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 12.2
First 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 03 – AIT

Due date of deliverable: month 36

Dissemination level:
Public

Table of content

1. Scope

This report intends to provide an overview on the status of software tools and standards relevant to the European Joint Programme on Rare Diseases (EJP RD). Specifically, it lists those elements which have been identified in preceding work as being relevant to the Virtual Platform (VP).

This collection of items will be designated as "the list" in the following.

The relevance to the VP of items in the list stems either from the VP design and implementation work or from the interaction with potential users of VP components and their existing or upcoming systems, respectively.

For the purpose of this document, "tools" are defined as all elements of the VP as described in the Virtual Platform Specification (VIPS)[1], except for standards and data sources.

This document is the first version and is to be updated by the Deliverable 12.3 "Second report on core set of FAIR software tools & on extended set of unified FAIR data standards applied in EJP RD", due in month 48 and will be completed by the Deliverable 12.4, "Report on extended set of FAIR software tools, applied in EJP RD, including overview of FAIRification guidelines for RD data managers", due in month 60.

Therefore, the list is expected to grow and the classification of items regarding their types and VP adoption status is expected to be updated/changed in subsequent releases.

2. Sources

This report builds on numerous preceding activities in general and the following documents and sources in particular.

Del 12.1 "Report on core set of unified FAIR data standards"

This deliverable was concerned with standards relevant to the VP. It identified: (a) existing standards that can be used/piloted directly; and (b) standards with a need to aggregate and/or map between and/or extend existing standards before they can be used in the VP context. Del 12.1 provided a very comprehensive list of standards, potentially relevant to the VP. Most of the standards included in Del 12.1 were considered for the list as well.

AD36 Additional Deliverable Virtual Platform Specification (VIPS)

The Virtual Platform (VP) provides RD stakeholders with resources relevant to RD research. The VIPS depicts the overall structure of the VP (architecture, functional and non-functional requirements), the use cases supported by the VP, the process how to link resources to the VP and the agreed guidelines and standards. Its level of detail aims at the managerial level, not at the implementation level. The VIPS provides the architectural framework for the VP and, as such, it defines a classification scheme for VP components and their underlying software/service developments, which have been adopted for the present report as well.

[1] VIPS first version: https://www.ejprarediseases.org/wp-content/uploads/2021/10/EJPRD_P2_AD36_PU_Virtual-Platform-Specification.pdf

The list of elements compiled by the FAIRification Work Focus
This list has been compiled by the FAIRification team while interacting with numerous stakeholders in EJP RD, and in particular the European Reference Networks (ERNs).

3. Main classification elements

The following paragraphs describe the classification elements used on the list to qualify the list entries with respect to different properties.

3.1. Type

The Type of list items can be one of the following elements

1. Standard
2. Software (needs to be installed)
3. Service (Software as a Service, SaaS, no installation needed)

Software and Services are defined as Tools in this document.

3.2. Development Status

The Status of software tools/services/standards refers to its state of usage with respect to the VP development and has been aligned to table 6 of the VIPS.

Status	Short Term	Description
1	Draft	The component being described has not been implemented yet or is under active development. However, it has not been shared or evaluated with stakeholders and end-users of the VP.
2	Trial use	The component being described is under active development and ready for trial use during the use case evaluation phase. However, no guarantees are made that future version remain compatible with older versions of the component.
3	Normative	The component described is in a stable state, can be used by clients and has guaranteed life cycle management process for future updates.
4	Deprecated	The component described is considered as being deprecated and might be removed from the document in the near future. For instance, a prototypic component developed in an iteration cycle might be dropped in favour for a more advanced component.

