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 12.4
Third Report on core set of FAIR software tools and on extended set of unified FAIR data standards applied in EJP RD

Organisation name of lead beneficiary for this deliverable:

Partner 64 – LUMC

Due date of deliverable: month 52

Dissemination level:
Public

Table of Contents

1. Scope

The Virtual Platform (VP) is a network of resources that have been prepared for automated applications across multiple resources, with access points for a wide range of user types: from clinician to computational scientists. A small number of 'VP-wide' components facilitate automation of communication between the resources. The final instalment of this series of reports on *the core set of FAIR software tools and extended set of unified FAIR data standards applied in the EJP RD* provides an overview of the software standards and tools that have been applied as building blocks for the VP. The scope of this deliverable is the minimized subset of standards and tools that have been tested by EJP RD partners as core building blocks for creating a functioning and extendable VP. The predecessors Deliverables D12.03 ([1]) and D12.02 ([2]) serve as reference to standards and tools that are used by members of the EJP RD community.

1.1. The core set of Standards and Tools utilized by onboarded resources on the VP, at three incremental levels

Three levels have been defined for Resource Providers to progressively add functionality to the VP, a process referred to as 'onboarding'. The levels pose incremental requirements for the application of standards and tools.

1. **Resource Discovery**
   Resources enable automated discovery by providing a minimal description in a standard, machine actionable way conforming to the FAIR Data Point (FDP) Specifications ([3], [4]). Resources thus guarantee that the metadata that describes them is at least provisioned in terms of the 'Data Catalog Vocabulary' (DCAT) and EJP RD recommended extensions of DCAT [5], navigable via Linked Data Platform predicates, and using the global Linked Data model RDF as the underlying data model. This metadata can be indexed by VP resource indexes. Indexes that are themselves a FDP provide a standard access point for computationally finding resources in the VP network. VP networks are advised to have at least one such governed index, maintained by a community partner. The EJP RD provided an index for the production deployment of the Virtual Platform. A second index was created on the "testbed" server, for pre-deployment testing of new resources and tools, prior to their deployment on the primary public Virtual Platform. This helps ensure that Resources discovered on the Virtual Platform are of high maturity and quality.
2. **Discovery by content**
   In addition to level 1, resources enable the discovery of their resource by content through providing additional metadata and implementing a common API conform to the GA4GH Beacon-2 framework ([6]). ERN RD Registries

provided a paradigm for discovery by data elements. They collect Common Data Elements conforming to a recommendation by the Joint Research Centre of the European Commission (Section 5.1.3.2; [7]). In addition, the EJP RD has developed a common informed consent form for RD registries and common consent elements with tools for authoring consent documents. The aim is to promote the harmonisation of access conditions and their provisioning conform to a machine-readable representation.

Data analysis in the VP Laboratories

In addition to levels 1 and 2, resources enable automated analysis over multiple resources by providing additional metadata via their FDP. Metadata about their access protocols, access conditions, the structure and semantics of their data, and functions that a resource can execute. Rich metadata provided in a standard way accommodates the diversity of data types and access protocols and conditions at this level of the VP network. This pertains to using DCAT DataService metadata standards, defined ontological annotations, and use of the openAPI standard for interface descriptions. Because of the diversity at this level, the VP onboarding requirements must expand incrementally, case-by-case, through co-creation (see section 1.2 and chapter 6). The EJP RD delivered the example of RD registries that provide a representation of the collected JRC Common Data Elements in terms of the Clinical and Registry Entries Semantic Model (CARE SM; section 5.1.4.1; [8], [9]), which in turn ensures that standards such as the Orphanet Rare Disease Ontology and ORPHACodes are used in a consistent and predictable way. While CARE-SM was primarily designed to represent clinical observations, it is under constant expansion and is capable of representing genetic and (to a lesser extent) genomic information. Nevertheless, none of the decisions made by the VP Labs limits the standards that can be used by the VP, and already there are nascent examples of using the VP for genotype/phenotype analyses independently of the CARE-SM model. The module can serve as a reference for other types of data by partially reusing the module, or by reusing the approach of building on existing semantic design patterns.

1.2. The core set of Standards and Tools for federated analysis and evolution of the VP: The VP Laboratories

Chapter 6 introduces standards and tools for federated data analysis and the evolution of the VP through interdisciplinary co-creation. The VP is designed to accommodate a functional, dynamic ecosystem of heterogeneous resources that evolves over time as more types of resources and more subcommunities connect. Moreover, it must accommodate that many of its parts, including data sources, VP-wide components such as the resource index, and access governance, are subject to active ongoing development, often at a global level. Therefore, the VP that the EJP RD delivers has been prepared for evolution through co-creation between providers of new resources and the architects of the VP overall architecture. The concept of *the VP Laboratories* (VP Labs) is of central importance. A VP lab represents an experimental version of a future Virtual Platform, or specific elements thereof, with the objective of exploring and testing potential enhancements and modifications to the existing VP network. In the final phase of the EJP RD this concept has been applied to

FAIR standards and tools that support the federated analysis of data records across multiple FAIR resources, thereby establishing the prerequisites of Level 3 onboarding. The evolution of the VP at this level depends on an experimental environment that enables the learning of how to address the complexity of data access and data interoperability challenges across a wide range of stakeholder communities.

2. Sources

This report builds on the previous deliverables and their sources (see [1] and references therein), and the extensive work done on onboarding guidance [10] and development of support for the three levels of onboarding in the final year and extension of the project.

3. Development status

In this document, a distinction is made between standards and tools that have been tested as normative core standards for automatic and programmatic identification and discovery of resources in the VP network (chapter 5), and standards and tools that are subject to trial use and community co-development in a VP Lab (chapter 6). These map to the 'Normative' and 'Trial use' development status introduced previously (section 3 in [1]).

4. Dissemination Format

Multiple methods for disseminating FAIR standards and tools have previously been provided ([1]). This report provides references with the descriptions of individual standards and tools in the sections below.

In line with the focus of this report on the minimal set of standards and tools to connect a resource to the VP network, we recommend the onboarding manual provided through GitHub as main starting point: see the EJP RD Virtual Platform: Resources onboarding manual [10]. This 'living document' is updated as the VP evolves. At the time of writing the manual focusses on onboarding at levels 1 and 2.

As a reference for preparing for the VP Labs to explore standards and tools that enable federated analysis of data records (level 3), and for co-creating future versions of the VP, we additionally refer developers to the EJP RD organisation main page on GitHub as a starting point [11].

5. The core set of Standards and Tools utilised by resources contributing to the VP network

5.1. Standards

5.1.1. The EJP RD Metadata Model: the common vocabulary to describe resources in the VP network

Providing a minimal description of a resource in terms of a common vocabulary that computers can process is a prerequisite for automated communication and discovery between resources in the VP. Providing metadata in terms of the *EJP RD Metadata Model* is therefore the entry level to VP compliance ([10]; sections 5.1.1.1-3 in [1]): by definition, all resources on the VP expose some information about themselves at least

in terms of this common vocabulary (e.g. patient registries, molecular and disease knowledge bases, and catalogues of biosamples, model systems, cell lines, but also bioinformatics tools, analysis workflows, and clinical research support tools such as the Clinical Trials Toolbox developed by the EJP RD). The process of providing descriptions of a resource, or in other words, 'annotating' information about the resource in accordance with the metadata model, is also referred to as 'instantiating' or 'populating' the metadata model. For example, the vocabulary term 'Theme' is instantiated (populated) with a specific Orphanet Rare Disease Ontology identifier to express the main theme of a disease registry database. A minimum set of information necessary for basic VP operations is defined in section 5.1.3.1).

- Further Information: ejp-rd-vp/resource-metadata-schema: Metadata model and schemas for the EJP virtual platform (https://github.com/ejp-rd-vp/resource-metadata-schema)

5.1.1.1. The EJP RD Metadata Model extends the DCAT vocabulary

The EJP RD Metadata Model extends the W3C-recommended Data Catalog Vocabulary, version 2 [12] (DCAT2; also see section **Error! Reference source not found.**). The world-wide use of DCAT across a wide variety of domains provides rare disease research a basis for automated discovery across the widest possible range of resources. DCAT is designed to be extended. Subclasses and sub properties were added for resource types that are specifically relevant for the VP (see section 5.1.1). DCAT2, its extensions and its instances are defined in the RDF data model [13], whereby the classes and properties that denote instances of metadata concepts and the relations between these instances are globally unique and referenceable identifiers. At a technical level this means that resources are *virtually linked* to other resources in the network through the identifiers of common metadata elements (terms, properties, instances), conform to the 'Linked Data' principle [14]. It is possible to serialise metadata in RDF in a variety of exchange formats (e.g. XML, JSON), although this may result in a loss of machine-readable semantics. A detailed description of the EJP RD Metadata Model and the ontologies that it builds on can be found on Github [5].

