

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 11.8

Second update Virtual platform of RD resources annotated with EJP ontological model

Organisation name of lead beneficiary for this deliverable:

Partner 01-b – INSERM-Orphanet

Contributors: AMC, BBMRI-ERIC, ELIXIR/EMBL-EBI, GUF, LUMC, UMCG,
UPM

Due date of deliverable: month 36

Dissemination level:

Public

Table of content

1. Introduction	2
2. Approach	3
2.1. FAIRification process based on the EJP metadata-model.....	4
3. Knowledge Bases	4
4. Integration of semantic resources into the EJP RD Virtual Platform	6
5. Updated Mapping services and Orphanet's API	9
5.1. ORPHAcodes mapping services update	9
5.2. Orphanet Registries & Biobanks API	10
6. Discussion and Next Steps	13

1. Introduction

During Year 1, based on the EJP RD Resources Metadata Model, a first implementation of a common model to capture descriptions of several catalogues was produced and made machine readable ([Deliverable 11.6](#)). Catalogue providers (RD-Connect registries and biobanks Finder, BBMRI, JRC-ERDRI, Orphanet) described and exposed their metadata elements in a standardized way.

During year 2 (Y2), the implementation work was reshaped according both to the adaptation of the Metadata Model to the DCAT 2.0 core model released in April 2020 and its adaptation to specific EJP resources description. A methodology to address the incorporation of new sources was proposed.

Furthermore, grounded to the EJP RD Resources Metadata Model, in a joint effort with Query Builder Work Focus (WF), a proof-of-concept (POC) was developed to present the catalogues information (Registries & Biobanks) accordingly (both through API and Semantic technology) and leverage the possibility for building a common query mechanism across resources. In order to extend the query engine, a dedicated cross medical terminologies mapping services API was developed, based on [ORDO](#) (Orphanet Rare Diseases Ontology), embedding a webservice to exploit the several rare diseases classifications. ([Deliverable 11.7](#))

As an extended version of the previous work, during the past year, we worked on the update of the EJP RD Metadata Model to add new resources (Knowledge bases, Innovation Management Toolbox, ...). An implementation of a semantic middleware was performed to ensure compatibility with API connexion and future integration of semantic components (FAIR Data Point, RDF triple stores). An updated version of the API catalogue was implemented, ([API V3](#)). Accordingly, the Orphanet catalogue API was modified to directly use different disease coding systems, and improve the compatibility with the semantic component, being also directly queryable by concept IRI (Internationalized Resource Identifier)

2. Approach

During Year 3 (Y3), based on the third version of the Ontological model of resources metadata (Del 11.3) we have implemented new resources description, including knowledge bases and directory of services: Cellosaurus, hPSCreg, Bio.tools and WikiPathways (figure 1)

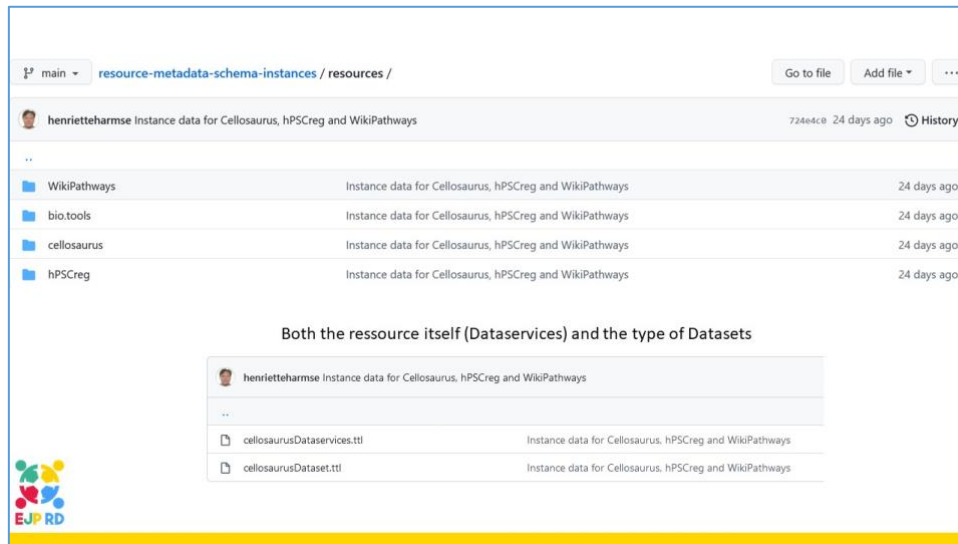


Figure 1. Update on resources description based on DCAT2.0 EJP RD Metadata Model (both resources and datasets are described)

Regarding the IMT (Innovation Management Toolbox), we have addressed the description of the content, following the methodology used in Year2 (Y2). This includes description of the content from ECRIN and EATRIS. (figure 2)