3.3. FAIR intent

The FAIR intent classifies each entry according to the main FAIR requirements that the item aims to help accomplish in a FAIRification process according to the submitter of the tool or standard:

1. Findable
2. Accessible
3. Interoperable
4. Reusable
    - For humans (including software engineers)
    - For Computers

3.3.1. FAIR for computers ('machine readable')

The term 'machine readable' is widely used in reference to implementing FAIR principles 'for computers'. Also, in this document, it can refer to *syntax* that computers can process (e.g., file formats and APIs) and *semantics* that computers can process (e.g., ontological models). The latter makes a resource 'actionable' for computers: a resource is 'understandable' for a non-human agent to autonomously act upon it (NB this includes *not* pursuing any action because the agent does not find appropriate access permissions). For the EJP RD VP 'machine readable' is interpreted in the context of the IRDiRC-recognized resource: 'The 15 FAIR Guiding Principles for scientific data management and stewardship'[2]. The intent of these principles taken together is that computers are able to process syntax *and* semantics of data and services. This enables Virtual Platforms to automatically 'adapt' their functionality to the FAIR resources that they are composed of. Not all standards and tools in this document refer to semantics. Each can help implement specific aspects of the FAIR principles and the EJP RD VP specifications.

3.4. Virtual Platform adoption status

The VP adoption status refers to whether a standard or tool is compatible with the EJP RD VP specifications. Compatibility with the specifications is recommended for contributing to the capabilities of the VP.

1. Part of and/or compatible with the latest version of the VP
2. NOT part of and/or compatible with the latest version of the VP
3. (undecided)

[2] Wilkinson, M., Dumontier, M., Aalbersberg, I. *et al.* 2016. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018. https://doi.org/10.1038/sdata.2016.18

4. Format

For the next versions of this deliverable the information may be presented as a list that can be searched, sorted, and filtered.

Classification of the FAIR maturity status[3] of the standards and tools is an ongoing process that is pursued for the EJP RD in the context of developing smart guidance for FAIR data stewardship, which is planned for 2022. Providing a *formal* FAIR maturity status for standards and tools is outside of the scope of the EJP RD project. Here, the main intended use is providing stewards an aid in delivering resources that are compatible with a FAIR-based VP. The status in this document reflects the intended use in FAIRification as indicated by the submitter of the tool or standard.

Continuous updating and refinement of the list is foreseen to be a rolling task and the knowledge required to compile and maintain this list requires the full breadth and depth of the EJP RD community expertise.

5. Used Standards and Tools

5.1. Standards

5.1.1. Meta-data standards

5.1.1.1. DCAT - Data Catalog Vocabulary (DCAT)

DCAT is a W3C-recommended vocabulary, defined in RDF, which is designed to facilitate interoperability between data catalogues published on the Web. The version that is currently used in the EJP RD is DCAT version 2.

- Status of development: Normative
- Further Information: https://www.w3.org/TR/vocab-dcat-2/

5.1.1.2. Dublin Core Metadata Terms and Element Set ('Dublin Core terms')

The *Dublin Core™ Metadata Terms*[4] are widely used terms to denote metadata elements. They are maintained by the Dublin Core™ Metadata Initiative (DCMI)[5]. Included are the fifteen terms of the *Dublin Core™ Metadata Element Set* (also known as "the Dublin Core")[6] plus several dozen properties, classes, datatypes, and vocabulary encoding schemes. The "Dublin Core" plus these extension vocabularies are collectively referred to as "DCMI metadata terms" ("Dublin Core terms" for short). They are expressed in RDF vocabularies for use in Linked Data. Creators of non-RDF metadata, such as in XML, JSON, UML, or relational databases, can use the terms by disregarding both the global identifier and the formal implications of term definitions in RDF, i.e., some of the machine-readable semantics will be lost. The terms are intended to be used in combination with metadata terms from other, compatible

[3] For example, see Bahim, C., *et al.*, 2020. The FAIR Data Maturity Model: An Approach to Harmonise FAIR Assessments. *Data Science Journal*, 19(1), p.41. DOI: http://doi.org/10.5334/dsj-2020-041

[4] https://www.dublincore.org/specifications/dublin-core/dcmi-terms/

[5] https://dublincore.org/

[6] https://www.dublincore.org/specifications/dublin-core/dces/ The fifteen-element Dublin Core has been formally standardized as ISO 15836, ANSI/NISO Z39.85, and IETF RFC 5013.

vocabularies in the context of application profiles. They are used as such in DCAT2 and thus the EJP RD Metadata model, as well as the ontologies that describe data elements, data access information, and provenance.

