D11.9
Third update Virtual platform of RD resources
annotated with EJP ontological model

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 11.9
Third 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 48

Dissemination level:
Public

D11.9
Third update Virtual platform of RD resources
annotated with EJP ontological model

Table of content

Table of Contents

1. Introduction

During the first 3 years of the project, subsequent implementations of the EJP RD metadata model to capture descriptions of a growing number of resources included in the Virtual Platform (VP) as a network were produced, introducing adaptations based on the evolution of the metadata model itself, from a first version released in Y1 to a DCAT-based, evolutive version, released from Y2 on. Common models and methodologies to capture and make discoverable resources' description were piloted from Y1, from a central data linkage pilot VP to a federated model based on a first version of the EJP RD API, demonstrating in year 3 a FAIR-compliant federated ecosystem based on FAIR Data Points. Webservices were progressively developed in order to implement the Orphanet nomenclature of rare diseases (ORPHAcodes) by the resources and by the human-readable interface (VP portal)

This deliverable presents the developments performed in Y4 to implement the VP as a network based on the evolutions of the metadata model, the incorporation of new resources and the implementation of a machine-readable and human-readable federated ecosystem.

During year 4 (Y4) further webservices were developed to fit the needs of functionalities in the VP ecosystem and especially the human-readable VP Portal. Beyond the update of the previous webservices provided by Orphanet, we added an API mechanism to ease the implementation of Orphanet's Rare diseases classifications and a new webservice which allows to search disorders by gene or genes by disorder based on the Orphanet Knowledgebase content.

Furthermore, based on the work previously done during Y3, we have developed a dynamic VP index concept, which is automatically generated from metadata that is provided in terms of the EJP RD metadata model by resources through a federation of FAIRdatapoint (FDP). This index is itself also an FDP, such that computational agents can identify resources

D11.9
Third update Virtual platform of RD resources
annotated with EJP ontological model

by their machine understandable descriptions. Any resource that provides DCAT metadata can be indexed and then be referenced by the VP Portal.

The VP Portal demonstrates the implementation of this mechanism in its backend. Moreover, the VP portal backend has been adapted to implement also new versions of the VP API specifications (https://github.com/ejp-rd-vp/vp-api-specs) and accordingly the front end will be adapted to reflect the new functionalities (search by genes, search by "group" of disorders...). A user interface working group was set up to design the new user interface.

2. Approach

During Year 4 (Y4) and following the fourth Annual Retreat decisions (April 2022), an iterative approach was taken to work toward the development of Version 1 of the Virtual Platform. A series of strategic workshops and hands-on hackathons took place in coordination with the Metadata, Overall Architecture, Query Builder, and Resources and Materials Work Foci. The process started with 2 strategic workshops in Heidelberg in September and in Vienna in November to refine the VP architecture and its respective components, and the levels of connection ranging from basic discoverability to federated analysis (4 levels).

The outcomes were handed over to the different teams to be executed in a series of hackathons and development meetings. The developments presented in this deliverable were designed and implemented during (and following) two of those meetings: the developers' meeting in Frankfurt in November in which the Orphanet mapping services and their use in expanded search were thoroughly discussed (refer to section 4) and the Resources Onboarding Hackathon in Barcelona in December in which the VP Index was connected to the VP Portal (refer to section 3).

In addition to the extensive development work, the Metadata team conducted individual meetings with 11 EJP RD resources between March and November 2022 to prepare them for the onboarding process by providing a demonstration of the Virtual Platform Portal in its beta version and to support them in the description of their metadata (refer to deliverable 11.04) for population on the VP Index. The process provided valuable feedback that drove the improvement of search functionalities and enabled a successful VP Index connection in December 2022.

3. The Virtual Platform Index and the FAIR-complaint approach for resource onboarding to the Virtual Platform

The FAIR Data Point (FDP) is a metadata service that, from one side, allows digital objects owners/publishers to expose the metadata of their digital objects following the FAIR principles and, for another side, allows digital objects' consumers to discover information (metadata) about the offered digital objects. Commonly, the FAIR Data Point is used to expose metadata of datasets but metadata of other types of digital objects can also be exposed such as biobanks, patient registries, tools, etc.