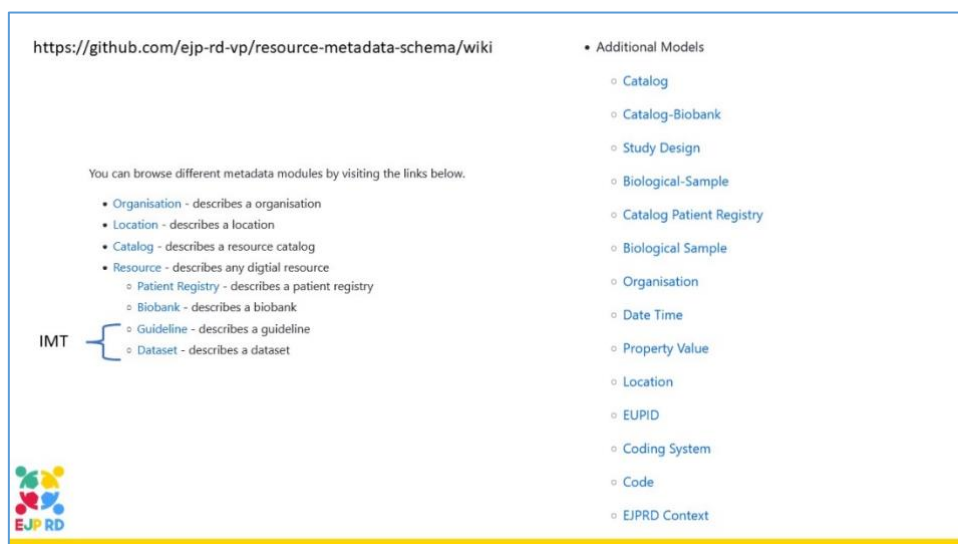


Figure 2. Representation of new resources metadata items (use for instance to describe IMT content)

Furthermore, during Y3, we explored technical solutions that can be compliant both with the semantic approach (FAIR Data Point, RDF and Triple store) and the REST API specification. This led us to use a middleware server "GRLC" to share the data with the virtual

platform. “GRLC” is a lightweight server that takes SPARQL queries, and translates them to Linked Data Web APIs. This enables universal access to Linked Data. Users are not required to know SPARQL to query their data, but instead can access a web API.

With a dedicated focus on implementation, we also worked closely with the Query Builder, Overall Architecture and Use Cases Work Foci in order to launch Version Zero (V0) of an EJP query portal compliant with the metadata model, available resources and API queries facilities.

A dedicated 2-day workshop bringing together the mentioned Work Foci was held in November 2021 in hybrid form. The workshop entailed collective brainstorming that resulted in the action plan for the launch of V0 and the future planning for the more comprehensive Version 1 expected in Year 4 of the project (Y4).

2.1. FAIRification process based on the EJP metadata-model

In Year 2 we FAIRified IMT resources. The process shown in Figure 3 was employed for this FAIRification. The description of IMT resources started from the scratch by the data provider, thus it gave us an opportunity to define an excel template together with data provider. The content of this excel template is later added to the FAIR Data Point with custom python. In the year 3 the resources to be integrated into the virtual platform have already achieved some degree of data FAIRness. For these resources we used FAIR Data Point to describe them and the way to access them. In the section 3 we give a detailed description of how these resources are described in the FAIR Data Point.

