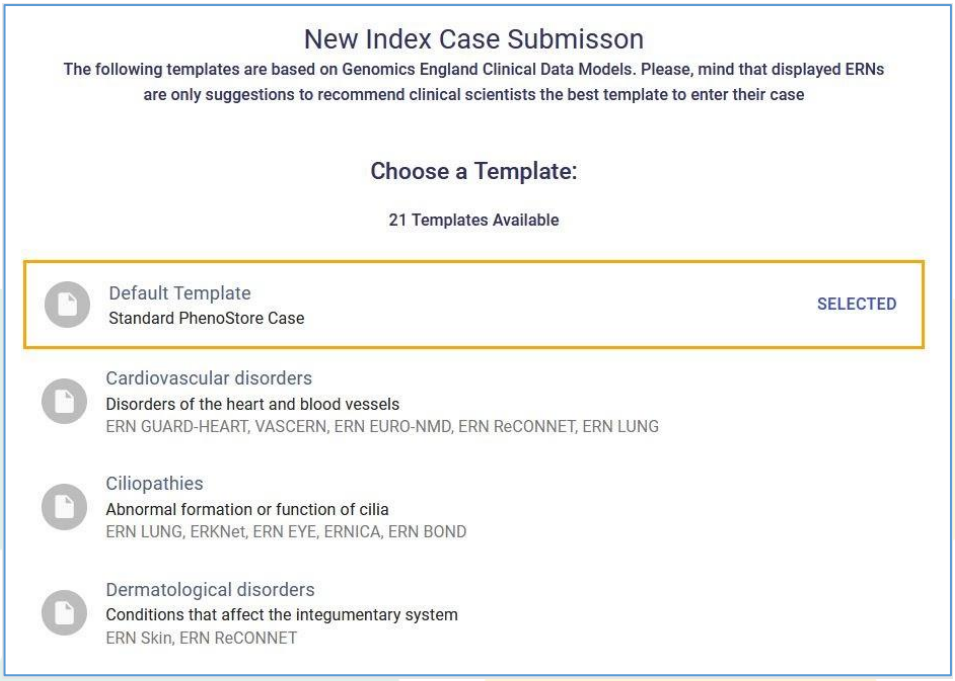


Figure 4. Screenshot showing 4 of the 21 available disease specific forms.

The front-end framework of the "Data Management" module was adapted from Django to React. This has reduced time-response for user's actions, and therefore it has improved the whole user experience.

The RD-Connect GPAP Welcome page was changed at https://platform.rd-connect.eu/ to increase clarity.

A new API was developed to improve the automatic generation of passwords for the Aspera data upload service provided by RedIRIS (the Spanish academic and research network). This has simplified the data submission workflow, both for the users and the RD-Connect GPAP.

2.2.3. Plans for improvement in 2021

The work on connecting the RD-Connect GPAP authorization system with the ELIXIR / LifeScience AAI[12] will be performed if prioritized by the corresponding subtask in T11.3. It will allow using the unified authentication system that will be implemented for EJP RD resources.

RD-Connect GPAP Data submission module will be renewed with a new front-end, which will integrate its appearance and internal structure with the newly developed RD-Connect GPAP-Phenostore module and will clarify data submission steps for users.

New user-guidelines will be generated for the renewal of the Data Submission module.

[12] https://elixir-europe.org/services/compute/aai

Working with INFRAFRONTIER to connect both platforms through an API will allow users from RD-connect GPAP to query candidate genes to gather information from the mouse models database.

The work that will be performed with hPSCreg to connect both platforms through an API will allow users from GPAP to discover relevant cell lines in hPSCreg from the data they will be analysing.

With EGA, working to link identifiers from both platforms will facilitate data submission to RD-Connect GPAP through the EGA.

A pilot project to use Privacy-Preserving Record Linkage (PPRL) using EUPID[13] services will be developed, which will allow testing these new technologies between RD-Connect GPAP and the Sample Catalogue.

2.3. DECIPHER

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

2.3.1. Resource data flow

DECIPHER[14] is a web platform that helps clinical and research teams to assess the pathogenicity of variants and to share rare disease patient records. Patient genotype and phenotype data is uploaded by academic clinical genetic centres worldwide, using the web interface, via bulk upload or through a deposition API. The DECIPHER web interface provides a suite of tools to assist users in assessing the pathogenicity of variants. Registered DECIPHER users at the depositing centre annotate the variants using the tools provided in DECIPHER. With explicit patient consent, the patient record is shared openly through the web portal. DECIPHER also supports the sharing of patient data between defined clinical genetic centres (consortium sharing)

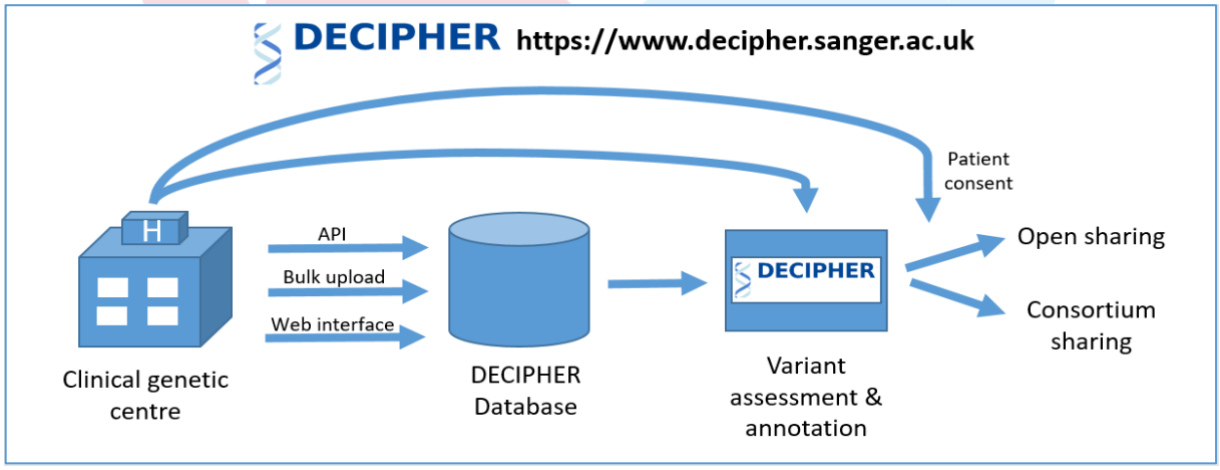


Figure 5. DECIPHER Data Flow.

[13] https://eupid.eu/#/home
[14] https://decipher.sanger.ac.uk/

2.3.2. Features or facilities added in 2020

During 2020 a major new version of DECIPHER was released. 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). DECIPHER also introduced the grouping of variants to allow the representation of, for example, compound heterozygous variants and rare pathogenic haplotypes. Online DECIPHER user guidelines, video tutorials and documentation for data deposition were updated with the release of the major new version. In addition, further predictive pathogenicity scores (CADD) and improved splice annotation (spliceAI) have been incorporated into the platform.