In a complex and heterogeneous ecosystem such as the one expected around the Rare Diseases Virtual Platform, certain levels of interoperability need to be agreed. In the case of the VP level 1 (discoverability) the agreement is to use the FDP approach to expose the metadata of the resources to be discovered through the VP.

While providing metadata in a FAIR way contributes to making the described resources more findable, accessible, interoperable and reusable, if no one knows where these

D11.9
Third update Virtual platform of RD resources
annotated with EJP ontological model

metadata are been offered does not help much. A fundamental part of the approach is to follow the FAIR principle F4 by "*registering or indexing the metadata in a searchable resource*". This indexing is the responsibility of the VP Index.

The VP Index works by being notified by willing FDPs of their existence and harvesting and indexing the available metadata from these metadata sources. We have adopted a distributed approach where the control and management of the metadata information remains at the source and the VP Index only keeps an updated index of these metadata records. In other words, the resource provider has total control over which information they would like to provide about their offered resources at any time. Once the connection is made between the participant FDP and the VP Index, the index is informed of changes in the metadata content in order to keep its metadata index up to date (refer to figure 1).

In order to provide this base level of metadata interoperability, the FDP provides:

- A common mechanism to access metadata about published digital objects following the HATEOAS REST approach;

- A common machine-actionable presentation format to express the provided metadata using W3C's RDF;

- A common approach to inform clients on how to navigate through the FDP's metadata structure using the container-member relations defined in the W3C's Linked Data Platform (LDP) specifications;

- A common representation format to express each metadata record's schema using W3C's Shapes Constraint Language (SHACL).

The specifications of the FDP can be found at https://specs.fairdatapoint.org, including examples of the metadata structure, navigation information and compliance criteria.

The main goal of the FDP is to establish a common method for metadata provisioning and accessing and, as a consequence, provide to client applications a predictable way of accessing and interacting with metadata content based on established standards.

The adopted strategy for the VP Index is to give the resource providers the freedom to choose how they will provide their metadata to the VP Index as long as it follows the minimal requirements of the FDP. To facilitate the onboarding, the EJP-RD program offered options ranging from simply filling an Excel sheet with the required information that is transformed through developed scripts into the FDP's RDF format, to deploying its own FDP with the required metadata schemas and filling the metadata entry forms, to using the FAIR-in-a-box or CDE-in-a-box packages provided by the project and using the provided FDP deployment. These options were detailed in the onboarding guidance provided to the resources in preparation for the Barcelona Hackathon in December 2022. This document contains a description of what the Virtual Platform is, its objectives, and the levels of connection available. The document also describes the technical options available for connecting to the Virtual Platform at levels 1 and 2; providing supporting material for implementation by the resource or by a service provider of their choice. The final version of the Onboarding the VP guidance document will be released in the EJP RD website.

During the 2-day onboarding hackathon in Barcelona in mid-December 2022, a number of resource providers engaged with the EJP-RD technology team to get their resources visible to the VP through the VP Index. This is explained further in section 4, along with other implementations at the Virtual Platform level.