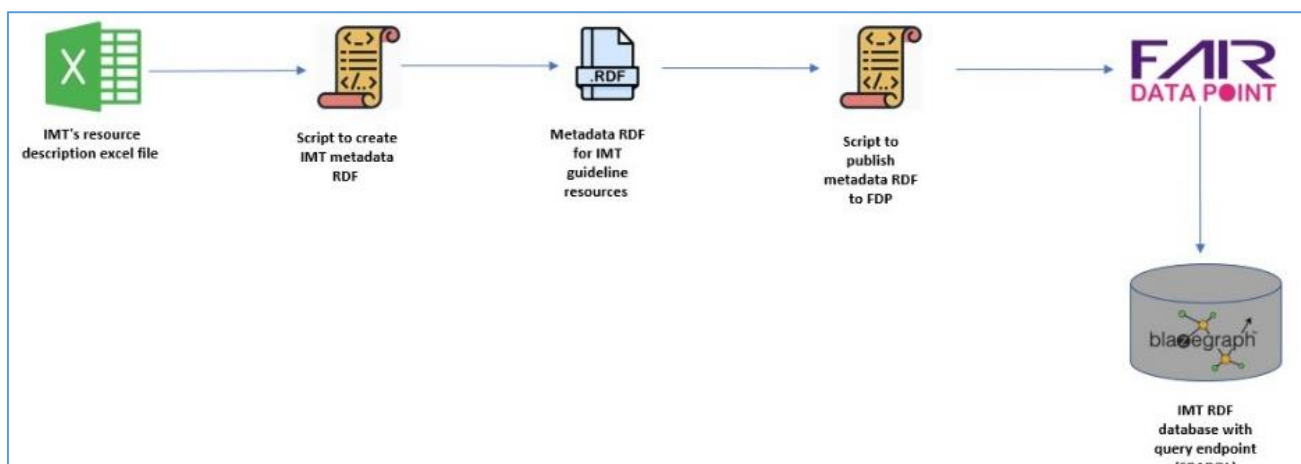


Figure 3. IMT resource FAIRification process

3. Knowledge Bases

During Y3 we used the updated metadata model to describe knowledge bases such as Cellosaurus¹, hPSCreg², Bio.tools³ and WikiPathways⁴. These resources are made available by the data providers in various formats. For instance, Cellosaurus content is made available via a webapp, a downloadable text file and a queryable distribution such as a SPARQL endpoint. In Y3 we setup a dedicated FAIR Data Point to describe these knowledge bases and the various distributions through which these resources are made available to data

¹ <https://web.expasy.org/cellosaurus/>

² <https://hpscereg.eu>

³ <https://bio.tools>

⁴ <https://www.wikipathways.org/index.php/Portal:RareDisease>

consumers. The knowledge bases' FAIR Data Point can be accessed [here](#). Figures 4 and 5 show the description of Cellosaurus and its SPARQL distribution in the FAIR Data Point respectively.

FAIR Data Point Search FAIR Data Point... AE

WP13 FAIR Data Point / Cell lines / Cellosaurus

Cellosaurus

Owner Edit Settings Delete

The Cellosaurus is a knowledge resource on cell lines. It attempts to describe all cell lines used in biomedical research. Its scope includes: 1) Immortalized cell lines 2) Naturally immortal cell lines (example: stem cell lines) 3) Finite life cell lines when those are distributed and used widely 4) Vertebrate cell line with an emphasis on human, mouse and rat cell lines 5) Invertebrate (insects and ticks) cell lines Its scope does not include: 1) Primary cell lines (with the exception of the finite life cell lines described above) 2) Plant cell lines