- Status of development: Normative
- Further Information and official definitions: https://dublincore.org/

5.1.1.3. EJP RD Metadata Model

The EJP RD Metadata Model is a machine-readable model for declaring information about a resource (e.g., a registry) for contributing to the functionality of the VP. The model extends DCAT2 (see 5.1.1.1) with resource types that are relevant for the VP. The model is defined in RDF by which it integrates with other models defined in RDF (a 'Linked Data' principle). Serialisation in non-RDF formats is possible (XML, JSON, etc.) at the expense of machine-readable semantics, similar to serialisation of the RDF definition of Dublin Core terms (see 5.1.1.2). Metadata about a resource that is standardised by the EJP RD Metadata Model (and thus also by DCAT2), and made accessible for machines with an interface that conforms to the FAIR Data Point (FDP) specifications (see 5.2.3.1) increases the Findability and Accessibility of the resource globally and in the VP, and enables the VP to dynamically adapt its functionality to the capabilities of its resources.

- Status of development: Trial use
- Further Information: https://github.com/S2Ola/EJP-RD-metadata-model

5.1.1.4. FASTA, FASTQ

FASTA and FASTQ files are flat file formats used to exchange nucleotide or amino acid sequence information. FASTQ files include a sequencing quality indicator.

- Status of development: Normative
- Further Information:
  - https://www.bioinformatics.nl/tools/crab_fasta.html;
  - https://support.illumina.com/bulletins/2016/04/fastq-files-explained.html

5.1.1.5. MIABIS 2.0

The Minimum Information About Biobank data Sharing (MIABIS) aims to standardize data elements used to describe biobanks, research on samples and associated data. The MIABIS Community Standards work on several granularity levels, with the aim to support interoperability between biobanks sharing their data. General attributes to describe biobanks, sample collections and studies at an aggregated/metadata level are defined in MIABIS Core 2.0 (Merino-Martinez et al., 2016). New MIABIS modules describing samples and sample donors at individual level have been approved by BBMRI-ERIC and are described here (Eklund et al., 2020).

- Status of development: Normative
- Further Information: https://github.com/MIABIS/miabis/wiki

5.1.1.6. RD3 (Rare Disease Data about Data) data model

Metadata database to track and find samples processed in the Sandbox and submitted to the EGA (see 5.2.6.2), including details on patient phenotypes, sample preparation, sequencing and information about files collected. RD3 is now being generalized with Dutch FAIR genomes initiative, http://fairgenomes.github.io in consideration of EJP RD metadata model. A reference implementation exists using MOLGENIS, see below.

- Status of development: Normative
- Further Information: https://github.com/molgenis/RD3_database

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

5.1.2.1. BED, GFF

The browser extensible data (BED) format is a concise and flexible way to represent genomic features and annotations

- Status of development: Normative
- Further Information:
  - https://bedtools.readthedocs.io/en/latest/content/general-usage.html#bed-format
  - https://bedtools.readthedocs.io/en/latest/content/general-usage.html#gff-format

5.1.2.2. CSV

Comma Separated Values is a common file format for tabular data, an old and simple flat-file format for representing data (text and values) in a rectangular matrix. It is not a formal "standard" as such but is very commonly used. A common alternative is the tab-delimited file format. RFC4180 describes the CSV format and mime-type for the internet community.

- Status of development: Normative
- Further Information: https://datahub.io/docs/data-packages/csv

5.1.2.3. ISA-Tab+serialisations (ISA-JSON etc)

The ISA Abstract Model, originally developed as a tabular format (ISA-Tab) since 2007, has been developed with several international collaborators and in synergy with related, domain-specific effort. ISA is supported as a tabular format (ISA-Tab) and a JSON format (ISA-JSON), with additional machine readable semantics as Linked Data (linkedISA), and by a programmable Python API (ISA API).