2.3.3. Plans for improvement in 2021

DECIPHER currently visualises genomic data in GRCh37 and is working towards visualizing the data in GRCh38. Visualising the data in GRCh38 will ensure that DECIPHER is using the most up-to-date gene build information and transcripts. DECIPHER will continue to support deposition in GRCh37. Using GRCh38 will permit DECIPHER to encourage the use of MANE Select transcripts (https://www.ncbi.nlm.nih.gov/refseq/MANE), a high-quality transcript which is 100% identical between Ensembl and Refseq. Online user guides and video tutorials will be updated throughout 2021 as new features are added to the website.

2.4. Joint Research Centre - European Rare Disease Registry Infrastructure (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: Esther van Enckevort (UMCG), Mary Wang (FTELE), Heimo Müller (BBMRI), Vittoro Meloni (BBMRI-CRS4), Robert Reihs (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[15]. 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

[15] http://www.orphadata.org/cgi-bin/index.php

information about the institute, the disease focus, available data and/or samples and related documents such as SOPs and Consent forms through an online questionnaire. All registries and biobanks are assessed by a panel and if they meet the minimal requirements for inclusion an ID-Card is created (See workflow from Gianotti, et. al., 2018).

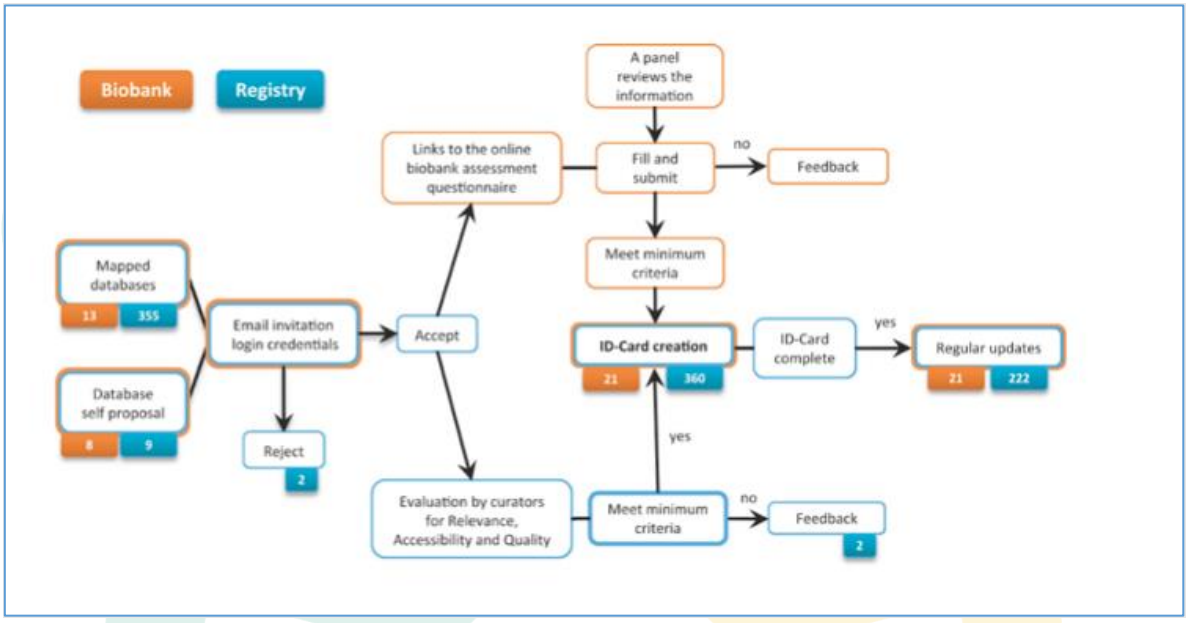


Figure 6. Inclusion of biobanks and registries in the Registry and Biobank Finder. The process of inclusion and evaluation of biobanks and registries in the Registry and Biobank Finder (mapped and self-proposed).
Gainotti, S., Torreri, P., Wang, C. M., Reihs, R., Mueller, H., Heslop, E., … Taruscio, 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 2020

Data from the Registry finder was converted into MOLGENIS[16] format so the data can be made seamlessly browsed with Sample Catalogue and Directory. In addition, quite a few technical improvements were made preparing for 2021. In particular (a) the user interface framework was reimplemented for data browsing in MOLGENIS allowing to more easily create new user interface that shows registry finder, sample catalogue and directory contents together; (b) a shopping cart system was implemented to enable individual samples to be selected with the aim to then directly send to negotiator. In addition, pilots were developed with Query Builder Work Focus to enable federated querying (joint effort with Sample Catalogue & Directory).

2.5.3. Plans for improvement in 2021

Deploy new data explorer user interface that replicates current Liferay[17] based browse experience on top of MOLGENIS. Migrate and align the RD-Connect Registry and Biobank Finder data model to the Molgenis EMX format to ease integrated search

[16] https://www.molgenis.org/
[17] https://www.liferay.com/

across biobanks, registries and biobanks. Evaluate the extension of the BBMRI-ERIC data model with a star schema to characterise collections which would enable fine-grained count searches to find how many relevant samples exist within a collection. Benefit from the merge with sample catalogue and directory to improve user experience via uniform search, filter and display behaviour. Connect to JRC and ERN registry projects to update the contents (joint effort with contents update on Sample Catalogue. Implement JSON-LD based semantic layer to enable federated querying, i.e., to add the EJP RD common data elements as semantic annotation. Continue participation in Federated Query Builder WF (part 1, resource level).
When collections or sub-collections (described with the star schema) are found, either locally or through a federated query in another catalogue, the user can start a negotiation with the owner of the resource though the BBMRI-ERIC negotiator. This important use case was already piloted in the RD-Connect project, in EJP RD this will be fully integrated, as both the RD-Connect Registry and Biobank Finder and the BBMR-EIRC directory will both use the same technical foundation, supporting the Life-Science AAI.

2.6. RD-Connect Sample Catalogue

Contributors: Esther 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.

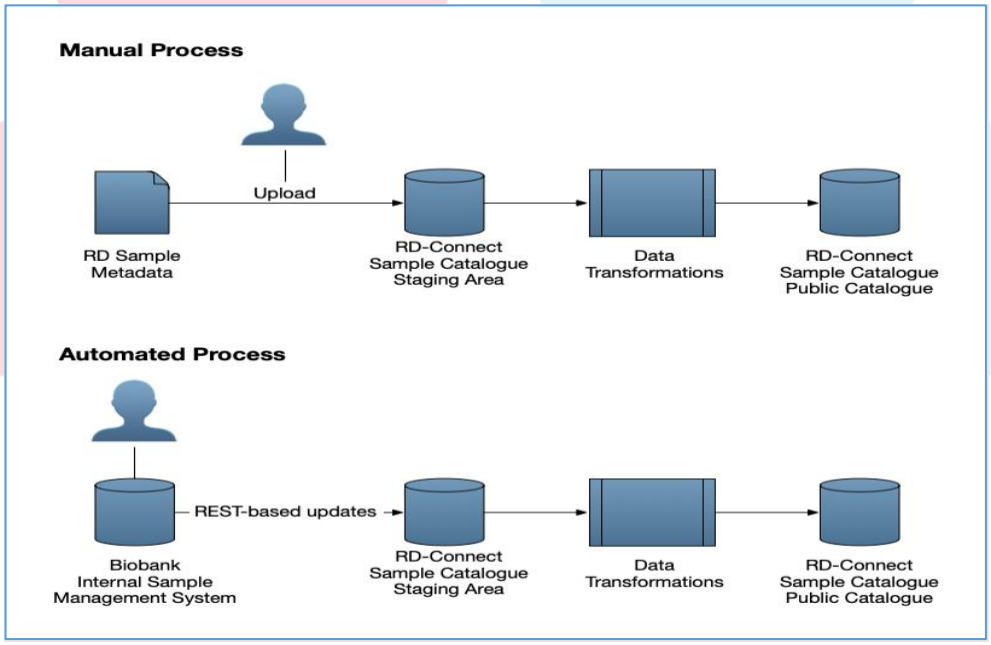


Figure 7. RD Connect Sample Catalogue Data Flow.