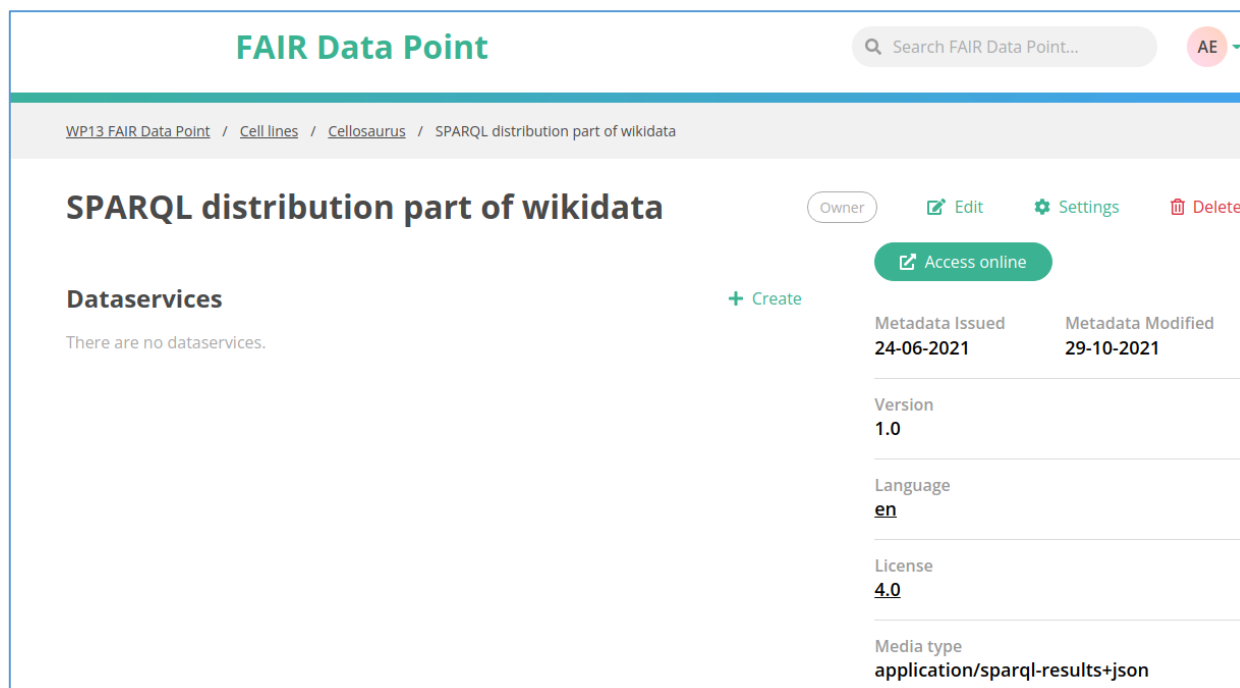
Distributions + Create

- EJPRD API**
Issued 29-10-2021 Modified 29-10-2021 Media Type application/json
- HTML distribution**
Issued 10-12-2021 Modified 10-12-2021 Media Type text/html
- SPARQL distribution part of wikidata**
Issued 24-06-2021 Modified 29-10-2021 Media Type application/sparql-results+json
- Text file**
Issued 24-06-2021 Modified 24-06-2021 Media Type text/plain

Metadata

- Metadata Issued: 24-06-2021
- Metadata Modified: 10-12-2021
- Version: 1.0
- Language: en
- License: 4.0
- Theme:
 - NCIT_C47846
 - SIO_010054
 - Q21014462
- Contact point
- Keyword:
 - Cellosaurus
 - Knowledge Base
 - cell line

Figure 4. Description of the Cellosaurus resource and its distribution in the FAIR Data Point



FAIR Data Point Search FAIR Data Point... AE

WP13 FAIR Data Point / Cell lines / Cellosaurus / SPARQL distribution part of wikidata

SPARQL distribution part of wikidata Owner Edit Settings Delete

Access online

Dataservices + Create

There are no dataservices.

Metadata Issued	Metadata Modified
24-06-2021	29-10-2021
Version	
1.0	
Language	
en	
License	
4.0	
Media type	
application/sparql-results+json	

Figure 5. Description of the SPARQL distribution of the Cellosaurus resource in the FAIR Data Point

4. Integration of semantic resources into the EJP RD Virtual Platform

As mentioned in section 2, during Y3 we used the GRLC server to integrate semantic resources to the virtual platform. The driving use case for this approach is the integration of the knowledge base resources described in section 4. The data providers of these resources made their data available as semantic resources (RDF, SPARQL endpoint). In order to use these resources in virtual platform we proposed a design shown in figure 6. We used this design to implement the EJP RD query builder API specification on semantic resources. In this design the GRLC server is used as a middleware application which translates REST API calls into SPARQL queries and SPARQL query responses to JSON documents. Out of the four knowledge base resources Cellosaurus and WikiPathways made their data available via SPARQL endpoints so we setup GRLC servers directly on the SPARQL endpoints provided by the data provider. In the case of hPSCReg and Bio.tools the data providers provided an RDF dump of their data: for these two resources we setup a triplestore to host the RDF data dump and we used this SPARQL endpoint to setup a GRLC server.