D11.9
Third update Virtual platform of RD resources
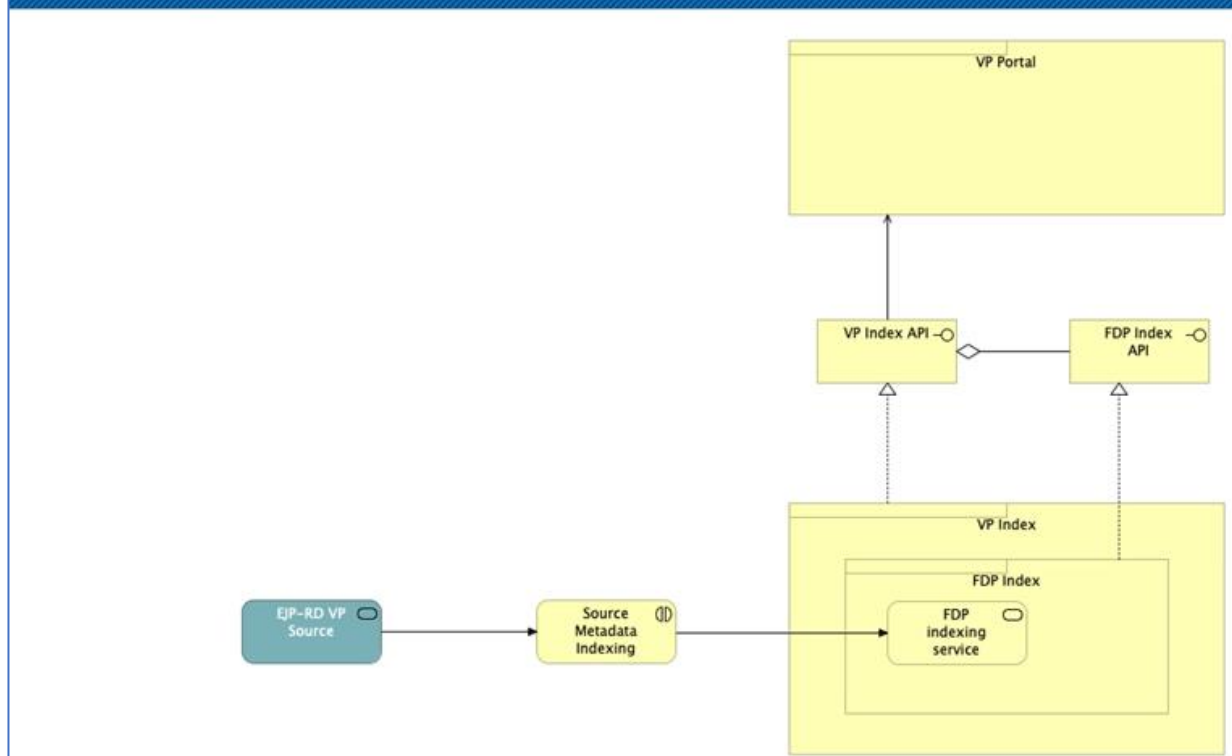annotated with EJP ontological model

Figure 1. Simplified Architecture of the VP Index

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

The Virtual Platform Portal (VP Portal) was further developed to enhance the execution of queries and the display of relevant the results. The key steps included:

- connecting the Virtual Platform to the VP Index described in section 3; which enabled the resources that are indexed to be dynamically discoverable via the VP Portal at level 1 (resources discovery through their metadata descriptors). The connection was established in December 2022 and will undergo minor interface improvements to ensure semantically-enhanced display of discoverable resources on the landing page of the Portal (see figure 2).

D11.9
Third update Virtual platform of RD resources
annotated with EJP ontological model

A collaborative platform for contributing and maintaining pathway information

ERKReg
The European Rare Kidney Disease Registry

VP Queryable

ERKReg, a Web-based registry for all patients with rare kidney diseases. The main objectives of this core registry are to generate epidemiological information, identify current patient cohort for clinical research, explore diagnostic and therapeutic management practices, and monitor treatment performance and patient's outcomes.

RD-Connect GPAP is a global genomics and clinical data platform for rare diseases: 1) a central and searchable hub for available hPSC-lines 2) a central registry of clinical studies based on hPSC-lines 3) a standardized identifier for hPSC-lines and register of these standard names 4) a trustworthy data source for hPSC-lines by verifying the ethical and biological conformity of registered lines based on community standards 5) comparability of quality standards in hPSC research 6) certification of quality and ethical provenance of hPSC-lines 7) a platform where any scientist or institution generating or working with PSC lines can register their cell lines, projects or clinical studies

Cellosaurus

VP Queryable

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

hPSCreg®

VP Queryable

hPSCreg is a global registry for human pluripotent stem cell lines (hPSC-lines). It aims to provide to the community 1) a central and searchable hub for available hPSC-lines 2) a central registry of clinical studies based on hPSC-lines 3) a standardized identifier for hPSC-lines and register of these standard names 4) a trustworthy data source for hPSC-lines by verifying the ethical and biological conformity of registered lines based on community standards 5) comparability of quality standards in hPSC research 6) certification of quality and ethical provenance of hPSC-lines 7) a platform where any scientist or institution generating or working with PSC lines can register their cell lines, projects or clinical studies

miRNome data from amniotic fluid samples of CAKUT and non-CAKUT pregnancies.