5.1.1.2. Classes denoting the resource type in support of automated discovery

The following classes are used to describe the initial types of resources populating the current VP:

- Patient Registry – this extension denotes that a resource is an instance of a patient registry
- Biobank – this extension denotes that a resource is an instance of a biobank
- Guideline – this extension denotes that a resources is an instance of a guideline
- Dataset – This core component of DCAT is used when the extensions above do not apply, and simply denotes that a resource is a data set.

all of these are subclasses of dcat:Resource. Instances of these subclasses of this class denote that a digital resource is of a specific type, implying it will include properties that are specific for that type, such as geographical distribution for biobanks and

registries[1]. Applications can use the additional precision, for instance to filter or classify specific types of resources in automated workflows.

5.1.1.3. Additional attributes used in resource descriptions in support of automated discovery

The full list of metadata classes and properties that describe a resource on the VP for automated applications can be found online [5], [15]. The following properties were added to support discovery:

- vCard-based contact information: resources on the VP denote contact information by classes defined in the vCard-rdf model that was developed by the W3C Semantic Web Interest Group [16]. The model covers a subset of the IETF vCard file format specification [17].
- Beacon-2 Discovery services: resources in the VP network are programmatically discoverable by minimally offering search operations conforming to the GA4GH Beacon-2 framework [18] and denoting this by instantiating a dcat:DataService with the *dcterms:type* property set to 'ejp:VP_Beacon2_<target>'. In the final release of the VP by the EJP RD '<target>' can be 'catalog' (searches on the description of a resource) and 'individuals' (searches on the types and sometimes values of data records that a resource gives access to under defined conditions). The set of targets is extended in the VP labs, starting with the resource types of section 5.1.1.2.
- VP Connection Governance: on the VP, resources inform clients which sections of metadata can be used for VP-wide discovery operations initiated from the VP Portal user interface. They do so by annotating an instance of a metadata subject (e.g. an instance of a subclass of dcat:Resource) with a 'vpConnection' property from the EJP RD Metadata model. In the final EJP RD release of the VP the defined value for that property is limited to 'VPDiscoverable'. Thus, the existence of this "tag" indicates to the VP that the resource is willing to be indexed and displayed on the VP. The tags are *informative*: by themselves they do not restrict metadata use outside of the Virtual Platform.

Beyond the Beacon2 discovery services, participants in the VP network may provide a wide range of data discovery or analytics services[2]. The rules around publishing these on the VP are:

- All operations are defined as an instance of dcat:DataService. The properties of dcat:DataService specify the type of the operation, and provide a pointer to the document containing the instructions for mechanized access to the tool.}
- Specifically, while Beacon is described using ontology terms minted by the EJP-RD project (i.e. Beacon2_individuals and Beacon2_catalog) other kinds of operations must be identified by an ontological type. The VP recommends that providers describe their operations in terms of the EDAM ontology or subclasses

[1] The specific properties that define a subclass are not always made explicit in a schema. Under the open world assumption of web ontologies, they can be implied or defined later.

[2] Not all services provided by resources to applications of the VP network are accessible via the Virtual Platform Portal, which the EJP RD has developed as a reference end-user interface portal. Through co-creation with service providers (Chapter 6), the functionality of this portal can be extended or other interfaces created to meet the needs of a target user community.

thereof [19], [20]. One of the children of the Operation class of the EDAM ontology can be used or extended to denote the type of operation (http://edamontology.org/operation_0004). For example, the Beacon-2 search operations are defined as subclasses of the database search operation class (http://edamontology.org/operation_2421).

5.1.1.4. Other metadata standards

The VP defines several metadata requirements that constrain what metadata features *must* be minimally present for inclusion in the VP Network (section 5.1.1 in [1]). There are a wide variety of other metadata standards that will co-exist with the VP requirements, and these are by no means excluded by the VP. The VP requirements simply allow the VP software to automatically discover appropriate resources, without the need to map between other standards and those selected by the VP (see chapter 6 for examples).

5.1.2. Standards on data file formats, markup, and annotation

Resources on the VP describe the standards that they use for data file formats, markup, and annotation schemas as part of the metadata that describes the resource. Specifically, the resource declares the availability of different distributions by instances dcat:Distribution with the dcat:mediatype property set to a globally unique identifier denoting the file format [21]. An example is 'https://www.iana.org/assignments/media-types/text/csv'. Henceforth, resources on the VP are not restricted to any specific file format, markup or annotation schema for their contents, although it is advised to use formats that are commonly used in an appropriate domain (e.g. the VCF file format in the domain of variant calling). Formats for various commonly used data types have been described previously in Deliverable D12.03 (section 5.1.2 in [1]).

Two exceptions pertain to general metadata and VP-wide resource discovery: (i) the general metadata that describes the resource is provisioned conform to the FAIR Data Point (FDP) specifications in RDF (section 5.1.6.1, [22]), (ii) the services that enable programmatic discovery conform to the GA4GH Beacon-2 framework, which includes a JSON -based API and schema for data exchange [18]. This combination ensures that computational agents have the means to discover how to interact with resources in the VP network.

5.1.3. Standards on metadata and data element sets

5.1.3.1. CME: Common Metadata Elements for resource discovery

Applying the EJP RD Metadata Model guarantees that for every resource on the VP, a description is available at least in terms of an agreed, standardised and machine actionable vocabulary (Section 5.1.1). However, successful discovery further depends on what information is minimally in these descriptions. Therefore, a minimal set of metadata elements was defined as essential for successfully running basic computational select and filter methods. 'Common Metadata Elements' have been defined within the following categories, where each category has mandatory, recommended, and optional properties to describe and categorize resources accurately:

- Organization (defines the ones providing resources like biobanks, patient registries, datasets, or data services)
- Patient Registries (describes patient registries in the context of rare disease resources)
- Biobanks (defines biobanks related to rare disease resources)
- Datasets (used to describe any rare disease-related dataset)
- Distribution (describes a specific representation of a dataset (e.g., csv and json)
- Catalogue (a bundle of numerous datasets, data services, biobanks, patient registries, or guidelines from a specific organization)
- Data Service (resources that provide access to data or analytical tools via some interface)
- Guideline (describes guidelines associated with the resource, such as "Biomarker Development Manual.")

Examples of common mandatory properties to all these categories include *title* and *description*. The properties *theme, keyword, publisher* and *contact point* are mandatory to most categories (i.e., patient registries, biobanks, datasets, catalogue, data service and guideline), while some properties are mandatory only to certain categories (e.g., population coverage – a property of specific for patient registries and biobanks). Although not mandatory, some properties are recommended, and others are optional. An example of recommended property includes *access rights* (recommended for distribution, guideline, biobanks, catalogue, dataset, patient registries and data services). It should be noted that different applications may have different requirements as to what is necessary and optional. For instance, the logo property may be necessary for visualisation in a user interface such as the EJP RD VP Portal, but it is optional for computational workflows.

- Further Information: https://vp-onboarding-doc.readthedocs.io/en/latest/level_1/properties.html

5.1.3.2. CDE - Common Data Elements for patient registries

Registries on the VP that collect observational data from patients facilitate automated applications across multiple registries if they collect a common set of data elements, and declare this in their metadata. The European Reference Networks of rare disease expert centres have been advised by the European Commission to collect at least the 16 Common Data Elements that were defined by the European Joint Research Centre [7]. The EJP RD followed this recommendation in building its support for ERNs, reflected in a semantic model (section 5.1.4.1) and tools (section 0). Also see section 5.2 in [1].

- Further Information: https://eu-rd-platform.jrc.ec.europa.eu/set-of-common-data-elements_en

5.1.4. Standards on data models

Resources on the VP extend discovery capabilities by providing data models that describe the data that they offer for research. Several data models exist (section 5.1.4 in [1]). For instance, rare disease registries on the VP network referring to the CARE-SM

model express that they provide Common Data Elements[3] in machine actionable terms (see section 5.1.4.1). Providing a reference to a data model that represents the resources' data in a standard way allows automated selection and filtering on data models. For example, the implementation of the Beacon-2 based discovery API that is part of FAIR-in-a-Box makes use of this [23]. Incorporating data model information for discovery in the metadata of a resource follows that of data formats (section 5.1.2). The method for federated analysis is described in section **Error! Reference source not found.**.

5.1.4.1. The Clinical And Registry Entries (CARE) Semantic Model

The Clinical And Registry Entries Semantic Model (CARE-SM) is a healthcare-based data model developed to standardize and enhance the representation of patient information stored in data registries ([9]). Leveraging the advanced capabilities of Semantic Web technologies and Linked data, CARE-SM is designed to facilitate interoperability over distributed or federated data sources.

CARE-SM evolved from the Common Data Element Semantic Model (CDE-SM), originally created to represent a set of concrete and standardized data elements, particularly in the context of rare disease registries as part of European Commission projects ([24]). However, as the need arose to model a broader and more complex array of healthcare data—such as treatments, interventions, imaging, and patient-reported outcomes—CARE-SM was developed to expand and improve upon the foundational CDE-SM. The new model maintains several of the original objectives of the data architecture, followed by model standardization and expansions.

CARE-SM integrates the SemanticScience Integrated Ontology (SIO; [25]) and the Open Biological and Biomedical Ontology (OBO; [26]) Foundry's terms to describe for upper-class definition and domain-specific data elements respectively, ensuring that the model is both precise and adaptable to various types of healthcare data. The use of standardized ontologies ensures that the data model is consistent and supports interoperability, making it easier for researchers and clinicians to transform, query, and analyse data across different systems.