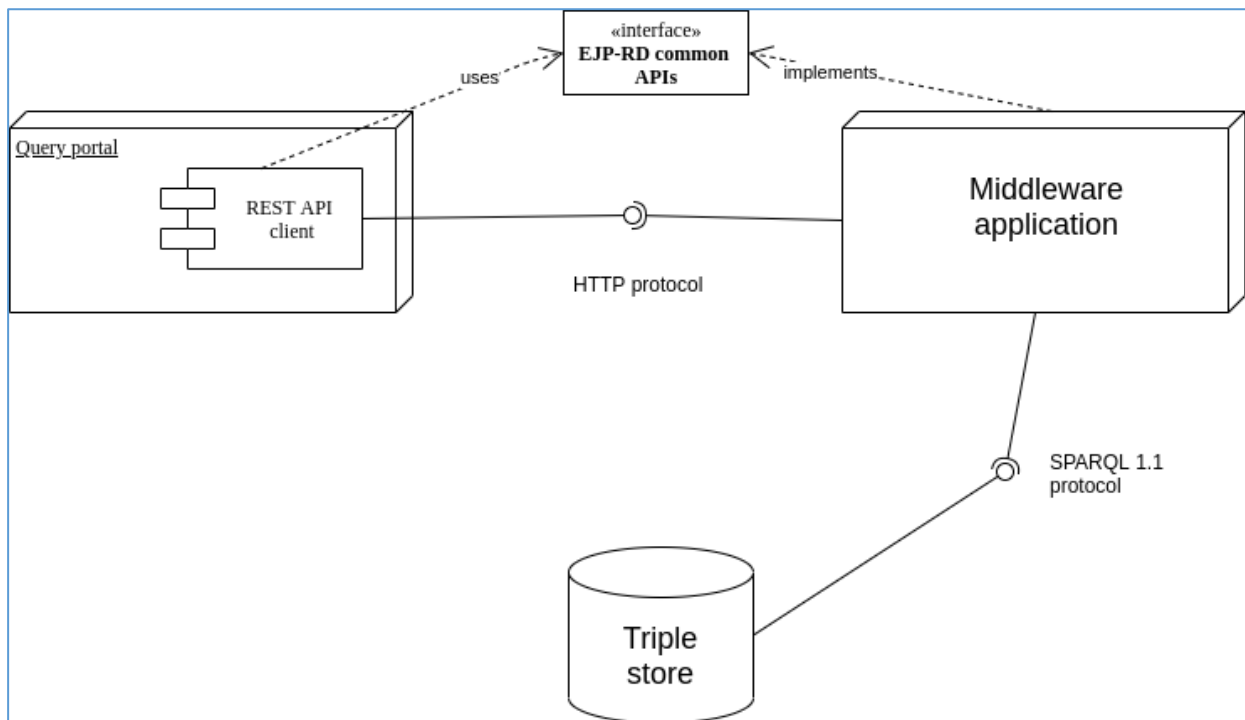


Figure 6. Design to implement the EJPRD query builder API specification on triplestore using a GRLC server

Figure 7 shows a list of knowledge bases and other resources that are currently queryable from the EJP RD virtual platform. In the current version of the virtual platform the Cellosaurus, hPSCreg, and WikiPathways knowledge bases are queryable with ORPHAcodes and all these three resources are using the design proposed in figure 6.

Figure 8 shows the search results for the term “Ovarian cancer” in knowledge bases. For this search term Cellosaurus and WikiPathways returned 607 cell lines and 8 pathways respectively. The GRLC server serving these resources is configured in a way that it can make use of the Orphanet Rare Disease Ontology’s hierarchical structure. The results returned by the virtual platform in this example contain results for any subclass of “Ovarian cancer” term.

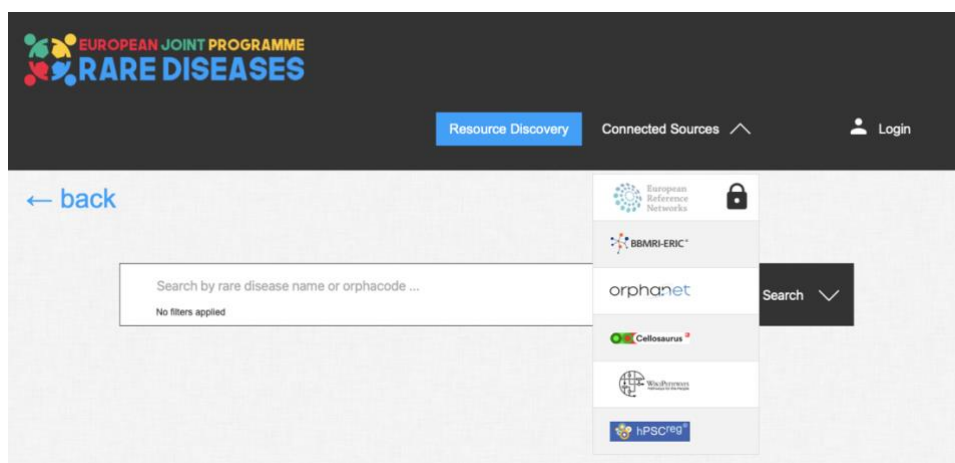


Figure 7. List of knowledge bases and resources made available to the EJP query portal