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 (CSV) 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 2020

The sample catalogue user interface was modernised using recent MOLGENIS developments, meaning that the outdated and boxy 'html 4.0' looking user interface was replaced with a more user-friendly layout based on Bootstrap styling framework, including minor refreshments of the layout. Upgrade to MOLGENIS 8.4.5 was performed. Time was mainly spent on maintenance of the data, and a technical merge with registry finder (above). All other improvements are still invisible to end-users, LifeScience AAI has been implemented for secure logins, and the FAIR data point framework was updated so directory and sample catalogue can integrate with GoFAIR/fairdatapoint.org (collaboration with LUMC). Piloting using new query engines (GraphQL & JSON-LD) as basis for mapping to the CDE of the Virtual Platform and preparation for VP federated search was perfomed. Finally, first ERN biobanks affiliated to the emerging ERN networks were reached out (interacting also with the FAIRification stewards). The catalogue now contains 52 biobanks, 353 registries and 68,454 samples related to RD.

2.6.3. Plans for improvement in 2021

Most importantly, a new data collection/update campaign is planned to update the catalogue contents. In particular, the ERN registry projects will be made aware of sample catalogue, registry finder and BBMRI-directory resources to ensure cross links

will be made. In addition, it is planned to redesign the user interface making use of directory and registry finder integration, and links to negotiator, and benefiting from MOLGENIS improvements; update data explorer, roll-out federated AAI, JSON-LD interface. Finally, the participation in the Query Builder Work Focus will continue to enable federated search.

2.7. BBMRI-ERIC Directory

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

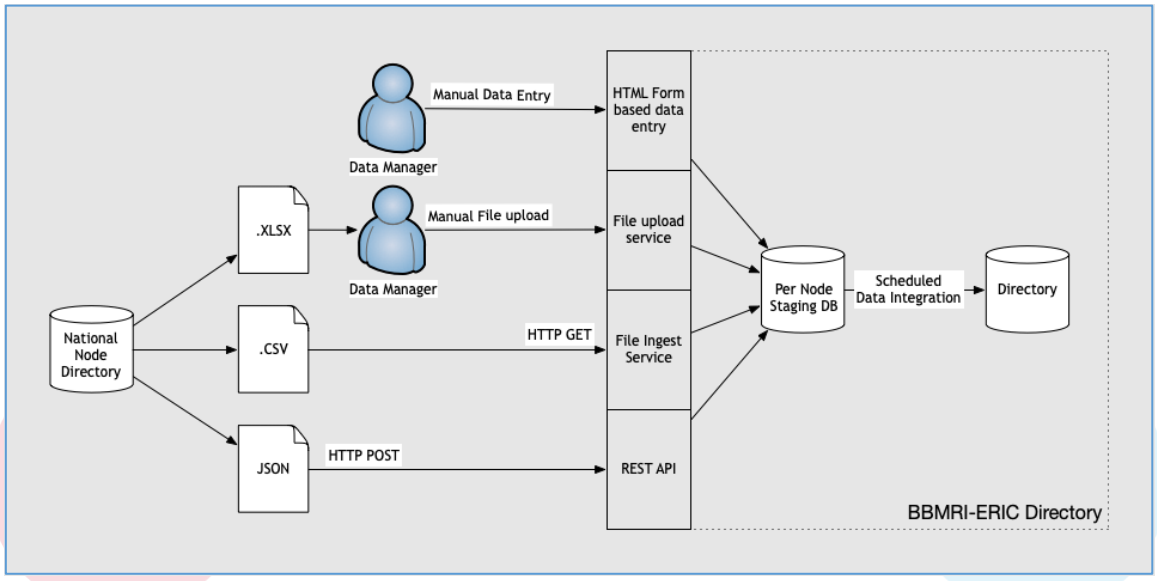


Figure 8. BBMRI-ERIC Directory data flow.

2.7.2. Features or facilities added in 2020

Multiple versions of the Directory were released, upgrading the user experience of browse and search scenarios including the improved (rare) disease filters, and technical integration with LifeScience AAI was made. Currently MOLGENIS 8.4.2 and with Directory user interface version 2.9.1. With co-funding of EOSC-life also COVID-19 sections were added. With co-funding of BBMRI-ERIC, a rebuilt of the selection +

request workflow with negotiator has been developed introducing a 'shopping cart' expected to be released in early 2021. The delivery on persistent ID and bioschema's has not been performed yet, however, JSON-LD and FAIR data point have been implemented as part of the FAIRification task of EJP RD. The FAIR data point is ready to deploy; JSON-LD (to be used for proper bioschema) will be deployed in 2021. Finally, a pilot implementation of the GA4GH standard DUO[18] was performed to enable rare disease researchers to better evaluate access conditions to the samples found in the directory.

2.7.3. Plans for improvement in 2021

The BBMRI-ERIC Directory development plan defines several topics to improve RD functionality with the following priorities for the Directory in 2021:

- Continue optimization of user experience for various browse and search scenarios in particular ability to count samples + filter on links to registry.
- Major update of RD contents in collaboration with sample catalogue this will include the extension of the data model towards attributes specific to registries and functionality to describe relationships between registries and biobanks
- Implementation of Query Builder WF pilots in particular the collection search box where search in Directory will also show ERDRI.dor and Orphanet hits and vice versa
- Make sure users start using DUO or other ontologies to describe data usage and access information in a machine-readable format to promote clear access conditions to be made available for rare disease samples (support of the use case as described in the "distributed consent control" use case)
- Implementation of Persistent Identifiers Policy to ensure cross references between EJP systems do not break. This is also important to identify and link resources, which are present in different catalogues (BBMRI-ERIC, ERDRI.dor, Orphanet).
- Integration of EJP RD resources in the BBMRI-ERIC negotiator workflow (pilot)
- Implementation of Bioschemas using JSON-LD to improve discoverability of RD resources in Google Data and other Bioschema enabled sources

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: Daphné Jaoui (INSERM-RaDiCo)

[18] https://www.ga4gh.org/news/data-use-ontology-approved-as-a-ga4gh-technical-standard/

2.8.1. RaDiCo introduction

RaDiCo is a national operational platform dedicated to the development, within a research framework, of many rare diseases (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)[19] 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.

[19] https://www.project-redcap.org/

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.).