CARE-SM presents a core entity-relationship structure. In the original CDE model, the relationships between entities were limited, primarily focusing on identifiers, roles, processes, and outputs. CARE-SM significantly extends this framework to include additional entities such as inputs, agents, targets, units and protocols, allowing a more comprehensive representation of clinical procedures and observations.

CARE-SM describes a contextual metadata layer, which provides the context for each data element, including temporal and event-based information. This layer allows the model to represent timelines of patient events and encounters, offering a richer and more detailed view of patient history information. For instance, the model can capture the sequence of clinical encounters, treatments, and outcomes, linking them together under a common clinical episode or event identifier. The contextual metadata layer also leverages RDF-Quads, an extension of the standard RDF (Resource Description Framework) triples, to include a fourth element—a context URI. This allows additional

[3] CARE-SM has been extended to cover more than the JRC CDEs and can be further extended.

metadata, such as temporal or administrative information, to be associated with each data element.

The CARE-SM implementation process involves generating common CSV templates for different data types, which are then transformed using YARRRML templates and conditionals to produce the final RDF representation. These components are integrated into a larger data transformation tool called FAIR-in-a-Box (FiaB; [23]). A video tutorial for CARE-SM is available ([8]).

5.1.5. Standards on data ontology, terminology and vocabulary

Resources on the VP have increased their usability and interoperability by applying concepts from community-agreed semantic models (ontologies, terminologies, and vocabularies) to denote the types and relations of the values of data elements in a set and in some cases to provide the values of qualitative data elements. Many semantic models exist, together covering most of the wide range of data types in biomedicine. Furthermore, most web ontology-based data models are constructed from multiple commonly used ontologies. The concepts have globally unique identifiers that are linked to the values in the data sets. Data sets that have some elements in common will also have ontology concepts in common. This is true in communities that converge on the ontologies that they use. Otherwise, there are semantic models to precisely define the semantic relation between related concepts in different ontologies ([27]), which provides a basis for mapping services in the VP Labs (see section 6.2.3). Henceforth, diverging data sets on the VP become linked via common identifiers from common ontologies (e.g. the identifier of 'ncit:Diagnosis Code' from the National Cancer Institute Thesaurus (OBO version) as the type of a value and the identifier for the disease code from the Orphanet Rare Disease Ontology as the value of that type virtually link resources that also use these ontologies to capture diagnosis information).

In this section we provide a table of the ontologies, terminologies, and vocabularies that are incorporated in the ontological models that resources minimally use in the VP: the EJP RD Metadata Model and CARE SM (Table 1). This pertains to concepts for types and relations, and in some cases to coding systems where ontological concepts are used as values. More information can be found via the provided references and in deliverable D12.03 ([1]). Resources on the VP and the VP Labs will often use additional ontologies, terminologies, and vocabularies for specific needs (e.g. see section in 5.1.5 [1]). Which ones are used is accessible in the resources' machine actionable metadata.

Table 1 – ontologies and terminologies used in data and metadata models of resources on the VP

Acronym	Name	References[1]	Incorporated in	Value or Concept[2]
DCAT-2	Data Catalog Vocabulary, version 2	• https://www.w3.org/TR/vocab-dcat-2/	EJP RD Metadata Model	[✓] Types [✗] Values
DCTerms	Dublin Core Metadata Initiative Metadata Terms	• https://www.dublincore.org/specifications/dublin-core/dcmi-terms/	EJP RD Metadata Model	[✓] Types [✗] Values

DUO	Data Use Ontology	• https://obofoundry.org/ontology/duo.html • https://www.ga4gh.org/product/data-use-ontology-duo/ • https://github.com/EBISPOT/DUO	EJP RD Metadata Model[3]	[✔] Types [✔] Values
EDAM	Ontology of bioscientific data analysis and data management	• https://edamontology.org/ • https://zenodo.org/records/3899895	EJP RD Metadata Model	[✔] Types [✘] Values
GENO	Genotype Ontology	• https://www.ebi.ac.uk/ols/ontologies/geno	CARE SM	[✔] Types [✘] Values
HGVS	Human Genome Variation Society Nomenclature	• http://varnomen.hgvs.org	CARE SM	[✘] Types [✔] Values
HPO[4]	Human Phenotype Ontology	• http://purl.obolibrary.org/obo/hp/hp-international.owl https://hpo.jax.org/ https://doi.org/10.1016/j.ajhg.2008.09.017 • [28]	CARE SM	[✔] Types [✔] Values
ICD codes	International Classification of Diseases codes	• https://www.who.int/standards/classifications/classification-of-diseases	CARE SM	[✘] Types [✔] Values
ICF	International Classification of Functioning, Disability and Health	• https://www.who.int/standards/classifications/international-classification-of-functioning-disability-and-health	CARE SM	[✘] Types [✔] Values
NCIt OBO edition	National Cancer Institute Thesaurus	• https://github.com/NCI-Thesaurus/thesaurus-obo-edition • http://purl.obolibrary.org/obo/ncit.owl	CARE SM	[✔] Types [✔] Values
OBI	Ontology for Biomedical Investigations	• https://obi-ontology.org/	CARE SM	[✔] Types [✘] Values
OBIB	An ontology built for annotation and modeling of biobank repository and biobanking administration	• https://obofoundry.org/ontology/obib.html	CARE SM	[✔] Types [✘] Values
OMIM	Online Mendelian Inheritance in Man	• https://www.omim.org/	CARE SM	[✘] Types [✔] Values
ORDO[4,5]	Orphanet Rare Disease Ontology	• https://www.orphadata.com/ordo/ http://www.orpha.net/ORDO	CARE SM	[✔] Types [✔] Values

ORPHAcodes	Orphanet nomenclature of rare diseases	• https://www.orphadata.com/orphanet-nomenclature-for-coding/ • [29]	CARE SM	[✗] Types [✓] Values
OWL	Web Ontology Language	• https://www.w3.org/OWL/	EJP RD Metadata Model CARE SM Semantic Phenopackets	[✓] Types [✗] Values
RDF	Resource Description Framework	• https://www.w3.org/RDF/	EJP RD Metadata Model CARE SM Semantic Phenopackets	[✓] Types [✗] Values
RDFS	RDF Schema	• https://www.w3.org/TR/rdf12-schema/	EJP RD Metadata Model CARE SM	[✓] Types [✗] Values
RefSeq accession numbers	NCBI Reference Sequence Accession Numbers	• https://support.nlm.nih.gov/knowledgebase/article/KA-03437/en-us	CARE SM	[✗] Types [✓] Values
SKOS	Simple Knowledge Organization System	• https://www.w3.org/2004/02/skos/ https://www.w3.org/TR/2009/REC-skos-reference-20090818/ • https://www.w3.org/2004/02/skos/vocabs	EJP RD Metadata Model	[✓] Types [✗] Values
SIO	The Semanticscience Integrated Ontology	• https://github.com/MaastrichtU-IDS/semanticscience http://semanticscience.org/resource/SIO_010043	CARE SM	[✓] Types [✗] Values
UO	Unit Ontology	• https://obofoundry.org/ontology/uo.html	CARE SM	[✓] Types [✗] Values
WikiData	WikiData	• https://www.wikidata.org/	CARE SM	[✗] Types [✓] Values

1. Several of the ontologies in the table incorporate other ontologies or mappings to other ontologies.
2. Ontologies are used in data sets in two ways: (i) to denote the types of data elements (classes) in a dataset and relations between them (properties); (ii) as values of variables (effectively, as "tags"). This is indicated by 'Types' and 'Values' respectively. The former relates to semantic models that represent the meaning of the elements in a data set. Using ontology terms only as values does not make a data set machine actionable. That requires ontological types and relations.
3. DUO in the EJP RD metadata model is a placeholder for populating the EJP RD metadata model with the specific access conditions of specific content of a resource (as part of the full metadata of the resource).
4. Because of the close relationship between clinical entities covered by ORDO and phenotypic abnormalities covered by HPO according to the frequency of occurrence in the disease, Orphanet also provides an ontological module of qualified relations between HPO and ORDO entities (see paragraph 5.1.5.9 in [1], and http://www.orphadata.org/cgi-bin/index.php#hoommodal)
5. ORDO includes alignments between ORPHAcodes and medical terminologies such as ICD-10, ICD-11, OMIM, MedDRA, UMLS, and MeSH. The alignments are maintained and curated by Orphanet.

5.1.6. Standards on FAIR data communication mechanisms

Resources on the VP can, by definition, provision a defined minimal set of metadata conform to the FAIR Data Point specifications for automated navigation (section 5.1.6.1). They can offer additional mechanisms of communication or exchange of metadata or data by including a reference in their metadata (e.g. via dcat:Distribution and dcat:DataService). A selection of commonly used exchange mechanisms (DOIP, FHIR, ISO /AWI TR 24305, mzML/mzIdentML, Phenopackets, REST) was described previously (section 5.1.7 in [1]).