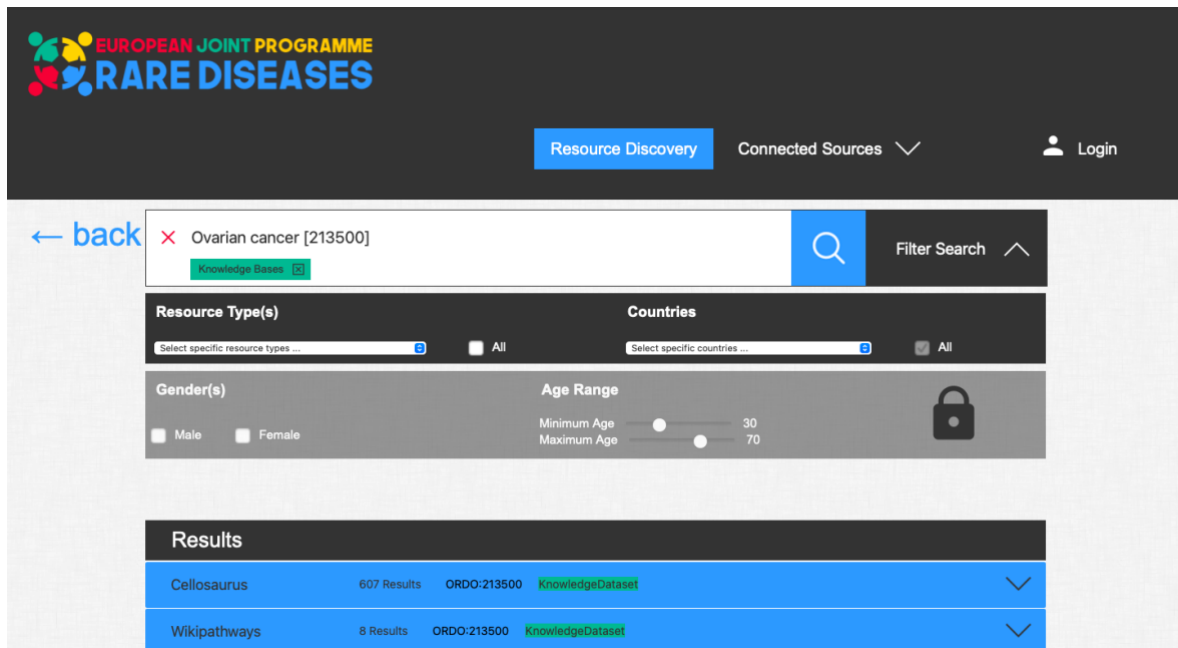


Figure 8. Search results for the term “Ovarian cancer” in knowledge bases. In this query the virtual platform uses the EJP RD metadata model to query the resources

The previous components, both resources API and integration of the semantic resources, compliant with the EJP-RD metadata model and API specifications are now integrated into the first release (V0) of the [EJP-RD Virtual Platform portal access point](#) (figure 9 and figure 10).