- Status of development: Normative
- Further Information:
  - https://isa-tools.org/format/specification.html
  - https://isa-specs.readthedocs.io/en/latest/isamodel.html

5.1.2.4. JSON LD

JavaScript Object Notation for Linked Data (RDF) that allows for semantic capture in the JSON syntax.

- Status of development: Normative
- Further Information: https://json-ld.org/

5.1.2.5. OML

Omics Markup Language (ISO/DIS 21393) is a data exchange format that is designed to facilitate exchanging omics data around the world without forcing changes of any database schema.

- Status of development: Normative
- Further Information: https://www.iso.org/standard/70855.html

5.1.2.6. OpenAPI Specification

The OpenAPI Specification, previously known as the Swagger Specification, is a specification for machine-readable interface files for describing, producing, consuming, and visualizing RESTful web services.

- Status of development: Normative
- Further Information: https://en.wikipedia.org/wiki/OpenAPI_Specification

5.1.2.7. Schema.org

Schema.org is a collaborative, community activity with a mission to create, maintain, and promote schemas for structured data on the Internet, on web pages, in email messages, and beyond.

- Status of development: Normative

Further Information: http://schema.org

5.1.2.8. XML

Extensible Markup Language (XML) is a markup language that defines a set of rules for encoding documents in a format that is both human-readable and machine-readable.

- Status of development: Normative
- Further Information: https://en.wikipedia.org/wiki/XML

5.1.3. Standard Data Element Sets

5.1.3.1. CCE - Common Condition of use Elements

This provides a core list of non-directional, atomic 'concepts' that specify types of use of an asset (e.g., the where, why, by whom, when etc). It aims to provide the basis for a common set of 'conditions of use' statements that resources (e.g., biobank, datasets, collections) can optionally assert, so that such information can be compiled, exchanged, aligned, and used for discovery and access decision making. These statements can be formulated as Digital Use Condition (DUC) constructs, and this is what is being employed presently to assess the utility of CCEs. Version 0.1 of CCE comprises 14 elements, applicable to registries and biobanks, which has been alpha-tested and is now in beta-testing.

- Status of development: Trial use
- Further Information:
  - https://docs.google.com/document/d/1Ejml2DFKcN5DMJ6qIIaAjqP2r35Nmi7RQmHrioQ8fhg/edit?usp=sharing

5.1.3.2. CDE - Common Data Elements

Widely used minimum set of data fields to be collected by EU Rare Disease registries comprising a set of 16 common data elements, released by the EU RD Platform aiming at increasing interoperability of RD registries. The CDEs ensure basic utility of RD registry data and help with standardization of datasets and basic interoperability in the sense of having the same types of data elements across registries. However, this provides no guarantee that the type of information is harmonized across registries by a uniform data model or that query APIs are uniform. This is achieved by making CDEs available in the form of the machine readable CDE semantic model (see 5.1.4.1) and providing a common query API.

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

5.1.4. Standards on Data Models

5.1.4.1. Basic Formal Ontology (BFO) and The Open Biological and Biomedical Ontology (OBO) Foundry ontologies

The BFO is focused on the task of providing the upper ontology as the foundation for all OBO Foundry ontologies that cover specific domains of scientific research, as for example in biomedicine. There are many OBO domain ontologies that together cover large parts of known biomedical reality. Among them HPO (Human Phenotypes), MONDO (Diseases), NCIT (Cancer and many associated concepts), and OBIB (biobanks). OBO foundry poses a set of restrictions and best practices to guarantee rigour and interoperability between OBO ontologies.

- Status of development: Normative
- Further Information: http://www.obofoundry.org/ontology/bfo.html; https://obofoundry.org/

5.1.4.2. Clinical Data Interchange Standards Consortium (CDISC)

The Clinical Data Interchange Standards Consortium (CDISC) develops and advances data standards to transform incompatible formats, inconsistent methodologies, and diverse perspectives into a framework for generating clinical research data.