On the VP Labs various mechanisms for exchange and communication are explored, including one where resources are 'visited' by an agent encompassing an algorithm and a description of that algorithm and information pertaining to access permissions (also known as 'FAIR Data Train'). In data visiting scenarios data stay where they are: data 'exchange' is limited to questions and answers. The messages are captured in RDF (see section 6.1.5.1).

5.1.6.1. FAIR Data Point specifications

The FAIR Data Point (FDP) is a metadata service that provides access to metadata following the FAIR principles. FDP uses a REST API for creating, storing and serving FAIR metadata. In order to provide metadata according to the FDP specifications [4], the following characteristics should be present:

- The root API URL must provide the metadata of the FAIR Data Point;
- The metadata content must be presented in, at least, RDF Turtle syntax. Other formats such as XML, JSON and other RDF syntaxes are allowed through content negotiation but the default media type for the HTTP9S) request must be RDF Turtle;
- The information about how to navigate the metadata content structure of the FDP must be provided in each metadata record using the Linked Data Platform (LDP) containment structure (see the FDP specs for examples [4]).

5.1.6.2. EJP RD Discovery API specifications based on the GA4GH Beacon-2 standard

Resources on the VP that provide the core discovery function for programmers, minimally implemented the EJP RD VP API specifications [30], [31]. Resources are recommended to implement the latest version, currently version 4 [31]. The API specification is defined complying with the latest Beacon v2 Specification of the Global Alliance for Genomics and Health [18]

5.1.7. Standards on security, authentication, and authorisation

In the federated VP network, the resources are responsible for the security of their data. Recommendations on authentication, authorization and governance are described in subsections 5.1.7.1 and **Error! Reference source not found.**. Furthermore, FAIR resources on the VP network minimally provide via their FAIR Data Point information to the VP network of their licenses, access protocols and access conditions in terms of the EJP RD metadata model. For instance, a recommended metadata property for data distributions is *dcat:accessRights*: a URL pointing to the location of information about who can access the resource or an indication of its security status. It is strongly recommended to point to a Data Use Conditions profile (also see section 6.1.2 and 6.1.2). The metadata property odrl:hasPolicy can furthermore be used to

provide a policy document conforming to the W3C recommended Open Digital Rights Language (ODRL; see section 6.1.2.1). All machine readable metadata pertaining to access is *informative* and does not in any way actively regulate access to the resource.

5.1.7.1. Life Science Login

Resources on the VP offering services that require authentication and authorisation minimally implement the Life Science Login (LS Login) protocol ([32], [33]), also known as Life Science Authentication and Authorization Infrastructure (LS AAI). The LS Login protocol requires information by which to authenticate and authorise users [34]. The VP Portal reference implementation implemented the Authentication protocol of the LS Login to provide authentication details of logged-in users to the resources. Resources on the VP Labs can apply various security mechanisms in automated analysis scenarios, including LS Login (see section 5.1.8 in [1] for four complementary mechanisms).

5.1.7.2. Governance of public data services

An alternative method is to provide a narrowly defined data service publicly (denoted by dcat:DataService) that has been vetted as 'safe to run' through a governance process. This process involves the authorities that are responsible for the data sharing policy of the resource (the data controllers). An example is the calculation of the time to diagnosis that the data controllers of a Duchenne patient registry agreed to offer as a public service. Data controllers can opt for this approach to offer services that do not require security.

5.2. Tools

A resource contributes to the VP if it minimally implements the standards described in section 5.1 (see the onboarding documentation for the latest updates [10]). Resources can provide and explore additional functionalities by implementing additional standards in the VP Labs (chapter 6).

5.2.1. Tools delivered by the EJP RD to help create a functional VP network

Deliverable D12.03 already listed a number of tools that resources can use to make their resource part of the VP or the VP Labs ([1]). Here, we tabulate the core tools that the EJP RD project delivered to help resources implement the core standards defined in section 5.1.

Table 2 – tools supporting the implementation of the standards to be part of the VP network, developed by or in conjunction with the EJP RD

Tool name	Description	References	Applies to … standard	
Deposition services with FDP and Beacon-2 functionality	Deposition services that implemented the core standards where appropriate (e.g. GPAP, WikiPathways, ERN	• https://index.vp.ejprarediseases.org/ (FDP index of resources with FDPs) • https://vp.ejprarediseases.org/discovery/sources	• EJP RD Metadata model • CARE SM (where appropriate) • Common Conditions of Use	

Tool name	Description	References	Applies to … standard	
	registries). Depositing data in these resources is a means to contribute a data set to the VP[1].	(user friendly list of onboarded resources)		
Document-based FDP	"Barebone" implementation of the metadata provisioning service conform to the FAIR Data Point specifications	• https://github.com/StaticFDP/static-fdp/blob/main/UserDocumentation.md • https://github.com/StaticFDP/fdpCloud • [35] • [4] • [3]	• EJP RD Metadata Model • FAIR Data Point specifications	
DUC Profiler	Tool to define a Data Use Condition (DUC) Profile using Common Conditions of Use	• https://duc.le.ac.uk • [36]	• Data Use Conditions and Common Conditions of Use (Section 6.1.2)	
FAIR Data Point Populator	Describe a resource by populating the EJP RD Metadata Model within a FAIR Data Point via github actions	• https://github.com/ejp-rd-vp/fdp-populator • [37]	• EJP RD Metadata Model • Metadata provisioning service conform to FAIR Data Point specifications	
FAIR Data Point index	Index of onboarded resources, i.e. resources that provide the minimally required description via a FAIR Data Point. The index itself is a FAIR Data Point.	• https://github.com/FAIRDataTeam/FAIRDataPoint (search for 'index' for the index related branches) • https://index.vp.ejprarediseases.org/ [2] • https://vp.ejprarediseases.org/discovery/sources	• EJP RD Metadata Model • Common Metadata Elements • FAIR Data Point specifications	
FDP Reference implementation	Docker images for deployment of a reference implementation of the metadata provisioning service conform to the FAIR Data Point specifications, with User Interface, query interface, REST API, editing functions, and RDF triple store backend.	• https://github.com/FAIRDataTeam/FAIRDataPoint • [38] • [4] • [3]	• EJP RD Metadata Model • FAIR Data Point specifications	

Tool name	Description	References	Applies to … standard	
FAIR evaluator	Tool that assembles automated tests of individual FAIR principles and applies them to a digital resource.	• Section 5.2.2 • https://w3id.org/AmIFAIR (public) • https://fairdata.systems (commercial)	• The FAIR Guiding Principles for scientific data management and stewardship [39]	
FiaB: FAIR-in-a-Box	Tool to deploy a FDP automatically with features to convert CDEs in CSV format to the CARE SM model and serve them via the EJP RD VP API v2 [30] and v4 [31] that are based on the Beacon-2 specs [18].	• https://github.com/ejp-rd-vp/FiaB • [40]	• EJP RD Metadata Model • Metadata provisioning service conform to FAIR Data Point specifications • CARE SM • Beacon2-based VP API version 2 and 4 over CARE-SM	
LS Login	Life Science Login, also known as Life Science AAI (Authentication and Authorisation Infrastructure). It reconsiled the AAI protocols of ELIXIR and BBMRI.	• https://lifescience-ri.eu/ls-login.html • Section 5.1.7.1	• Life Science AAI • EJP RD Metadata Model • Metadata provisioning service conform to FAIR Data Point specifications	
MOLGENIS with FDP and Beacon-2 functionality	Customizable platform for managing (scientific) data and implementing FAIR principles.	• https://vp-onboarding-doc.readthedocs.io/en/latest/level_2/solutions/molgenis.html • https://github.com/molgenis/molgenis-emx2 • [41]	• EJP RD Metadata model • CARE SM (where appropriate) • EJP-RD Beacon-2 API • FDP specifications • OpenID-Connect • LifeScience AAI	
RD-Nexus with FDP and Beacon-2 functionality	Dedicated discovery software to enable safe findability in a federated manner, whereby no data or metadata leave servers controlled by the resource.	• https://vp-onboarding-doc.readthedocs.io/en/latest/level_2/solutions/rd_nexus.html	• EJP RD Metadata model • CARE SM (where appropriate) • EJP-RD Beacon-2 API • FDP specifications • OpenID-Connect • LifeScience AAI	
VP front-end Portal Reference implementation	Reference implementation of a portal on top of the FAIR-based VP infrastructure	• https://github.com/ejp-rd-vp/vp-dp-frontend • https://vp.ejprarediseases.org/	• EJP RD Metadata Model and FAIR Data Point Specifications • Beacon-2 framework	

Tool name	Description	References	Applies to … standard

1. The original source is not enhanced: VP compliance depends on the persistence of the deposition service that implemented FDP and Beacon-2 functionality.
2. The reference provided points to the official VP-wide resource index as provided by the EJP RD project. This deployment also serves as a reference for FDP indexes.

5.2.2. The FAIR Evaluator

The FAIR Evaluator is a tool that assembles automated tests of individual FAIR principles and applies them to a digital resource (e.g., a registry metadata record) to determine the resource's level of compliance with the FAIR Principles. A public version of The Evaluator is available to test open-access resources. A commercial version is available for deployment inside sensitive data spaces, or for cases where the result of the evaluation should not be made public. The FAIR evaluator has been applied to resources that are compliant with the EJP's DCAT-based Metadata model, and achieve among the highest scores recorded for digital objects in the health space, indicating that the decisions of the EJP VP have resulted in a notable degree of "FAIRness". Beyond this, a customized version of the FAIR Evaluator was constructed to assist the onboarding team to quickly identify failures of a resource to comply with VP requirements, thus hastening the process of debugging onboarding attempts. This has proven extremely useful for new resources as they come into the VP network.