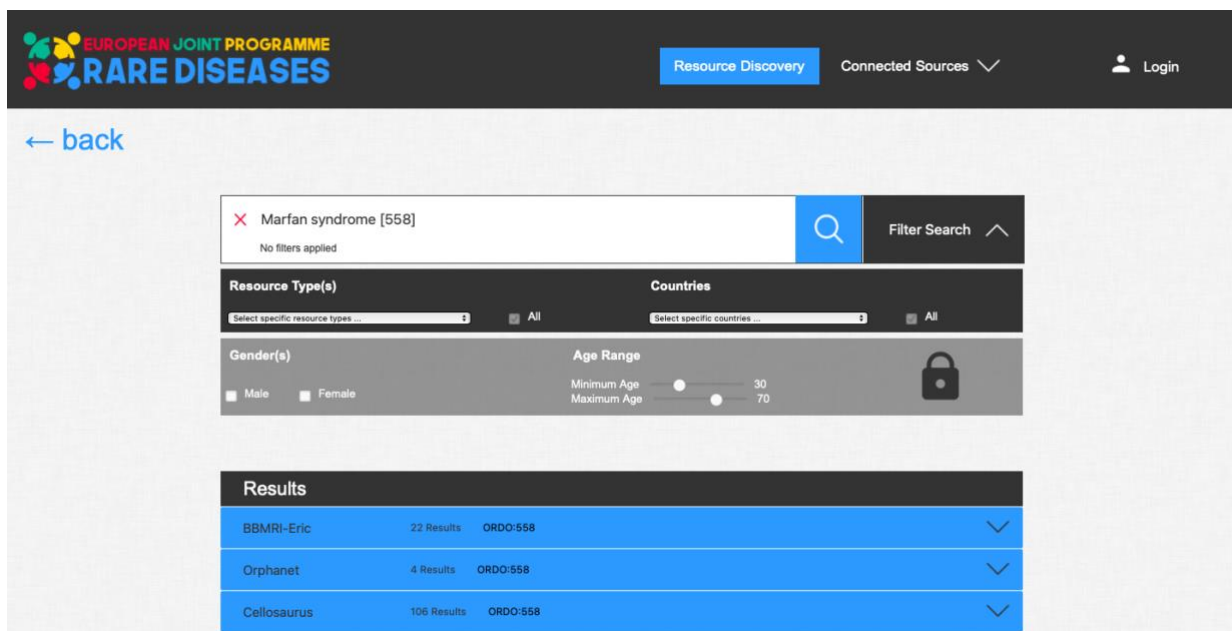


Figure 9. Search result sample from V0 of the EJP-RD Virtual platform

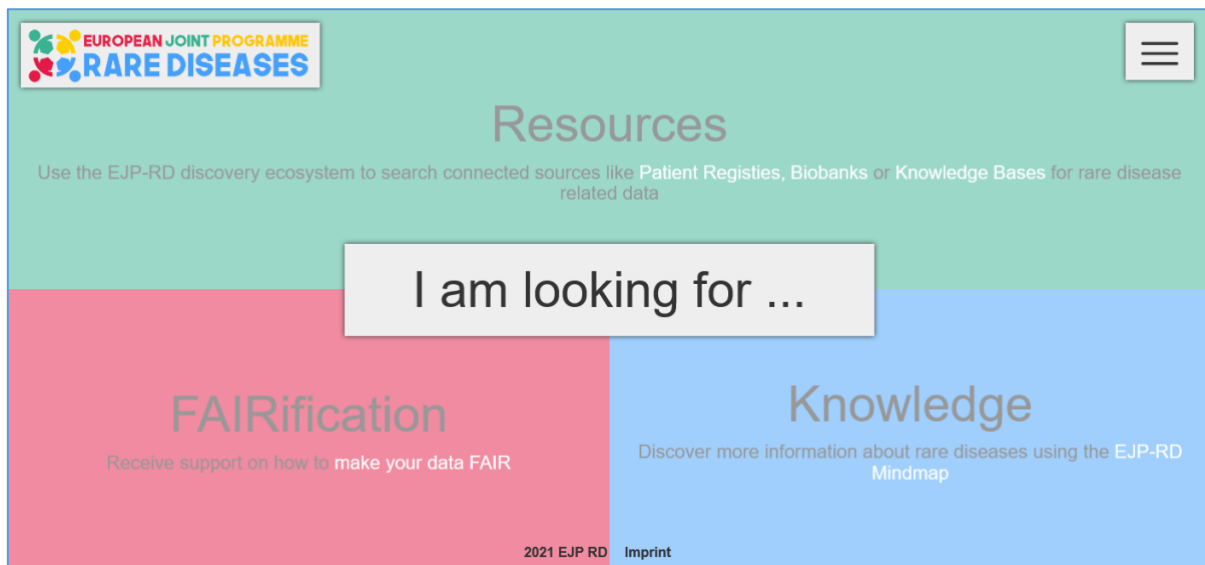


Figure 10. V0 of the EJP-RD Virtual platform access point

5. Updated Mapping services and Orphanet's API

Both the Orphanet's mapping services and the Orphanet Registries & Biobanks catalogues were updated. Especially, the Orphanet catalogue's API is fully compliant with the mapping services

5.1. ORPHAcodes mapping services update

Orphanet provides mappings between the Orphanet nomenclature (ORPHAcodes) to other terminologies, including OMIM, ICD-10, MeSH, MedDRA, UMLS. These mappings are released as downloadable files in Orphadata (www.orphadata.org) and are integrated in ORDO ([Orphanet Rare Diseases Ontology](#)). The [mappings](#) are expert-reviewed, using semi-automated mapping tools and are validated by the Orphanet curation team. Each mapping is assessed and identified as an Exact match or not with respect to the respective ORPHAcodes(s).

The alignments specify the comparability between terminologies by defining if the concepts are perfectly equivalent (exact mapping) or not.

- *E (Exact mapping: the two concepts are equivalent).*
- *NTBT (ORPHAcodes's Narrower Term maps to a Broader Term).*
- *BTNT (ORPHAcodes's Broader Term maps to a Narrower Term).*

The mappings are embedded directly on the Orphanet Knowledge base and made available through several tools provided by Orphanet (for instance OMIM and ICD-10 mappings are provided through the [RD-Code project API](#) released on an annual basis)

When using the ontology, the mappings are directly available (with an exception with SNOMED CT mappings, due to specific licensing conditions).

Nevertheless, the integration of an ontology needs specific skills and a technical setup (SPARQL query tools). In the context of the EJP-RD, Orphanet provides a dedicated API to query the mappings. Both API or ORDO mappings can be used.

An updated version of the services has been made available. The description of the [API mappings](#) service is accessible in the EJP GitHub repository.

Useful links are provided in table 1 below.

Table 1. Orphanet Mapping and API Parameters

<u>Element</u>	<u>Link(s)</u>	<u>Details</u>
API parameters	/mapping/: get: description: "Get information about the API."	Get information about the API
	/mapping?form={origin}&code={code}&to={destination} will return results for a clinical entity (resources), from any terminology listed to another (orphanet, omim, umls, mesh, meddra, icd). Based on Orphanet's mappings. "	Get mappings with a query for a clinical entity (resources). From any terminology listed to another (orphanet, omim, umls, mesh, meddra, icd). Based on Orphanet's mappings.
	<u>/mapping?form={/mapping?form={origin}&code={code}&to={destination}</u>	will return results for a clinical entity (resources), from any terminology listed to another (orphanet, omim, umls, mesh, meddra, icd). Based on Orphanet's mappings.
Querying using a different coding system	http://purl.org/orphanetws/Mapping/map?from=icd&code=Q87.4 http://purl.org/orphanetws/Mapping/map?from=omim&code=154700	The webservice is queryable by using different coding systems.
Parallel mapping to more than one coding system at once using the API mapping webservice	http://purl.org/orphanetws/mapping?from=orphanet&code=558&to=omim&to=icd	The API could be used to obtain codes at once

5.2. Orphanet Registries & Biobanks API

Furthermore, this mappings API webservice was also implemented into the Orphanet's catalogues (Registries & Biobanks) API. Accordingly, the catalogue API is usable with any code. The expected parameter is the code IRI, which is an output of the mapping webservices.