- Status of development: Normative
- Further Information: https://www.cdisc.org/standards

5.1.4.3. bioCADDIE Data Tag Suite (DATS)

DATS, which stands for DAta Tag Suite, is a data description model designed and produced to describe datasets being ingested in DataMed, a prototype for data discovery developed as part of the NIH Big Data 2 Knowledge bioCADDIE project.

- Status of development: Normative
- Further Information: https://datatagsuite.github.io/docs/html/

5.1.4.4. DUC – Digital Use Conditions

This provides a standardised structure for expressing statements that specify conditions of use or consent. It comprises a Header (metadata) section, and a body of at least one statement. Each statement refers to a non-directional, atomic 'type of use' concept (see CCE), termed a 'Condition', with an optional 'Condition Detail' modifier that gives further details of that instance of that type of use. This is then given directionality via a 'Rule' that states whether that form of use is Obligated, Permitted or Forbidden, and this Rule can also be elaborated by an optional 'Other Considerations' option. Regularised statements about "Applicability" and "Scope" can also be provided. Version 0.1 of DUC using CCE has been alpha-tested and will now be piloted as a basis for representing CCE statements to underpin VP v1.0 discovery activities.

- Status of development: Trial use
- Further Information:
  - https://docs.google.com/document/d/1EjmI2DFKcN5DMJ6qIlaAjqP2r35Nmi7RQmHrioQ8fhg/edit?usp=sharing
  - https://docs.google.com/spreadsheets/d/1P23mP1ZCIcLPq50aIilg0yusMWOW8HewgIzp8TKNQ-s/edit?usp=sharing

5.1.4.5. JRC Common Data Element Semantic Data Model

Semantic Model for the set of 16 Common Data Elements (CDEs) for Rare Diseases Registration released by the EU RD Platform to increase interoperability of RD registries (see 5.1.3.1). It maps the CDEs and qualified relationships between them to standard ontologies[7]. The basis for each data element module is a semantic design pattern

[7] Rajaram Kaliyaperumal, Mark D. Wilkinson, *et al.*, Semantic modelling of Common Data Elements for Rare Disease registries, and a prototype workflow for their deployment over registry data. https://doi.org/10.1101/2021.07.27.21261169

defined by the Semanticscience Integrated Ontology (SIO)[8]. The model aims to represent, in a globally understood, machine readable language, what the values of CDE data records in registry datasets mean: a uniform way for computers to 'understand' data records across multiple registries. By annotating data records with the model *at source*, e.g., an RD registry, data records become globally linkable (interoperable) and machine actionable ('ontologised data'). The model is defined in RDF, similar to the other semantic models in this document (e.g., DCAT2, the EJP RD Metadata model, aforementioned ontologies). Serialisations in other formats are possible at the expanse of machine-readable semantics. It should be noted that the access conditions that apply to the original data should also be applied to the 'ontologised data'.

- Status of development: Normative
- Further Information: https://github.com/ejp-rd-vp/CDE-semantic-model

5.1.4.6. OMOP (OHDSI object model)

The OMOP Common Data Model allows for the systematic analysis of disparate observational databases. The concept behind this approach is to transform data contained within those databases into a common format (data model) as well as a common representation (terminologies, vocabularies, coding schemes), and then perform systematic analyses using a library of standard analytic routines that have been written based on the common format.

- Status of development: Normative
- Further Information: https://www.ohdsi.org/data-standardization/the-common-data-model/

5.1.4.7. Resource Description Framework (RDF) and Linked Data

The Resource Description Framework is a World Wide Web Consortium (W3C) Recommendation for representing information in the Web and expressing machine readable descriptions of resources. The building blocks are subject-predicate-object triples, where the elements may be hyper-links (typically Uniform Resource Identifiers, URIs), blank nodes, or data-typed literals (e.g., the values in registry records). RDF triples form graphs that express meaning for computers. The Web Ontology Language (OWL), used by most biomedical ontologies, is defined in RDF. By using URIs for connecting nodes and edges, RDF builds on web standards that have proven to scale to globally federated networks. RDF is closely associated with the Linked Data Principles: (i) use URIs as names for things; (ii) use HTTP URIs so that people can look up those names; (iii) provide useful information when someone looks up a URI, using the standards (RDF, SPARQL[9]); (iv) include links to other URIs, such that they can discover more things. In summary, the properties and purpose of RDF make it the default backbone for machine readable semantic models, including those used in the VP. NB RDF should not be confused with file formats: RDF is a data model. RDF graphs can be