6. Standards and tools in the VP laboratories for evolving the Virtual Platform for Rare Disease research

Below we list standards and tools that are available for evolving the VP through a process of co-creation between new data providers and VP architects. The tools function in a VP Lab, an enhanced, but more experimental mirror of the VP network (see section 1.2).

6.1. Standards

6.1.1. EJP RD Metadata model for the VP Labs

The EJP RD Metadata model was created in anticipation of further evolution by a collaborative effort between future stakeholders. In the VP labs the latest version of the metadata model is in use [5].

Operating the EJPRD Metadata model in the VP Laboratories

- The VP Labs already implements the set of metadata extensions that have been created for the VP, but are not yet operational. These include the Beacon-2-based VP API Biosamples endpoint, and the implementation of the Beacon-2-based VP API v4 messaging specification, including the addition of v4-specific ontology "tags" to identify v4-compliant data services.
- At this time, there are no other operational extensions to the EJPRD Metadata model in-use by VP Labs. Section 5.1.1 describes the metadata extensions that

are already anticipated to fulfill the deeper discovery requirements of "Level 3" analyses.

6.1.2. DUC, CCE – Data Use Conditions and Common Condition of use Elements

Resources on the VP network are strongly encouraged to include a Data Use Conditions (DUC) profile in the resource description to increase the transparency of what uses of their (meta)data the data provider allows. It is often critical for applications that this information is available, for example to filter by access conditions or to facilitate post-hoc audits by logging access conditions as machine-readable provenance during an automated workflow. *Common Conditions of use Elements* (CCE) have been defined to aid the community of data producers and data consumers in converging to a workable number of reusable use conditions. CCEs and DUC profiles have been described previously (section 5.1.3.1 and 5.1.4.5 in [1]). DUC/CCE profiles are *not normative, nor legally binding* – they are intended to be used for automated search filtering to assist in finding data providers who appear to match the desired data usage activities.

- Further Information: doi: 10.1038/s41597-024-03279-z [36]

6.1.2.1. Digital rights policies in terms of the Open Digital Rights Language (ODRL)

On the VP labs, data providers and users can test uses of the metadata property odrl:hasPolicy that refers to a machine readable policy document in terms of the Open Digital Rights Language (ODRL). ODRL, a W3C-recommendation, provides a way to richly express the rights and/or responsibilities associated with access to and/or use of the resource. The range of the property is a URL pointing to the location of the RDF document containing the ODRL statements.

- Further information: ; https://www.w3.org/TR/odrl-model/

6.1.3. Co-created Data Models for federated analysis (VP Labs)

Beyond the JRC Common Data Elements for patient registries (section 5.1.3.2), no further data element sets have been defined as a general standard for the VP; however, the VP Labs are prepared for adding support for highly specialized datatypes (e.g. interventions, imaging, genomics) together with the data producing partners in the Rare Disease community. The CARE-SM model can support hundreds of specialized datatypes via the specialization of its 22 core models, but it should be emphasized that CARE-SM is intended to be a convenience, not a network or platform requirement – that is, resources that choose to conform to CARE-SM (e.g. via the pre-constructed FiaB transformation pipeline) will be able to take immediate advantage of the VP Labs Level 3 tooling, while others will need to map the tooling into their local data representation. Assuming that not all resources will conform to CARE-SM, it will become necessary to define a mechanism for communicating to an automated agent what the internal data model is, for each site. To date, this has been achieved via the DCAT "conformsTo" property, which is often used to point to a shape definition (e.g. in SHACL or ShEx) that represents the nature of the internal data model (section 6.1.4). This, however, has not been standardized for the VP, nor is it likely to be sufficient to fully communicate a resource's holdings to a visiting agent such as a FAIR Data

Train[4]. As such, VP Labs partners are advised to build strawman proposals to fully understand the requirements and build prototype tools that take advantage of these behaviours on the VP Labs infrastructure. The VP governance process will determine when these are sufficiently stable to move into the production VP.

6.1.4. Shape definitions denoting modelled data

Resources in a VP Lab that is set up for federated analysis on heterogeneous data should provide, via their FDP, an indication of the "shape" of the contained data – that is, the schema to which the contained data adheres [22].

For data that is represented in RDF there are two related shape definition languages. Resources on the VP use the W3C-recommended Shapes Constraint Language (SHACL) by default [42]. Optionally, for features unavailable in SHACL, resources can provide Shape Expressions (ShEx) Language definitions, next to the minimally expected SHACL definitions. SHACL (and ShEx) makes use of the ontological representation of the data in RDF to define the constraints that the properties of the data elements must comply with. Moreover, while the semantic model is an external vocabulary to which the resource refers , the shape definition acts as a more structural model, and allows concrete determinations of the elements that *are* contained in the data, versus the elements that *may be* contained [42].

The exact approach to the use of shapes on the VP Labs is an ongoing exploration, given that (a) not all data is in RDF format, and (b) there are limitations in the expressivity of DCAT regarding its referencing of standards. DCAT recommends the use of the *conformsTo* property to point to the standard; however, a given dataset may conform to multiple standards, with respect to both content and structure. For example, a microarray dataset might structurally conform to XML, and content-wise conform to the MIAME minimal information model.

6.1.5. Semantic Phenopackets model - version 2

Resources on the VP Labs that offer data sets containing genotype and phenotype information can use the ontological Phenopacket model in the same way that CARE-SM modules are used for JRC Common Data Elements ([43], [44]). The GA4GH Phenopackets standard was created to facilitate exchange of data elements relevant for genotype-phenotype analysis (Section 5.1.1.9 in [1]; [45]). It provides insight into the data elements that genotype-phenotype studies require. The ontological model enables a federated approach, where resources provide a definition of their data sets and data elements for visiting agents, instead of sending potentially sensitive data around the network. The ontological Phenopackets model follows the same design pattern of CARE-SM, hence the model is based on terms from the Semanticscience Integrated Ontology [25]. Additionally, for some common predicates and those originating from biomedical domain knowledge other ontologies have been added such as the NCI Thesaurus [46], and Human Phenotype Ontology [47] and Genotype Ontology [48]. Shape definitions (section 6.1.4) for 16 Phenopacket data elements were created in SHACL for the EJP RD [49] and ShEx for

[4] In analogy to FAIR Data 'stations', a FAIR Data 'train' denotes a general concept for FAIR transport containers that communicate information between stations.

the BIND project [50]. A proposal for referencing the model into resource metadata is described in the next paragraph.

6.1.5.1. Relating OMICS and genotype-phenotype metadata in resource descriptions

For FAIR-based federated omics analysis, for instance to automatically decide which statistical method to apply on which resources, it is necessary to refer to omics metadata and phenotype data from the resources' DCAT-based metadata. As mentioned in section 6.1.2, the '*dcat:conformsTo*' property is not sufficient to describe this relationship. Co-creating and testing the use of the 'dcat:Relationship' class as an extension of the EJP RD Metadata Model is necessary. As an example of co-creating new models in the VP Labs, a strawman model was created for connecting data elements in OMICS and genotype-phenotype data sets (see Figure 1). Minimal metadata was described in terms of commonly used ontologies such as the Orphanet Rare Disease Ontology (ORDO) and the NCBITaxon ontology. Phenotypic information has been added to the model in a machine understandable way by defining an additional DCAT Dataset and Distribution. The DCAT 'Relationship' class was used to denote the relationship between the omics data set and the data set with phenotype information. The data are virtually linked at record level by common semantic classes for data elements such as the person. The omics data are incoherent without this virtual connection.