Based on that, when querying the catalogue API, the ORPHAcodes are used as a backbone for mappings, but are not mandatory to perform the query. Some examples are included in table 2 and a sample output in Figure 11 hereunder.

Table 2. Sample Queries of Orphanet Registries and Biobanks

Coding System	Example query
ICD-10	http://155.133.131.171:8080/Orphanet/resource/search?code=http://identifiers.org/icd/Q87.4
MedDRA	http://155.133.131.171:8080/Orphanet/resource/search?code=http://purl.bioontology.org/ontology/MEDDRA/10026829
MeSH	http://155.133.131.171:8080/Orphanet/resource/search?code=http://id.nlm.nih.gov/mesh/D008382
UMLS	http://155.133.131.171:8080/Orphanet/resource/search?code=https://www.ncbi.nlm.nih.gov/medgen/C0024796
OMIM	http://155.133.131.171:8080/Orphanet/resource/search?code=https://www.omim.org/entry/610168

```

155.133.131.171:8080/Orphanet/resource/search?code=http://identifiers.org/icd/Q87.4
JSON  Données brutes  En-têtes
Enregistrer Copier Tout réduire Tout développer Filtre le JSON
▼ resourceResponses:
  ▼ 0:
    id: "73803"
    type: "BiobankDataset"
    description: "Biobank of MD-NET: Muscle Tissue Culture Collection (MTCC) (EuroBioBank partner)"
    name: "MD-NET: Muscle Tissue Culture Collection (MTCC) (EuroBioBank partner)"
    homepage: "http://www.orpha.net/consor/cgi-bin/OC_Exp.php?Lng=en&Expert=229349"
    location:
      country: "DE"
  ▼ 1:
    id: "67063"
    type: "PatientRegistryDataset"
    description: "PatientRegistry of RaDiCo-MARFAN: National cohort on Marfan syndrome and apparent diseases"
    name: "RaDiCo-MARFAN: National cohort on Marfan syndrome and apparent diseases"
    homepage: "http://www.orpha.net/consor/cgi-bin/OC_Exp.php?Lng=en&Expert=205886"
    location:
      country: "FR"
  ▼ 2:
    id: "141038"
    type: "PatientRegistryDataset"
    description: "PatientRegistry of SACHER: Swiss Adult Congenital HEart disease Registry"
    name: "SACHER: Swiss Adult Congenital HEart disease Registry"
    homepage: "http://www.orpha.net/consor/cgi-bin/OC_Exp.php?Lng=en&Expert=563035"
    location:
      country: "CH"
  ▼ 3:
    id: "101485"
    type: "BiobankDataset"
    description: "Biobank of Marfan biobank"
    name: "Marfan biobank"
    homepage: "http://www.orpha.net/consor/cgi-bin/OC_Exp.php?Lng=en&Expert=362573"
    location:
      country: "HU"
  apiVersion: "v0.3"
  ▼ page:
    number: 0
    size: 10
    totalPages: 1
    totalElements: 4
  
```

Figure 11. Output of Orphanet catalogue API, ICD-10 code as input parameter

6. Discussion and Next Steps

The work will be pursued during next year, in line with the Y4 objectives.

Nevertheless, during the 2 previous years we faced various challenges during the inclusion of new resources into the EJP RD metadata model and their technical implementation; particularly in resources without a pre-existing structured model (such as Innovation Management Toolbox), for which a methodology was devised in collaboration with Work Package 19 partners ([DEL 11.7](#)).

During Year 3, as we addressed several resources with their own data model and metadata model, the approach was focused on the adaptation to and compliance with the EJP RD metadata model.

Based on both types of experiences, we foresee the need to provide a dedicated methodology guideline. This methodology guideline will leverage the capabilities of the future integration of any resource into the EJP RD virtual platform ecosystem and can be complementary to the [VIPS document](#) (Virtual Platform Specifications).

Furthermore, with the release of the EJP RD Virtual Platform access portal (V0), which includes real resources metadata and accessible datasets, we will need to iterate our work in an adaptive way. Based on the users' feedback and input from the Use Cases WF, new functionalities and improvements will need to be developed. This requires close collaboration with existing resources.

The launch of the successive versions of the portal will reshape the coming work and enable us to proactively ensure continuous integration of end-user and resources feedback. This will lead to a new prioritization process to ensure both alignment with Year 4 workplan and coordinated work of the Work Foci to meet the end-user's needs.

During next steps we will also dedicate efforts to shape efficiently DCAT2.0 described resources and record-level available datasets within a continuum to ensure the link between high level metadata description and access to data.