A metabolomic study of urinary changes in type 2 diabetes in human compared to the control group

EMT/MET cell line model to demonstrate that epithelial mesenchymal plasticity occurring in normal cells generates co-existing phenotypically and functionally divergent cell subpopulations with increased tumorigenic potential in vivo

The Duchenne Data Platform (DDP) is a patient-led registry. Patients are the owners of their data and have their own 'data lockers' where they can not only collect their Patient Reported Outcomes (PROs) but can also upload their data from other sources such as hospitals and wearables.

Figure 2. Display of Resources Connected to the VP via the VP Index

Based on the decisions taken in the Virtual Platform workshop series conducted by Pillar2 teams between September and December 2022, a modification of the API (https://github.com/ejp-rd-vp/vp-api-specs) at the back-end of the VP has started to ensure compliance with the new Beacon-2 framework specifications for Level2 discovery and query of resources and of records. The update has been partially implemented in the VP and the fully updated API will be ready in Quarter 1 of Year 5.

- An extensive review of search and filtering options has been done during the developers' workshop in Frankfurt in November 2022, which resulted in agreement

D11.9
Third update Virtual platform of RD resources
annotated with EJP ontological model

on the design of the following functionalities and implementing them using Orphanet services:

- o Hierarchical search of rare diseases; enabling the end user to search groups of diseases and expand the search upward or downward in the hierarchy using the Classification API of Orphanet.

- o Ability to search using other common disease codification systems: for Version 1 of the VP Portal, Orphanet mapping services will be used to enable search by ORPHAcodes, OMIM codes, and ICD-10 codes.

- o Search by gene functionality; which allows the end user to use a gene symbol in the search field resulting in a list of resources associated with that gene (specifically, associated with the diseases of ORPHAcodes mapped to those genes). The details of this gene-disease link are further described in the following section (section 5.2).

Throughout the execution of the functionalities, the metadata work focus closely collaborates with the Overall Architecture and Query Builder teams to ensure that the resulting design of the user interface accurately reflects the implementation of semantic approaches and the use of existing interoperable tools and services for the enhancement of the VP network of resources as a whole. Updated Orphanet's web services including Mapping service, genes-disorders service, hierarchies service and Orphanet's Registries & Biobanks API

Orphanet's catalogue of Registries and biobanks has been updated (twice a year) as well mappings services. We have setup a new service that allows to search by disorder concept annotated with genes concepts, and to search by genes annotated with disorder concepts, using ORPHAcodes. We have produced a service to ease the process of implementing the Orphanet's rare diseases hierarchies, which allows to perform queries retrieving disorders concepts belonging to Orphanet classification branches. The semantic ORPHAcodes mapping services was updated

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.com) and are integrated in ORDO (Orphanet Rare Diseases Ontology) twice a year. 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 (including Narrower-to-Broader and Broader-to-Narrower) with respect to the respective ORPHAcode(s).

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

Description of the service is available on GitHub: https://github.com/ejp-rd-vp/Orphanet-Mapping-API.

Useful links to services are provided in table 1 below.

D11.9
Third update Virtual platform of RD resources
annotated with EJP ontological model

Table 1. Orphanet Mapping and API Parameters

Element	Parameters & Link(s)	Details
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 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.
	/mapping?form={/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.
Querying using a different coding system	http://purl.org/orphanetws/mapping?from=icd&code=Q87.4 http://purl.org/orphanetws/mapping?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

4.1. Orphanet Hierarchies service

This Service dedicated to support implementation and usage of Orphanet's classifications. The API focuses on Rare Diseases classifications based on ORDO (Orphanet Rare Diseases Ontology)

A clinical entity (group of disorders, disorder or subtype of disorder), for which an ORPHAcode exist, could belong to several medical domains (classifications)

This service help to navigate from an ORPHAcode through the levels in the classifications' trees.