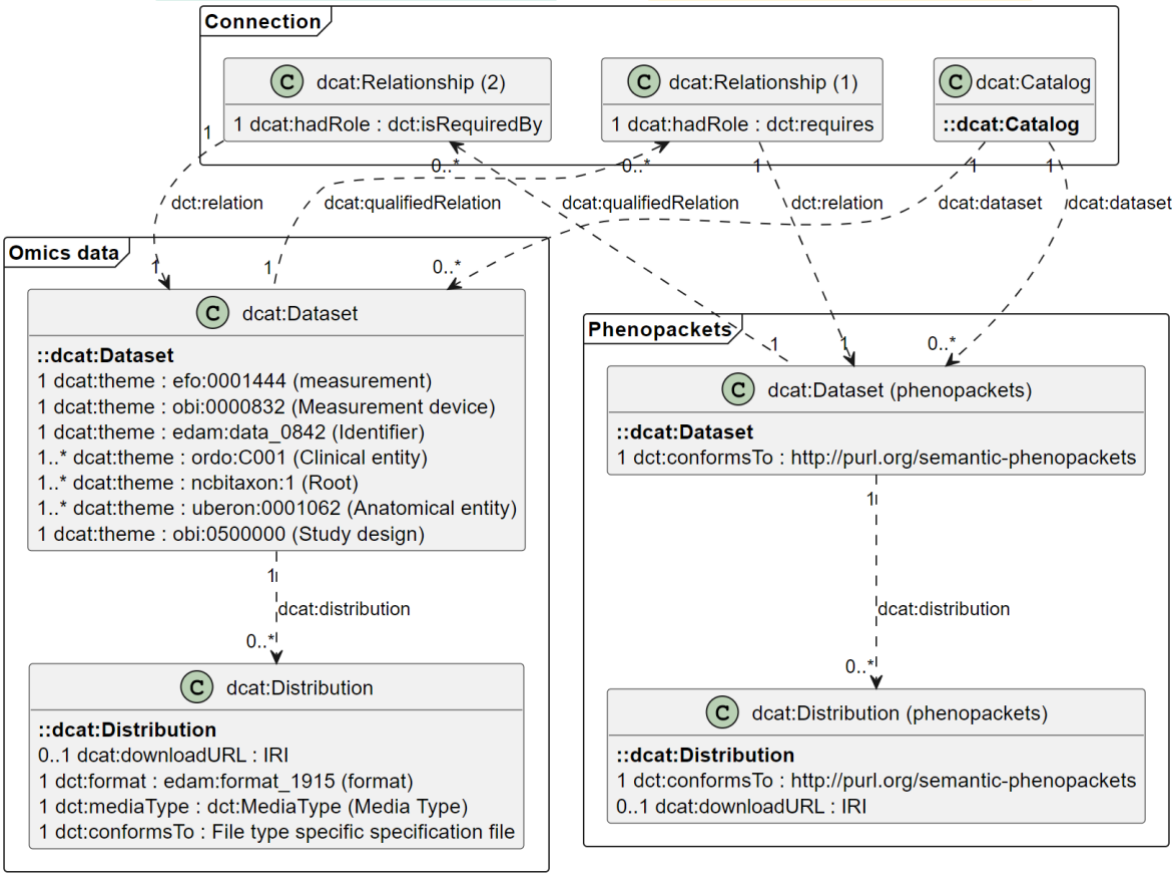

Figure 1 – Example of a strawman solution for connecting Omics and Phenopackets metadata via dcat:Relationship. At data record level, additional functionally meaningful links are created by the use of common ontological terms in the semantic models denoting the data elements in Omics, genotype, and phenotype data sets (not shown here).

6.2. Tools

6.2.1. The VP-labs testbed

The EJP RD delivered a VP-labs testbed as a Virtual Machine where partners can instantiate demonstrator or prototype VP components in a public manner, containing only mock, synthetic, or public data. It was originally created for testing methods for FAIR-based federated analysis. It contains the core components of the VP Network (FAIR Data Points, and a FAIR Data Point index – fully independent of those used to run the production EJP RD-VP) as well as a novel Virtual Platform that is aimed more at technical developers, as it exposes the VP tool interfaces in a less user-friendly, but more granular, manner. This allows tool builders to deeply test and troubleshoot their VP tools prior to registering them on the production Virtual Platform. It is also being used to host local copies of public datasets (e.g. subsections of the MONARCH dataset) to allow for more efficient federated query during the building and troubleshooting of novel analytical environments. EJP RD partners will keep the testbed available for future projects as a springboard for further development of a rapid prototyping environment.

One of the novel tools already deployed on the VP Labs testbed is the FLAIR-GG Virtual Platform. This is a parallel attempt to build a VP that can communicate with the EJP VP Network (i.e. the FDP Index and the network of participating FAIR Data Points). Funded by the Gobierno de España (independent of the EJP RD), FLAIR-GG is specifically designed for L3 interactions, and thus, is tuned to the discovery and federated invocation of data services and feeding the federated results into pre-defined Analytics applications (in the form of JupyterLite notebooks). FLAIR-GG is, itself, an API, allowing most/all VP functionality to be accessed without the use of the Web interface, thus VP exploration can be done from inside of the Jupyter analytics environment.

6.2.2. Tools to prepare data for federated analysis on FAIR data on the VP network, example: interpretation of pathogenic variants

In chapter 5 several tools have been described that facilitate the implementation of the core standards by which resources start to contribute to the VP network. On the VP Labs, new tools to apply these standards or tools to apply candidate VP standards are co-developed in collaboration with data producers, data users, and experts of the principles of the VP infrastructure. For example, tools have been developed to prepare data sources in a VP laboratory in order to investigate a scenario in which a clinical geneticist with an undiagnosed patient orders that an algorithm is run on genome and clinical data sources in the VP network to propose diagnoses or leads with explanations (see [51] for the scenario). The clinician is explicitly *not* asking to get data from the sources: the compiled answer suffices. Everything else remains unseen by the clinician. It is assumed that the analysis workflow will use the VP discovery services to automatically select appropriate resources, after which the algorithms will run locally or under the auspices of the local resource (e.g. via a temporary cache in a trusted analysis environment). The preparatory workflow that was implemented to demonstrate the ability to get answers from federated ontologised data uses the Semantic Phenopacket model based on CARE-SM as reference (section 6.1.520). Shape definitions were defined for the data elements that a resource contains, which

in turn were used to map between the original data format (e.g. a locally used JSON schema) and the RDF representation in terms of the Semantic Phenopacket model (Figure 2, Table 3). The workflow involves automations and two steps to be performed by people with a data steward role. The first step is to structure or export local data conforming to an expected structure (here a JSON format, in the case of FAIR-in-a-box this is a preformatted CSV file). The second is for a FAIR data steward to define the shape definitions of the data elements at hand (here Phenopackets data elements). The workflow (Figure 2) and the FAIR principles-based tools and standards used to accommodate it (Table 3) are provided as a paradigm for the future evolution of the VP network.

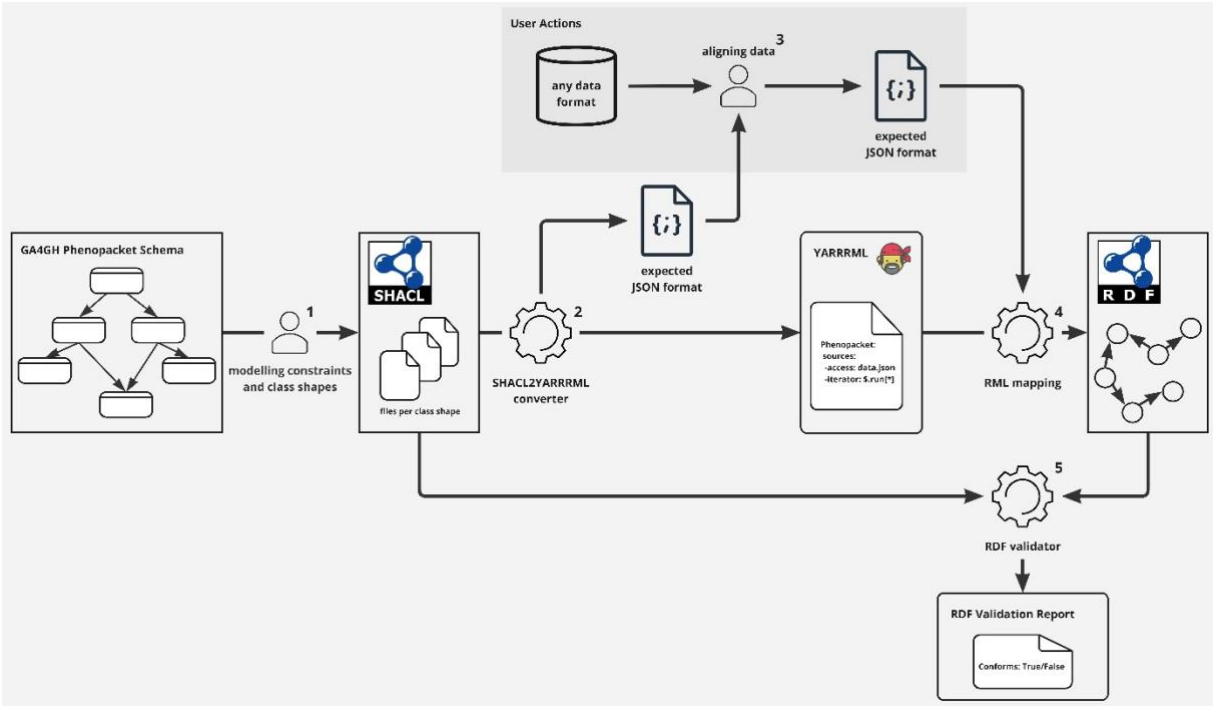

Figure 2 - Overview of FAIR standards and tools for federated analysis on FAIR data. The process involves transforming SHACL specifications into YARRRML mapping rules. The SHACL specifications are also used to generate a JSON template to which data must be mapped. The tool is then able to automatically generate RDF triples from the populated JSON. To ensure compliance, the tool validates the result against the original SHACL definitions.

Table 3 - Tools to map data from a local JSON format to RDF data in terms of the Semantic Phenopacket model based on CARE-SM.