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:** 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)[20] 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)[21]:** 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[22] and ORDO[23]. However, the Bioportal gives access to more than 800 available ontologies.
- **Thériaque[24]:** Thériaque is a database of all medicines available in France for health professionals.

[20] https://hpo.jax.org/app/
[21] https://www.meddra.org/
[22] https://www.orpha.net/consor/cgi-bin/index.php
[23] https://www.ebi.ac.uk/ols/ontologies/ordo
[24] https://www.theriaque.org/apps/contenu/accueil.php

2.8.2.2. RaDiCo Information System user's organization

RaDiCo Information System (IS) user's organization reproduces the cohort's organization in the RaDiCo Information System (IS) through identified roles. Each user with access to RaDiCo cohorts online has a defined role. Each role has precisely defined rights and access to medical data.
More generally, RaDiCo IS user's organization follows the principles of Attribute Based Access Control (ABAC)

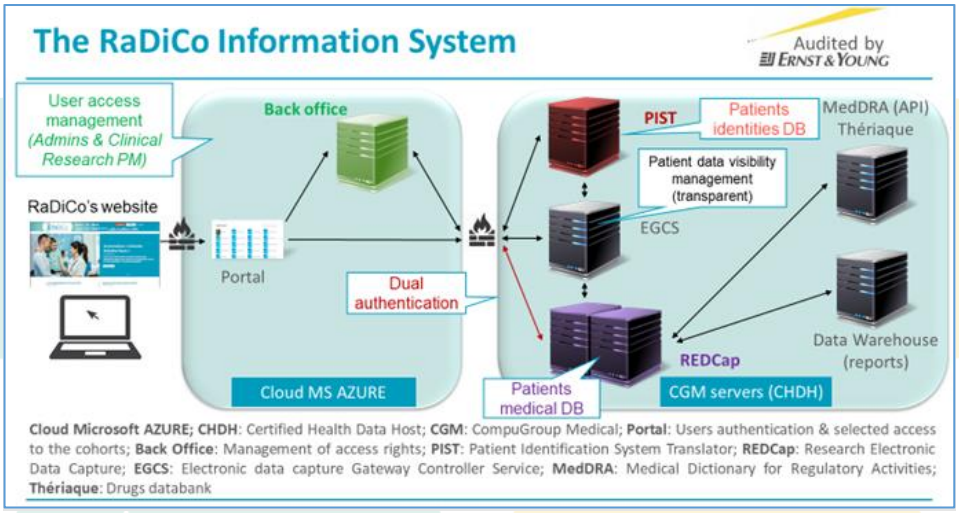


Figure 9. The RaDiCo Information System.

Clinical Research users:

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

IT users:

- Informatic System Administrator
- Software developer

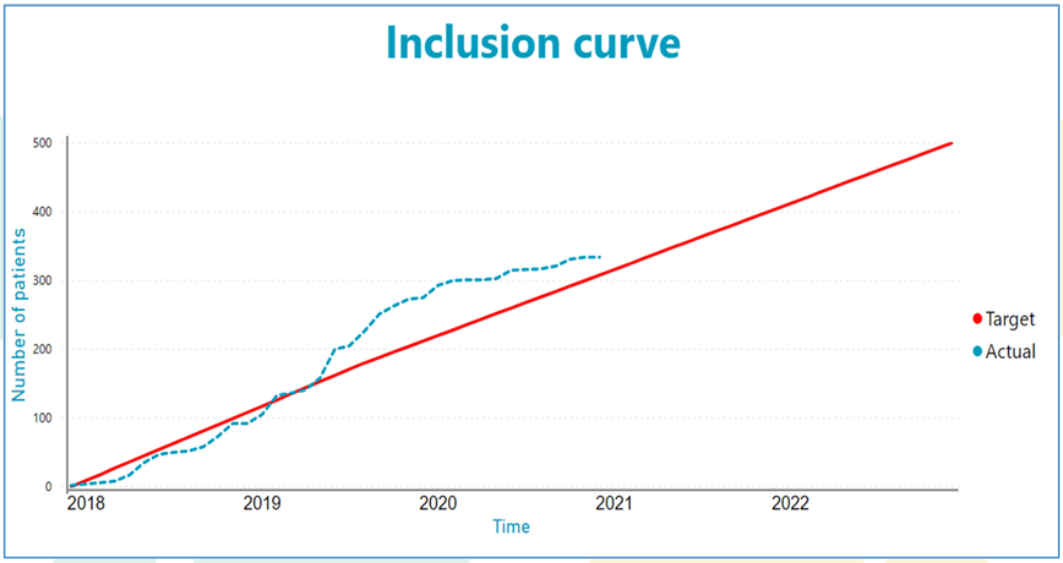
eHealth users:

- eHealth project managers

2.8.3. Features or facilities added in 2020

RaDiCo factorizes, mutualizes, and industrializes production processes for rare diseases (RD) cohorts. As an EJP RD member, RaDiCo already collaborates with some ERNs.
It improves reporting tools by introducing interoperability concepts (standardization, FAIRification, etc.). It uses the cloud computing (Infrastructure as a Service) to build and follow-up RDs e-cohorts, while respecting the European General Data Protection Regulation (GDPR).
The introduction of the concept of country within European multi-centres e-cohorts make it possible to manage patients' data followed-up in European centres and in non-EU centres as well. RaDiCo thus contributes to the improvement of the study of RDs e-cohorts.

For the year 2020, it was planned to add a Business Intelligence platform that will provide resources for statistical analyses, data management and Key Performance Indicators (KPIs).
This decision-making platform is destined to clinical research managers and to data managers. They can have an overview of all cohort progresses in the same tool.



These centralized automatic KPIs are now available in our platform and allow us to monitor every cohort in a centralized tool.