API parameters

{ORPHAcode} : an ORPHAcode of a clinical entity as a node to start with (Mandatory)

IRI or code are accepted.

e.g.: /traverse?code=http://www.orpha.net/ORDO/Orphanet_558

D11.9
Third update Virtual platform of RD resources
annotated with EJP ontological model

/traverse?code=558

{way} : from the node, path to levels. Expected 'up' (to "parents"), 'down' (to "childs") (optional. Without, retrieve +1/-1 from node) /traverse?code=558&way=down will return for each level, all the descendants of the specified Code

{level} : level to reach from the ORPHAcode as starting point (optional. Without, return all available levels)

{classif} : specify 1 to n Orphanet classifications ID to traverse (optional. Without, return results from all classifications)

Set of examples
   Endpoint: [http://155.133.131.171:8080/ClassifTraversal/hierarchies]
   Supported Classifications list: /ClassifTraversal/hierarchies/list
http://155.133.131.171:8080/ClassifTraversal/hierarchies/list

   e.g.: ORPHAcode 558 (Marfan syndrome)
http://155.133.131.171:8080/ClassifTraversal/hierarchies/traverse?code=http://www.orpha.net/ORDO/Orphanet_558
http://155.133.131.171:8080/ClassifTraversal/hierarchies/traverse?code=558

(Marfan Syndrome starting node, 2 classifications used)
http://155.133.131.171:8080/ClassifTraversal/hierarchies/traverse?code=558&classif=199&classif=189

(Marfan Syndrome starting node, way down)
http://155.133.131.171:8080/ClassifTraversal/hierarchies/traverse?code=558&way=down

(Marfan Syndrome starting node, all the ascendants of the specified Code.)
http://155.133.131.171:8080/ClassifTraversal/hierarchies/traverse?code=558&way=up

(Marfan Syndrome starting node, way up to 3 level maximum, through 2 classifications
http://155.133.131.171:8080/ClassifTraversal/hierarchies/traverse?code=558&way=up&level=3&classif=199&classif=189

This service is described in the according GitHub repository : https://github.com/ejp-rd-vp/orphanet_hierarchies_service

4.2. Orphanet Genes/Diseases Mapper API
This API focuses on Diseases and Genes related to them, Orphanet provides a link between genes and diseases related to them and this API implements that link.

Gene to disease(s)
The service allows query based on HGNC codes or gene symbol

API parameters

'gendis/find?by={hgnc code or Symbol}&input={the value of hgnc code or Symbol}
e.g.: [http://purl.org/gendis/find?by=hgnc&input=3603]
e.g.: [http://purl.org/gendis/find?by=symbol&input=FBN1]

D11.9
Third update Virtual platform of RD resources
annotated with EJP ontological model

You will obtain the list of diseases linked to the specific gene (by code or symbol)

Returns OrphaCodes and label:
    orphaCode
    "orphaCode":
    "http://www.orpha.net/ORDO/Orphanet_284963", "label":"Marfan syndrome type 1"
    "inferredResourceResponses":

In Orphanet Knowledge base, genes are linked to specific clinical entities such as "subtype of disorder". For instance, FBN1 gene will be linked to Marfan type 1 and TGFBR2 gene will be linked to Marfan type 2. However, resources are often annotated at the level of "disorder" clinical entities. (Marfan syndrome in this example)

The API will return a result expanded to this "upper" level (from subtype to disorder level) when relevant as "inferredResourceResponses" :
    "inferredResourceResponses":
    {"orphaCode":"http://www.orpha.net/ORDO/Orphanet_558", "label":"Marfan syndrome"}
    "orphanetEndpoint"

We added a feature that allows to query the Orphanet Resources Endpoint (Biobanks and Registries) based on the returned list of Orphacodes
    "orphanetEndpoint":
"http://155.133.131.171:8080/Orphanet/resource/search?code=300382&code=91387&code=2623&code=2833&code=284963&code=284979&code=3449&code=969&code=1885&code=558"

Disease (ORPHAcode) to Genes
This service allows query based on ORPHAcodes to retrieve genes linked, with HGCN code and Symbol.

API parameters
'gendis/find?by=orphacode&input={orphacode or IRI}
e.g.:
[http://155.133.131.171:8080/GENES/gendis/find?by=orphacode&input=http://www.orpha.net/ORDO/Orphanet_284963]
e.g.: [http://155.133.131.171:8080/GENES/gendis/find?by=orphacode&input=284963]

You will obtain the list of genes linked to the clinical entity orphacode concept, with HGNC code, symbol and "relation" between orphacode concept and each gene.
    directLinkedResponses hgnc "3603" symbol "FBN1" relation "Disease-causing germline mutation(s) in" inferredResponses [] apiVersion "v0.1"

    directLinkedResponses : the clinical entity concept (orphacode) is annotated with the gene in the Orphanet knowledge base
    inferredResponses : the clinical entity concept (orphacode) is not directly annotated with the gene but a subtype of the entity is.