Name tool or standard	Description	References	Related FAIR … standard	
Semantic Phenopackets model	Ontological model representing the data elements represented in the GA4GH Phenopackets exchange format.	• Section 6.1.5 • https://github.com/rosazwart/phenopackets-v2-rdf-schema	• CARE-SM	
FAIR Data Steward	Human FAIR data expert involved in creating the shape	• https://doi.org/10.1186%2Fs13023-022-02558-5	• Semantic Phenopackets based on CARE-SM	

Name tool or standard	Description	References	Related FAIR … standard	
	definitions of the genotype-phenotype data that a resource offers	• https://doi.org/10.1162/dint_a_00028 • https://doi.org/10.1038%2Fs41597-022-01325-2		
Phenopacket SHACL definitions	SHACL shape definitions for Phenopacket data elements	• https://github.com/rosazwart/phenopackets-v2-rdf-schema/tree/main/phenopacketv2_shacl	• Semantic Phenopackets based on CARE-SM • EJP RD Metadata Model	
SHACL2YARRRML converter	Script to generate YAML mapping files to convert from JSON to RDF using shape definitions and the RDF Mapping Language	• https://github.com/rosazwart/phenopackets-v2-rdf-schema/tree/main/shacl2yarrrml	• Semantic Phenopackets based on CARE-SM	
YARRRML2RDF	Script to execute a conversion from JSON to RDF using a YAML template that defines the mapping between JSON and RDF shape definitions for (i) clinical data elements, (ii) candidate pathogenic variants	• https://github.com/rosazwart/phenopackets-v2-rdf-schema/tree/main/yarrrml2rdf	• Semantic Phenopackets based on CARE-SM	
Matey	A browser-based IDE to generate RDF triples given a data set and YARRRML rules. It is used by the YARRRML2RDF script.	• https://rml.io/yarrrml/matey/	• Semantic Phenopackets based on CARE-SM	
RDF Validator	A script to validate a given RDF file against the SHACL files containing all the class shapes.	• https://github.com/rosazwart/phenopackets-v2-rdf-schema/tree/main/rdfvalidator	• Semantic Phenopackets shape definitions in RDF	

6.2.3. Tools complementing the core FAIR standards and tools of the VP ecosystem

There are many standards and tools that are relevant for resources on the VP that are not listed in this document. A short list of tools with an indirect relation to core FAIR standards and tools is highlighted for further evolution of the VP via VP Labs.

In [IOS Press Ebooks - Privacy-Preserving Linkage of Distributed Pseudonymised Datasets in a Virtual European Rare Disease Platform] we have demonstrated how the EUPID Services, a pseudonymization

service supporting privacy-preserving record linkage, can be linked to the EJP RD VP on level 1 and level 2, by implementing an FDP and a Beacon-2 API.

Tool name	Description	References	Related FAIR standard
ERDRI[1]	European Platform on Rare Disease Registration (EU RD Platform), including a privacy preserving data linkage service (SPIDER), a directory of registries (ERDRI.dor), and a registry of data element metadata (ERDRI.mdr).	• https://eu-rd-platform.jrc.ec.europa.eu/	• EJP RD Metadata Model • Standards on data models, ontology, terminology and vocabulary • Metadata provisioning service conform to FAIR Data Point specifications • Standards on data communication • Standards on data security
EUPID	European Patient Identity Services, tool to link data sets via a privacy preserving person identifier.	• https://services.eupid.eu/ • Section 5.2.3.1 in [1] • [52], [53], [54]	• CARE-SM • Metadata provisioning service conform to FAIR Data Point specifications • Standards on data communication, including the EJP RD Beacon-2 based API • Standards on data security
CARE-SM to OMOP-CDM Mapping Service	Semi-automated service to map data in the CARE-SM format to the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) of the Observational Health Data Sciences and Informatics initiative (OHDSI) [55]	• https://github.com/ejp-rd-vp/CARE2OMOP • [56], [57]	• CARE-SM • Metadata provisioning service conform to FAIR Data Point specifications
PROV-O	The PROV Ontology	• https://www.w3.org/TR/prov-overview/ • http://www.w3.org/TR/prov-o/	• CARE-SM • Metadata provisioning service conform to FAIR Data Point specifications • Standards on data communication • Standards on data security
RD Code Mappings (files, API, web service)	Orphanet nomenclature files for coding, intended to be used to implement the Orphanet nomenclature in Health Information	• https://github.com/orphanet-rare-diseases-issues/RD-CODE • https://api.orphacode.org/ • https://mappings.orphacode.org/	• EJP RD Metadata Model • CARE-SM • Disease-related data and ontological models

Tool name	Description	References	Related FAIR standard
	Systems for codification purposes		
SSSOM	A Simple Standard for Sharing Ontological Mappings	• https://github.com/mapping-commons/sssom • [27]	• All ontology-based models

7. References

[1] Austrian Institute of Technology (AIT), 'Second Report on core set of FAIR software tools and on extended set of unified FAIR data standards applied in EJP RD', INSERM (Coordinator), Project Deliverable D12.03, Dec. 2022. [Online]. Available: https://ec.europa.eu/research/participants/documents/downloadPublic?documentIds=080166e5f717ae53&appId=PPGMS

[2] Austrian Institute of Technology (AIT), 'First report on core set of FAIR software tools and on extended set of unified FAIR data standards applied in EJP RD', INSERM (Coordinator), Project Deliverable D12.02, Apr. 2021. [Online]. Available: https://ec.europa.eu/research/participants/documents/downloadPublic?documentIds=080166e5eaae53ad&appId=PPGMS

[3] L. O. B. Da Silva Santos, K. Burger, R. Kaliyaperumal, and M. D. Wilkinson, 'FAIR Data Point: A FAIR-Oriented Approach for Metadata Publication', *Data Intell.*, vol. 5, no. 1, pp. 163–183, Mar. 2023, doi: 10.1162/dint_a_00160.

[4] 'FAIR Data Point specifications'. Accessed: Aug. 18, 2024. [Online]. Available: https://specs.fairdatapoint.org/fdp-specs-v1.2.html

[5] EJP RD Metadata work focus group, *ejp-rd-vp/resource-metadata-schema*. (Aug. 12, 2024). EJP RD Pillar 2: Coordinated Access to Data and Services. Accessed: Aug. 16, 2024. [Online]. Available: https://github.com/ejp-rd-vp/resource-metadata-schema

[6] Global Alliance for Genomics and Health, *Unified repository for Beacon v2 Code & Documentation*. (Aug. 02, 2024). Global Alliance for Genomics and Health. Accessed: Aug. 16, 2024. [Online]. Available: https://github.com/ga4gh-beacon/beacon-v2

[7] Joint Research Centre, *Set of Common Data Elements*. [Online]. Available: https://eu-rd-platform.jrc.ec.europa.eu/set-of-common-data-elements_en

[8] SWAT4HCLS, *Take CARE of your patient data. Clinical And Registry Entries (CARE) Semantic Model*, (Apr. 01, 2024). Accessed: Aug. 31, 2024. [Online Video]. Available: https://www.youtube.com/watch?v=XTghx4h7jyI

[9] *CARE-SM/CARE-Semantic-Model*. (Aug. 27, 2024). CARE-SM. Accessed: Aug. 31, 2024. [Online]. Available: https://github.com/CARE-SM/CARE-Semantic-Model

[10] 'EJP RD Virtual Platform: Resources onboarding manual — EJP RD Onboarding Document documentation'. Accessed: Aug. 17, 2024. [Online]. Available: https://vp-onboarding-doc.readthedocs.io/en/latest/index.html

[11] 'EJP RD Pillar 2: Coordinated Access to Data and Services', GitHub. Accessed: Aug. 27, 2024. [Online]. Available: https://github.com/ejp-rd-vp

[12] R. Albertoni, D. Browning, S. Cox, A. Gonzalez Beltran, A. Perego, and P. Winstanley, 'Data Catalog Vocabulary (DCAT) - Version 2', Data Catalog Vocabulary (DCAT) - Version 2. [Online]. Available: https://www.w3.org/TR/vocab-dcat-2/

[13] 'RDF - Semantic Web Standards'. Accessed: Aug. 16, 2024. [Online]. Available: https://www.w3.org/RDF/

[14] T. Berners-Lee, 'Linked Data - Design Issues'. Accessed: Aug. 16, 2024. [Online]. Available: https://www.w3.org/DesignIssues/LinkedData.html

[15] 'EJP RD Metadata Schema — EJP RD Onboarding Document documentation'. Accessed: Aug. 17, 2024. [Online]. Available: https://vp-onboarding-doc.readthedocs.io/en/latest/level_1/metadata.html#