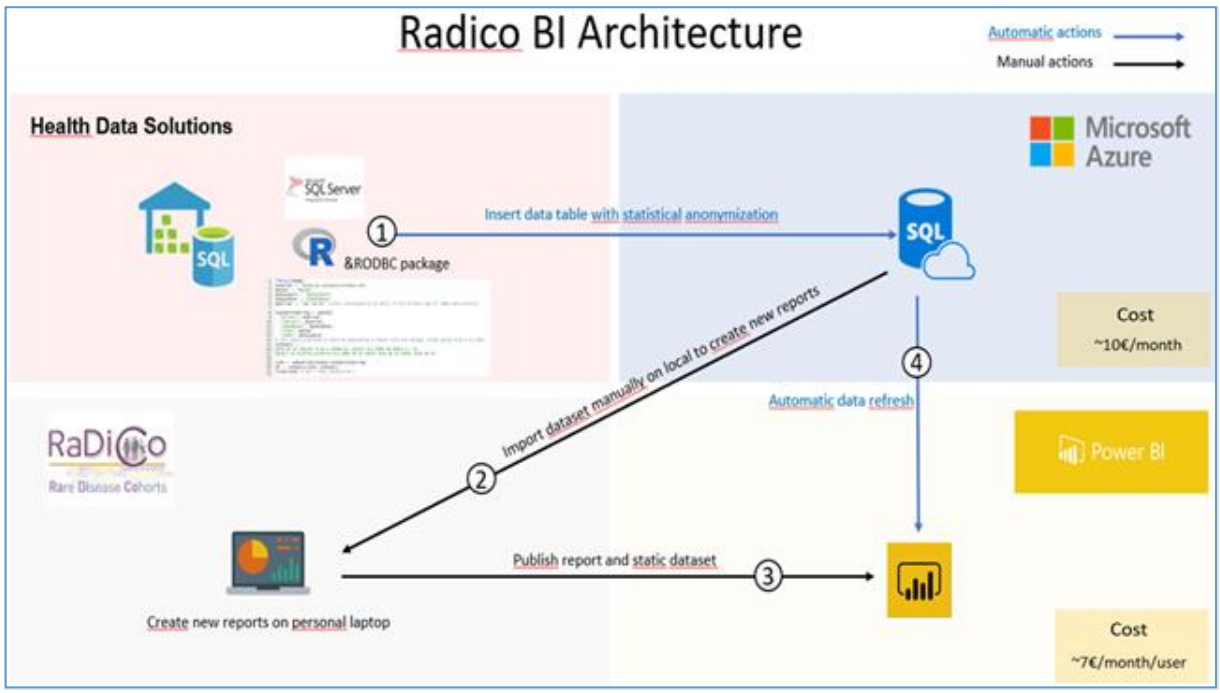


Figure 10. RaDiCo KPI tracking system.

Moreover, in line with the Certified Health Data Hosting, the recording and signalling logging system has been improved, including patients' inclusion and subsequent consultations. This tool is used by IT managers to track dysfunction or "malevolent" actions.

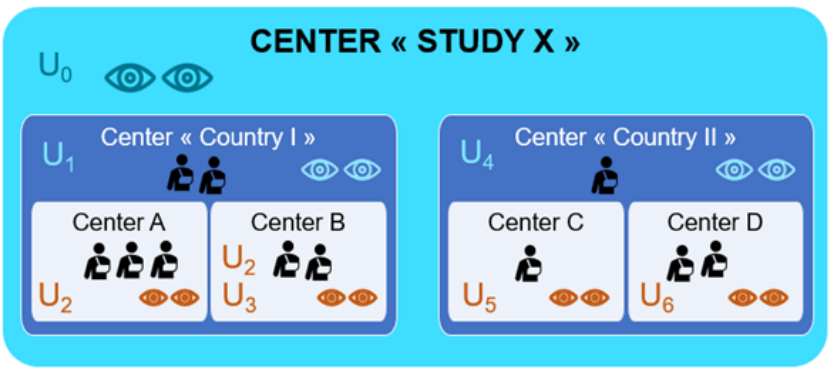
Auto-generated calendar module has also been added. This module improves patient follow-up in the longitudinal studies.

Different processes concerning patient data quality control have been identified:

- Duplicate management: this process allows to treat duplicates which could have been inadvertently created by users, and in accordance with the GDPR regulation. It allows us to move duplicates to another database, not accessible by external users and to have a treatment historicity
- Identity rectification management: this process allows to make patient identity data rectification in the platform, in accordance with the GDPR regulation
- Relocation of patient management (to another centre) our platform allows us to manage multicentric cohorts. When there is a patient's move to another centre, our system manages patient data transmission to the new user(s) from the new centre, where the patient just arrived.

Finally, the patient/user grouping system has been improved by adding notions of study, country and centre.

- ∉ "Study centre", where users have access to all patients of a given cohort. It is the case for data managers, project managers and statisticians.
- ∉ "Country centre" where users have access to all patients from a given country in a targeted cohort. It is the case for a geneticist who is making all genetic analyses for a given disease
- ∉ "Centre" in which users have access to all patients from a given centre in the cohort. It is the case for any investigator or Clinical research technician who is working in a centre.



2.8.4. Plans for improvement in 2021

- Implement a patient centre "historicity" (i.e., in the case where a patient is supported by a given centre and thereafter by another centre, for example when a young patient becomes an adult), identifying the current centre and previous centre(s)
- Set-up a procedure to reduce the creation of duplicate patient, while respecting patient's medical data confidentiality

- Set-up a procedure for sharing confidential documents with users, through REDCap.

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[25] 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.

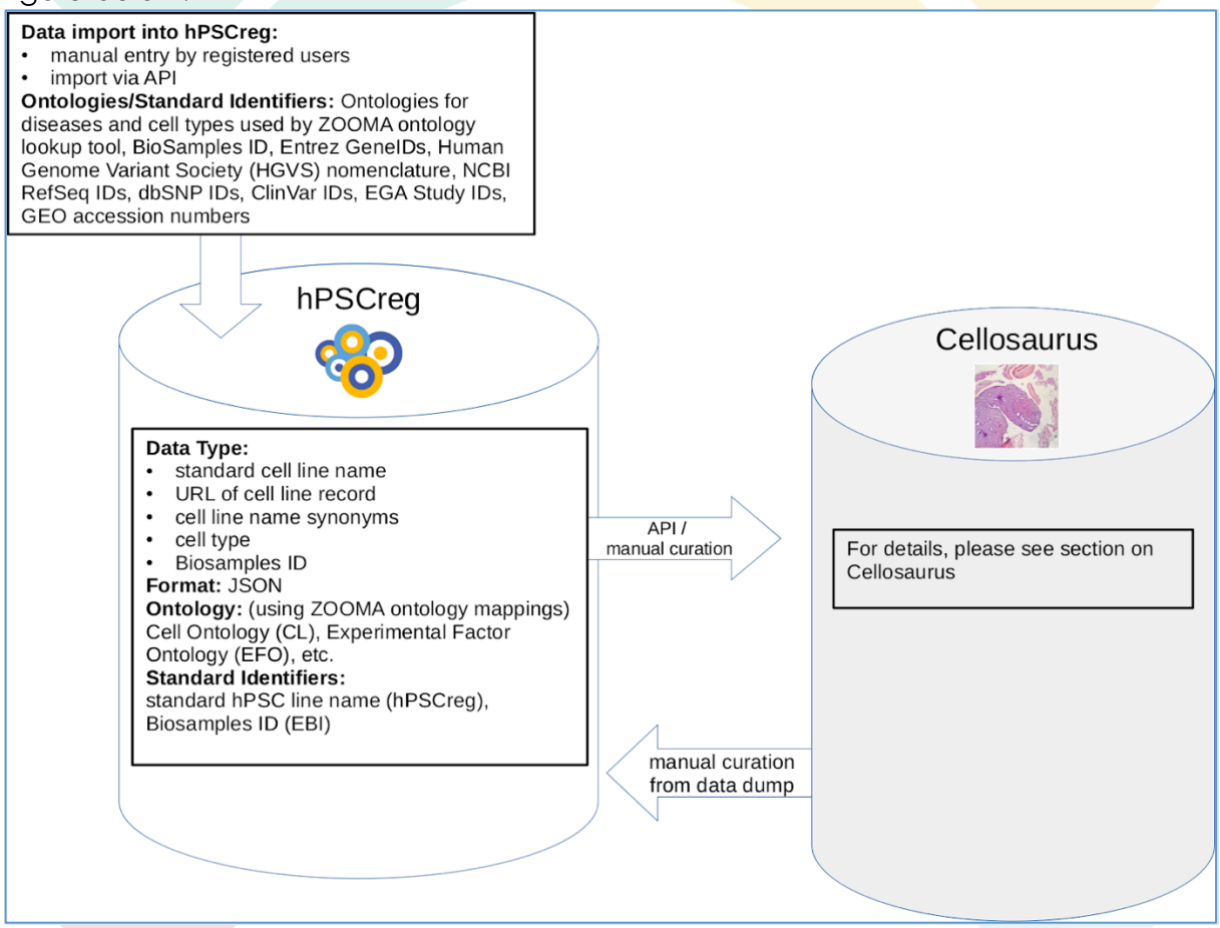


Figure 11: Data flow for the human Pluripotent Stem Cell registry (hPSCreg).