Example: FBN1 gene is linked to Marfan type 1. TGFBR2 gene is linked to Marfan type 2. The Marfan syndrome (orpha:558) is not linked directly to any gene but has "Marfan type 1" and "Marfan type 2" as subtypes. Therefore, in order to obtain genes related to "Marfan Syndrome" we propose an inferred response including subtypes.

e.g.: [http://155.133.131.171:8080/GENES/gendis/find?by=orphacode&input=558]

D11.9
Third update Virtual platform of RD resources
annotated with EJP ontological model

"inferredResponses"
"orphaCode": "284963"
"label": "Marfan syndrome type 1"
linkedResponses
"hgnc":"3603", "symbol":"FBN1", "relation": "Disease-causing germline mutation(s) in"
"orphaCode":"284973"
"label": "Marfan syndrome type 2"
linkedResponses
"hgnc": "11773", "symbol": "TGFBR2", "relation": "Disease-causing germline mutation(s) in"

The service is described in GitHub (https://github.com/ejp-rd-vp/Orphanet-GenesDisease-Mapper-API)

4.3. Orphanet Registries & Biobanks API

The content was updated from the Orphanet's catalogue of Registries and Biobanks. We updated the service allowing to use directly different nomenclature codes, including ORPHAcodes. The service could be used using IRI from nomenclatures or value of codes. Results are returned in JSON format.

The service was initially developed following the specifications of Query Builder previous version (https://github.com/ejp-rd-vp/query_builder_api) but will be reshaped accordingly to the updated Beacon-2 framework-based API specification.

Table 2. Sample Queries of Orphanet Registries and Biobanks

Coding System	Example query
ORPHAcode	http://155.133.131.171:8080/Orphanet/resource/search?code=http://www.orpha.net/ORDO/Orphanet_558
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

D11.9
Third update Virtual platform of RD resources
annotated with EJP ontological model

Figure 3. Example of output of Orphanet catalogue API, ICD-10 code as input parameter

D11.9
Third update Virtual platform of RD resources
annotated with EJP ontological model

5. Discussion and Next Steps

As more resources are addressed, to be part of the VP ecosystem, we face new challenges.

Firstly, the adaptation to the new VP index infrastructure, based on FAIR Data Point which needs for each resource to provide their own FDP or rely on a proposed third party to host their metadata in a shared FDP. This step is mandatory to be at least listed as an accessible and discoverable resource in the VP ecosystem (Level 1 of connexion in the VP) A set of methodologies are therefore proposed to facilitate this process, which is necessary to have an adaptive, evolving and scalable VP.

In Y4 the VP API specifications were modified (https://github.com/ejp-rd-vp/vp-api-specs) to use a GA4GH Beacon-2 API framework for Level 2 record-level queries from the VP Portal and for cross-queries in the VP network. The main objective is to propose a single merged API mechanism to be implemented and presented into the VIPS document, in order to avoid several layers of different APIs within the VP ecosystem. This will need adaptation also for some RD resources but moreover it needs to reshape the backend and partially the front end of the VP Portal along the process.

From an organisational point of view, in order to accelerate progress in the implementation phase of the project, a dedicated user interface working group was set up to design the new user interface: the EJP RD VP Portal.

The VP Portal also gained an important status within the project as it reflects a reference implementation process for any other "portal" which aims to be connected to the VP ecosystem. Therefore, during the last phase of the project a dedicated joint effort must be pursued to ensure agility and quality in the development process, as a too limited manpower won't be able to manage both backend and frontend functionalities.

Finally, as stated and started during the Barcelona Resource Onboarding Hackathon (December 2022), in the final stage, a dedicated effort to accompany the resources to be onboarded must be pursued as well. Along the process, we will also adapt accordingly the DCAT-based EJP RD model as well as to the record-level data model, to reflect the diversity of resources, including metadata related to legal statement, licence, and access rights, both for resources, distribution and accessibility of datasets.