[16] R. Iannella and J. McKinney, 'vCard Ontology - for describing People and Organizations'. Accessed: Aug. 16, 2024. [Online]. Available: https://www.w3.org/TR/vcard-rdf/
[17] Internet Engineering Task Force, 'vCard Format Specification'. Aug. 2011. [Online]. Available: https://datatracker.ietf.org/doc/html/rfc6350
[18] Global Alliance for Genomics and Health, 'Unified repository for Beacon v2 Code & Documentation'. Jul. 19, 2024. [Online]. Available: https://github.com/ga4gh-beacon/beacon-v2/
[19] M. Black *et al.*, 'EDAM: the bioscientific data analysis ontology (update 2021)', 2022, *F1000 Research Limited*. doi: 10.7490/F1000RESEARCH.1118900.1.
[20] J. Ison, Matúš Kalaš, H. Ménager, E. Willighagen, B. Grüning, and Albangaignard, *edamontology/edamontology: EDAM 1.25*. (Jun. 18, 2020). Zenodo. doi: 10.5281/ZENODO.822690.
[21] 'Distribution — EJP RD Onboarding Document documentation'. Accessed: Aug. 19, 2024. [Online]. Available: https://vp-onboarding-doc.readthedocs.io/en/latest/level_1/properties/distribution.html
[22] 'Usage — FAIR Data Point 1.16 documentation'. Accessed: Aug. 18, 2024. [Online]. Available: https://fairdatapoint.readthedocs.io/en/latest/usage/usage.html#resource-definitions
[23] *ejp-rd-vp/FiaB*. (Aug. 19, 2024). Ruby. EJP RD Pillar 2: Coordinated Access to Data and Services. Accessed: Aug. 23, 2024. [Online]. Available: https://github.com/ejp-rd-vp/FiaB
[24] R. Kaliyaperumal *et al.*, 'Semantic modelling of common data elements for rare disease registries, and a prototype workflow for their deployment over registry data', *J. Biomed. Semant.*, vol. 13, no. 1, p. 9, Mar. 2022, doi: 10.1186/s13326-022-00264-6.
[25] M. Dumontier *et al.*, 'The Semanticscience Integrated Ontology (SIO) for biomedical research and knowledge discovery', *J. Biomed. Semant.*, vol. 5, no. 1, p. 14, Mar. 2014, doi: 10.1186/2041-1480-5-14.
[26] R. Jackson *et al.*, 'OBO Foundry in 2021: operationalizing open data principles to evaluate ontologies', *Database*, vol. 2021, p. baab069, Sep. 2021, doi: 10.1093/database/baab069.
[27] N. Matentzoglu *et al.*, 'A Simple Standard for Sharing Ontological Mappings (SSSOM)', *Database J. Biol. Databases Curation*, vol. 2022, p. baac035, May 2022, doi: 10.1093/database/baac035.
[28] 'The Human Phenotype Ontology: A Tool for Annotating and Analyzing Human Hereditary Disease - ScienceDirect'. Accessed: Aug. 20, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0002929708005351?via%3Dihub
[29] *orphanet-rare-diseases-issues/RD-CODE*. (Aug. 14, 2024). orphanet-rare-diseases-issues. Accessed: Aug. 20, 2024. [Online]. Available: https://github.com/orphanet-rare-diseases-issues/RD-CODE
[30] *ejp-rd-vp/vp-api-specs*. (May 09, 2024). EJP RD Pillar 2: Coordinated Access to Data and Services. Accessed: Aug. 27, 2024. [Online]. Available: https://github.com/ejp-rd-vp/vp-api-specs
[31] 'ejp-rd-vp/vp-api-specs at v4.0_spec', GitHub. Accessed: Aug. 27, 2024. [Online]. Available: https://github.com/ejp-rd-vp/vp-api-specs
[32] 'Overview of the AAI', ELIXIR. Accessed: Aug. 22, 2024. [Online]. Available: https://elixir-europe.org/platforms/compute/aai/overview
[33] 'LS Login | LifeScience RI'. Accessed: Aug. 26, 2024. [Online]. Available: https://lifescience-ri.eu/ls-login.html
[34] 'LS Login : Life Science Login Attribute Requirements | LifeScience RI'. Accessed: Aug. 26, 2024. [Online]. Available: https://lifescience-ri.eu/ls-login/documentation/service-provider-documentation/life-science-login-attribute-requirements.html
[35] L. O. Bonino Da Silva Santos *et al.*, 'Frugal FAIR Data Point: A Document-Based Implementation for Enhanced Interoperability and Accessibility', 2024, doi: 10.4126/FRL01-006473140.
[36] M. del C. Sanchez Gonzalez *et al.*, 'Common conditions of use elements. Atomic concepts for consistent and effective information governance', *Sci. Data*, vol. 11, p. 465, May 2024, doi: 10.1038/s41597-024-03279-z.

[37] *ejp-rd-vp/fdp-populator*. (Aug. 26, 2024). Python. EJP RD Pillar 2: Coordinated Access to Data and Services. Accessed: Aug. 27, 2024. [Online]. Available: https://github.com/ejp-rd-vp/fdp-populator
[38] 'Local Deployment — FAIR Data Point 1.16 documentation'. Accessed: Aug. 23, 2024. [Online]. Available: https://fairdatapoint.readthedocs.io/en/latest/deployment/local-deployment.html
[39] M. D. Wilkinson *et al.*, 'The FAIR Guiding Principles for scientific data management and stewardship', *Sci. Data*, vol. 3, no. 1, p. 160018, Mar. 2016, doi: 10.1038/sdata.2016.18.
[40] O. M. Benhamed *et al.*, 'The FAIR Data Point: Interfaces and Tooling', *Data Intell.*, vol. 5, no. 1, pp. 184–201, Mar. 2023, doi: 10.1162/dint_a_00161.
[41] EMBL-EBI (EGA), 'Fourth version Additional facilities integrated to resources regarding data deposition and access, including user guidelines and documentation.', INSERM (Coordinator), Project Deliverable D11.14, Dec. 2022. [Online]. Available: https://ec.europa.eu/research/participants/documents/downloadPublic?documentIds=080166e5f71798ec&appId=PPGMS
[42] 'Shapes Constraint Language (SHACL)'. Accessed: Aug. 18, 2024. [Online]. Available: https://www.w3.org/TR/shacl/
[43] 'Welcome to the documentation for the phenopacket-schema! — phenopacket-schema 2.0 documentation'. Accessed: Aug. 19, 2024. [Online]. Available: https://phenopacket-schema.readthedocs.io/en/v2/
[44] rosazwart, *rosazwart/phenopackets-v2-rdf-schema*. (Aug. 22, 2024). Python. Accessed: Aug. 27, 2024. [Online]. Available: https://github.com/rosazwart/phenopackets-v2-rdf-schema
[45] J. O. B. Jacobsen *et al.*, 'The GA4GH Phenopacket schema defines a computable representation of clinical data', *Nat. Biotechnol.*, vol. 40, no. 6, pp. 817–820, Jun. 2022, doi: 10.1038/s41587-022-01357-4.
[46] 'OBO Foundry'. Accessed: Aug. 28, 2024. [Online]. Available: https://obofoundry.org/ontology/ncit.html
[47] 'OBO Foundry'. Accessed: Aug. 28, 2024. [Online]. Available: https://obofoundry.org/ontology/hp.html
[48] 'OBO Foundry'. Accessed: Aug. 28, 2024. [Online]. Available: https://obofoundry.org/ontology/geno.html
[49] 'phenopackets-v2-rdf-schema/phenopacketv2_shacl at main · rosazwart/phenopackets-v2-rdf-schema'. Accessed: Aug. 28, 2024. [Online]. Available: https://github.com/rosazwart/phenopackets-v2-rdf-schema/tree/main/phenopacketv2_shacl
[50] 'Brain Involvement iN Dystrophinopathies | BIND Project | Fact Sheet | H2020', CORDIS | European Commission. Accessed: Aug. 27, 2024. [Online]. Available: https://cordis.europa.eu/project/id/847826
[51] R. Zwart and A. Waagmeester, 'Welcome to the EJP-RD FDP Infrastructure and VP Portal Demonstrator — EJP-RD FDP and Virtual Portal demonstrator'. Accessed: Aug. 25, 2024. [Online]. Available: https://ejp-rd-vp.github.io/DistributedAnalysisDemonstrator/
[52] M. Baumgartner *et al.*, 'Health data space nodes for privacy-preserving linkage of medical data to support collaborative secondary analyses', *Front. Med.*, vol. 11, p. 1301660, Apr. 2024, doi: 10.3389/fmed.2024.1301660.
[53] D. Hayn *et al.*, 'Use Cases Requiring Privacy-Preserving Record Linkage in Paediatric Oncology', *Cancers*, vol. 16, no. 15, p. 2696, Jul. 2024, doi: 10.3390/cancers16152696.
[54] D. Hayn *et al.*, 'Privacy-Preserving Linkage of Distributed Pseudonymised Datasets in a Virtual European Rare Disease Platform', in *Digital Health and Informatics Innovations for Sustainable Health Care Systems*, IOS Press, 2024, pp. 1442–1446. doi: 10.3233/SHTI240683.
[55] 'Data Standardization – OHDSI'. Accessed: Aug. 27, 2024. [Online]. Available: https://www.ohdsi.org/data-standardization/
[56] N. Queralt-Rosinach *et al.*, 'Mapping OHDSI OMOP Common Data Model and GA4GH Phenopackets for COVID-19 disease epidemics and analytics', Nov. 25, 2022, *OSF*. doi: 10.37044/osf.io/ep3xh.

[57] N. Benis *et al.*, 'EJP RD meets OHDSI: enabling interoperability for rare disease research', Mar. 17, 2023. doi: 10.5281/zenodo.7745968.