2.9.2. Features or facilities added in 2020

Connecting human pluripotent stem cell (hPSC) lines and their data from research applications to downstream uses such as hPSC-derived cell types for research or clinical translation gives comprehensive overview of hPSC research outputs for

[25] https://hpscreg.eu/

researchers, clinicians, patients, industry representatives and funders. In 2020, 217 new lines were submitted by users as of November 2020, of which 58 lines (27%) originated from donors with rare disease. To improve its visibility to the rare disease community, hPSCreg became a recognized resource of IRDiRC, and hPSCreg has presented at rare disease biobank training workshops to encourage hPSC lines as valuable tools for RD research. The clinical study database of clinical trials involving hPSC-derived cell types for interventional treatment was published in July 2020. To further develop the database of regulatory "primed" hPSC lines intended for clinical application, hPSCreg has garnered feedback from the stem cell community to guide its development.

2.9.3. Plans for improvement in 2021

While hPSCreg now has a registry of hPSC lines and a database of clinical studies involving hPSC-derived lines, a resource for hPSC-derived cell types is needed to fill the gap in the complete path to translational application of these lines. To this end, hPSCreg will add a database of hPSC-derived cell types to the overall suite of hPSC resources hosted at https://hpscreg.eu/. FAIRification of hPSCreg's data will continue and will facilitate hPSCreg's planned integration into the Virtual Platform in 2021.

2.10. Cellosaurus

Contributors: Amos Bairoch (SIB [ELIXIR-CH])

2.10.1. Resource data flow

Cellosaurus is a manually curated resource concerning cell lines. It provides a wealth of information on more than 125'000 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[26].

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.

[26] https://web.expasy.org/cellosaurus/

2.10.2. Features or facilities added in 2020

- Cell lines relevant to rare diseases have been mapped to the Orphanet ORDO nomenclature. There are currently about 32'000 cell lines that are mapped to a bit over 1'100 ORDO terms
- A new information topic concerning *Karyotypic information* was introduced.
- Cross-references were added to the Kerafast cell line collection and to the Liver Cancer Model Repository (LIMORE).
- Mapping disease variations to ClinVar[27] has been initiated.
- A massive retrofitting of the tissue origin of a cell line has been started.

2.10.3. Plans for improvement in 2021

It is planned to complete the retrofitting of ClinVar mappings to the variants that exist in ClinVar. The mapping of the tissue origin of cell lines to Uberon will be started.

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 about more than 7600 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.

[27] https://www.ncbi.nlm.nih.gov/clinvar/

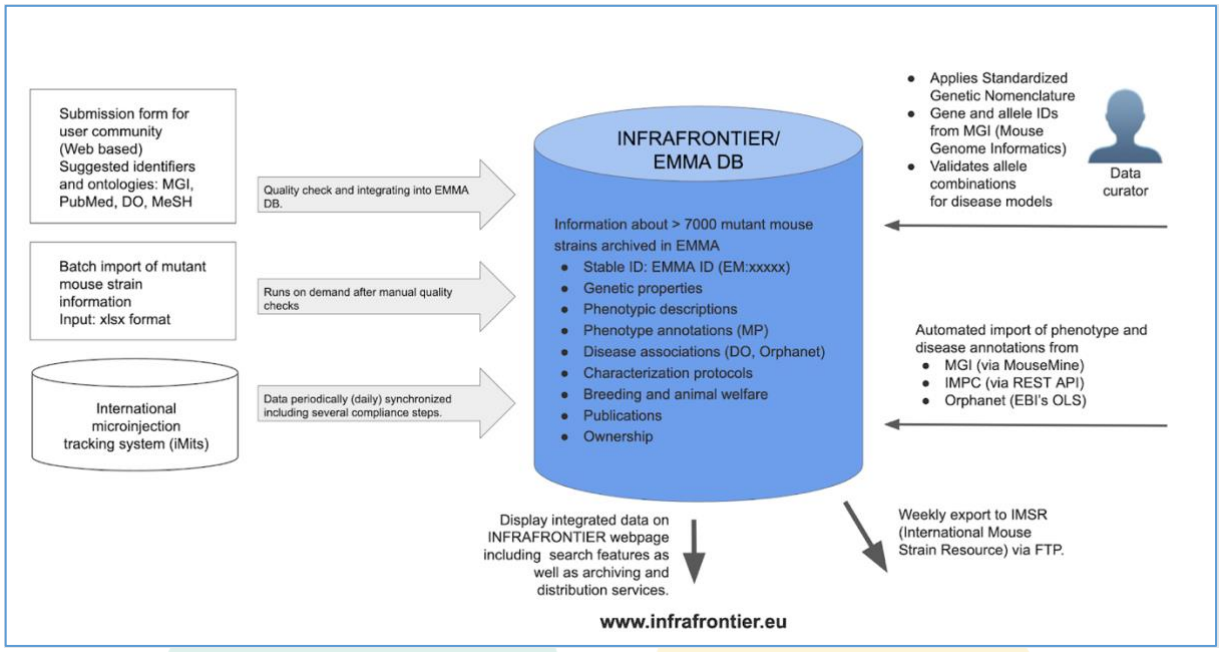


Figure 12. Data flow of EMMA.

2.11.2. Features or facilities added in 2020

The list of EMMA strains that are potentially interesting for rare disease researchers on INFRAFRONTIER's rare disease landing page[28] set up in 2019 has increased. Currently the EMMA repository holds 1551 mouse strains that carry mutations in 933 genes that have been implicated to play a role in rare diseases (1263 different rare diseases).

Although we managed to draw much attention to the dedicated landing page, we still wanted to include rare disease information directly on INFRAFRONTIER's central mouse strain search page[29], which is the one with the second most visits (after the home page). To further increase visibility of rare disease aspects at INFRAFRONTIER, rare disease links (via ORPHAcodes) have been integrated directly on this main EMMA search. It is now possible to search by rare disease names and IDs there as well.

A first virtual meeting between CNAG-CRG and INFRAFRONTIER to start defining the requirements for the API needed for the integration of INFRAFRONTIER data in the EJP RD virtual platform was conducted in November 2020.

2.11.3. Plans for improvement in 2021

From INFRAFRONTIER's side, the activities already started in year 2 will be continued in year 3. The focus will be on developing the API (in collaboration with the metadata, use cases and overall architecture work foci) for the integration of INFRAFRONTIER in the EJP RD virtual platform.

[28] https://www.infrafrontier.eu/infrafrontier-and-rare-diseases
[29] https://www.infrafrontier.eu/search

To improve the analysis capabilities in rare diseases, we will initiate a collaboration between RD-Connect GPAP and INFRAFRONTIER platforms. This collaboration will allow users from RD-Connect GPAP with a candidate gene to query INFRAFRONTIER for their database of mouse models; this query will return the available mouse models with alterations in the corresponding mouse gene of interest, which can help initiate new lines of research. In addition, we will expand queries using ontologies like HPO or ORDO, which will allow for genome-phenome comparisons between both species. This can provide valuable information to, for instance, define human candidate genes from a mouse model phenotype. Finally, we will study the possibility to expand these queries from RD-Connect GPAP to mutant mouse strains stored at INFRAFRONTIER.

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

2.12. 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[30]. 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.

[30] http://www.proteomexchange.org/

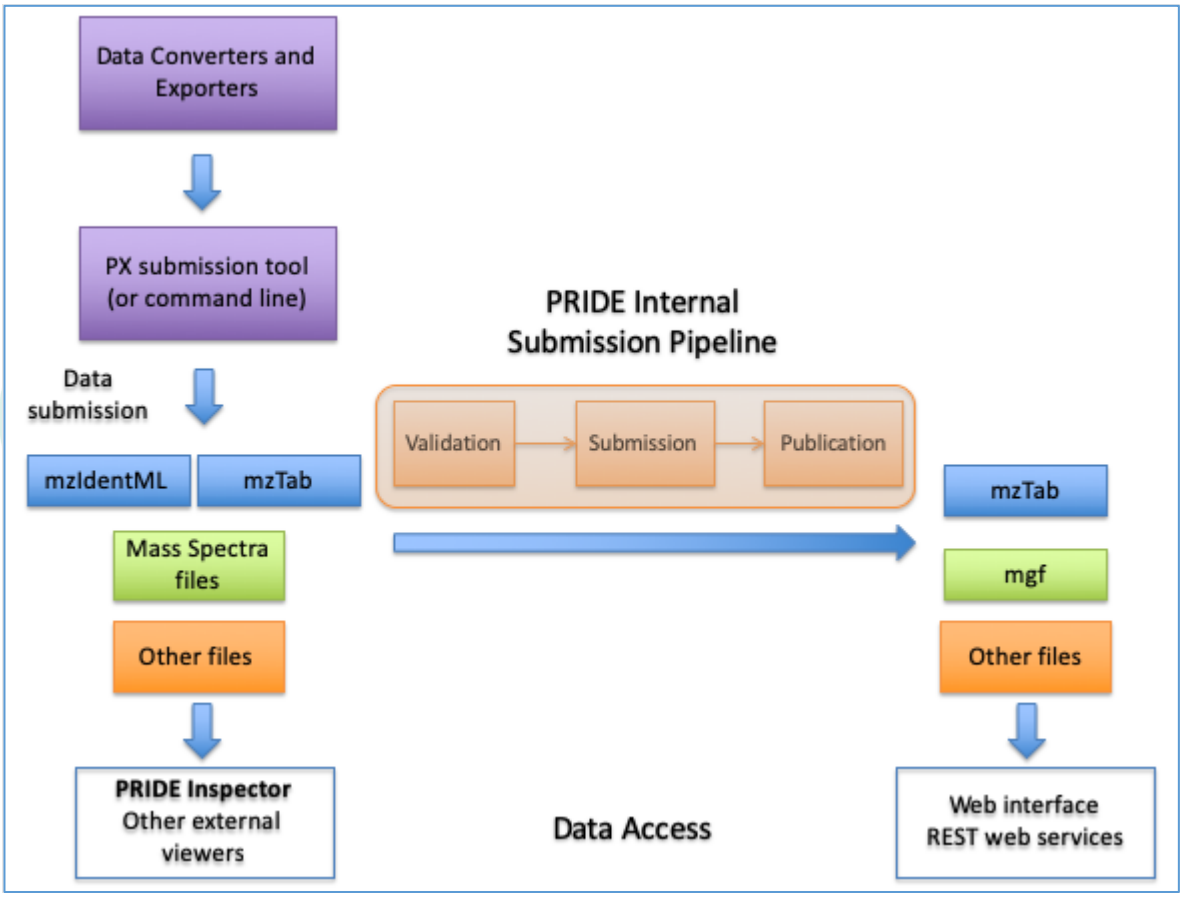


Figure 13: Data flow for PRIDE.

2.12.2. Features or facilities added in 2020

PRIDE has further improved its technical infrastructure in 2020 to support better its users and keep supporting the increase in submitted datasets (more than 450 datasets per month on average during 2020). The main improvements have been related to the robustness of the PRIDE system, involving its deployment in two data centres, and making the system much more "agnostic" in terms of underlying infrastructure, moving towards Kubernetes and postgres-QL. Re-uses of datasets available in the literature have started to be tracked (e.g. https://www.ebi.ac.uk/pride/archive/projects/PXD000561).

Additionally, support for Universal Spectrum Identifiers (USI, individual identifiers for mass spectra) has been implemented for "complete" submissions (e.g. https://www.ebi.ac.uk/pride/archive/spectra?usi=mzspec:PXD019317:sh_5282_HYK_101018_Mac_D_25mM.mzML:scan:38109:YMKYEK[MS:1001460]SYR/3).

Also, several improvements have been done in the PRIDE submission tool, involving support for checksums (to ensure the integrity of the uploaded files). Initial support for the SDRF-Proteomics format (for improving metadata annotation of datasets, see https://github.com/bigbio/proteomics-metadata-standard) has also been implemented. Additionally, PRIDE COVID-19 related datasets (32 public ones at the moment of writing) were manually tagged and made also available via the EMBL-EBI COVID-19 data portal (https://www.covid19dataportal.org/).

An increasing number of re-analysed quantitative proteomics (~50) datasets are either already available or will be soon released via the EMBL-EBI's Expression Atlas (https://www.ebi.ac.uk/gxa/experiments?experimentType=baseline&species=homo+sapiens, in the technology column, please filter by technology "Proteomics").

2.12.3. Plans for improvement in 2021

The SRDF-Proteomics format will be formalised as the format to provide metadata annotations for PRIDE dataset, and the associated open software will be released to facilitate annotation and re-analysis of public proteomics datasets. In this context, the linking between PRIDE and EBI's Biosamples database (https://www.ebi.ac.uk/biosamples/) will be established. As a key point, support for re-analysed datasets in the context of the originally submitted ones will be implemented. As part of this work, PRIDE submission and validation data pipelines will be redeveloped to better support handling and visualisation of proteomics MS data. Also, in the context of data re-analysis, dissemination of proteomics data into other data resources will continue, involving Expression Atlas (e.g., DIA quantitative datasets), UniProt (datasets involving Post-translational modifications) and Ensembl (for proteogenomics data).

2.13. MetaboLights

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

2.13.1. Resource data flow

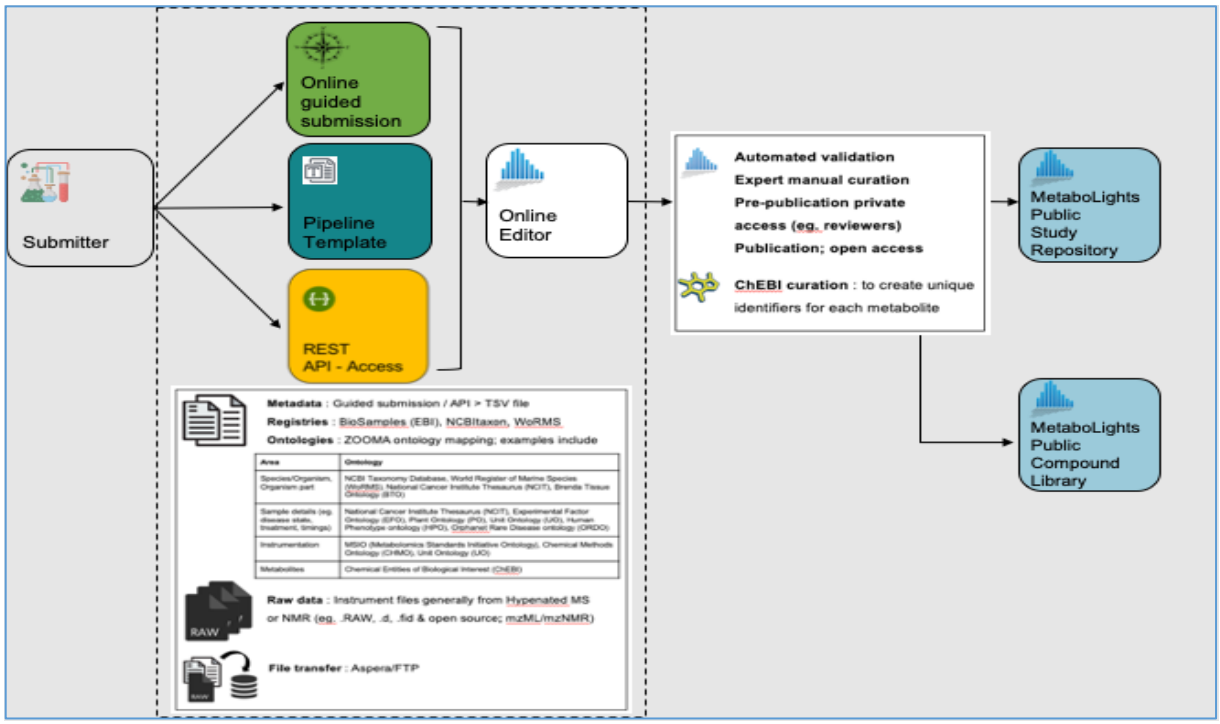


Figure 14. Dataflow for MetaboLights.

MetaboLights is a data repository for metabolomics data. Each new study is assigned a unique and persistent identifier. Submitters can choose to use the online guided

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

2.13.2. Features or facilities added in 2020

In 2020 we performed an audit of the online submission system which initially went live as a beta version in 2018. Results of the audit together with feedback from submitters have resulted in further improvements to the submission process. A significant effort has been put into the validation infrastructure and more specifically, handling of larger data sets often associated with human / clinical studies, to ensure accurate real time validations. The new validations will provide those submitting rare disease metabolomics data with faster, more intuitive responses to guide their submission.
We have also improved API access for study deposition which is available to all submitters. This has allowed us to collaborate with commercial metabolomics service providers resulting in an increased uptake in open-source deposition including rare disease studies.
To further support the submission of clinical or sensitive data, we have initiated the development of a controlled access instance of MetaboLights which will support deposition of studies with governance requirements.
The first COVID data set became public in September and several more are in the process of being submitted. Public studies will also be accessible via the EMBL-EBI COVID-19 data portal (https://www.covid19dataportal.org/). This has the benefit of extending our experience in handling clinical data.

2.13.3. Plans for improvement in 2021

The aim in 2021 is to continue the activities to improve the submission process for larger studies and the development of the controlled access instance of MetaboLights.
With the database developments, and other improvements over the last 2 years, including the addition of more specific ontologies for the rare disease community, we will now look to enhance the data discoverability. This will include the redevelopment of MetaboLights search capabilities and API access.
We will also update online guides to complement the new additions, including written guides and video tutorials, and will continue to support submission of rare disease studies with manual curation and active helpdesk.

While we have observed a slight increase in submission of studies associated with rare disease in the past year, we believe there is more opportunity to increase these submissions and provide the community with a comprehensive metabolomics repository. Feedback so far would indicate the controlled access aspect will be particularly important in that regard. We hope to further reach out to those researching metabolomics in rare disease and will participate in the EJPRD rare disease webinar series to both encourage data submission and demonstrate rare disease data availability with MetaboLights as well as providing training opportunities with metabolomics experts in our introductory course.

3. Section 3

3.1. Conclusion

Despite the challenges of 2020 the named deposition resources of EJP RD have continued to make improvements and new features to address rare disease use-cases. The resources still offer multi-faceted submitter and access routes, but standards are being broadly adopted to aid the RD community specific use-cases in the submission and access process. For example, the EGA has adopted the GA4GH Passport and Visas to allow interoperability with the ELIXIR AAI and Life Science AAI (standards) which allows users to link their EGA identity to an ELIXIR identity. This will ensure compatibility with the LifeScience AAI in future and will allow single sign on between the GPAP and EGA once the GPAP supports the LifeScience AAI (work due to be completed 2021).

Additionally, EGA has added CRAM support for the GA4GH standard htsget (currently available in test and due to be rolled out early 2021) which will allow users to download read data for subsections of the genome in which they are interested. It will save significant storage and compute power by avoiding the need to download whole files thereby potentially expediting research for the RD community. Also, by using the range request functionality of the htsget standard, the Data API has also been demonstrated to work with the GPAP, allowing ERNs to log into the GPAP and visualise the bam files in EGA using the Integrative Genomics Viewer (IGV) within the GPAP

A key goal for WP11 is to enhance cross linkage between rare disease deposition resources to enable researchers to easily access relevant data types presented through different resources (e.g., a mouse model with the same phenotype as rare disease patient). In EJP RD, examples from 2020 include how the RD-Connect GPAP will connect with INFRAFRONTIER to facilitate the cross querying of candidate genes to gather information from the mouse models database and also connect with hPSCreg to aid discovery of relevant cell lines within hPSCreg from the data they are analysing. Additionally, RD- connect GPAP will link to identifiers from EGA, which will help to facilitate data submission to RD-Connect GPAP through the EGA.
Over the coming year each of the resources will also be providing a webinar describing their service and offering users the opportunity to ask questions and provide valuable feedback.

Noteworthy, EJP-RD AWP3 plans to make RD-Connect GPAP, Cellosaurus, hPSCreg and INFRAFRONTIER represented through their metadata in the VP metadata model. This will be the first step towards VP findability and query ability.

Finally, the Work Focus 'Resources for the sharing of experimental data and materials' will continue to meet over the course of 2021 in order to update and discuss the progress of the resources based on this deliverable and to continue identifying improvements for the RD community in data submission and access processes.