[8] Dumontier, M., Baker, C.J., Baran, J. *et al.* The Semanticscience Integrated Ontology (SIO) for biomedical research and knowledge discovery. *J Biomed Semant* **5,** 14 (2014). https://doi.org/10.1186/2041-1480-5-14
[9] https://www.w3.org/TR/sparql11-overview/

serialised and exchanged in various formats (XML, JSON-LD, Turtle, etcetera). The W3C-recommended 'SPARQL' language is the default query language for RDF graphs.

- Status of development: Normative
- Further Information: https://www.w3.org/RDF/

5.1.5. Standards on data ontology, terminology and vocabulary

5.1.5.1. Anatomical Therapeutic Chemical (ATC)

The Anatomical Therapeutic Chemical (ATC) Classification System is used for the classification of active ingredients of drugs according to the organ or system on which they act and their therapeutic, pharmacological and chemical properties. It is controlled by the World Health Organization Collaborating Centre for Drug Statistics Methodology (WHOCC) and was first published in 1976.

- Status of development: Normative
- Further Information:
  - https://www.who.int/tools/atc-ddd-toolkit/atc-classification
  - https://bioportal.bioontology.org/ontologies/ATC

5.1.5.2. Data Catalogue Vocabulary (DCAT)

Data Catalog Vocabulary (DCAT) is an RDF vocabulary designed to facilitate interoperability between data catalogues published on the Web. By using DCAT to describe datasets in catalogues, publishers increase discoverability and enable applications to consume metadata from multiple catalogues. It enables decentralized publishing of catalogues and facilitates federated dataset search across catalogues. Aggregated DCAT metadata can serve as a manifest file to facilitate digital preservation

- Status of development: Normative
- Further Information: https://www.w3.org/TR/vocab-dcat-2/

5.1.5.3. Data Use Ontology (DUO)

DUO is an ontology which represent data use conditions. DUO allows to semantically tag datasets with restriction about their usage, making them discoverable automatically based on the authorization level of users, or intended usage. It is a GA4GH approved standard

- Status of development: Normative
- Further Information: https://www.ga4gh.org/news/data-use-ontology-approved-as-a-ga4gh-technical-standard/
  https://obofoundry.org/ontology/duo.html

5.1.5.4. Gene Ontology (GO)

The Gene Ontology (GO) knowledgebase is the world's largest source of information on the functions of genes. This knowledge is both human-readable and machine-readable, and is a foundation for computational analysis of large-scale molecular biology and genetics experiments in biomedical research.

- Status of development: Normative
- Further information: http://geneontology.org

5.1.5.5. GENO

Genotype Ontology, an integrated ontology for representing the genetic variations described in genotypes, and their causal relationships to phenotype and diseases.

- Status of development: Draft
- Further Information:
  - https://www.ebi.ac.uk/ols/ontologies/geno
  - https://bioportal.bioontology.org/ontologies/GENO

5.1.5.6. HGNC (HUGO)

HUGO Gene Nomenclature Committee is the resource for approved gene nomenclature, and is part of the CDE for rare disease registration (see 5.1.4.1).

- Status of development: Normative
- Further Information: https://www.genenames.org/

5.1.5.7. Sequence Variant Nomenclature (HGVS)

The HGVS Nomenclature is a set of recommendations when describing sequence variant in a consistent and unambiguous manner to facilitate the report and exchange of information on the analysis of a genome. It is an IRDiRC Recognized Resource and is part of the CDE for rare disease registration (see 5.1.4.1).

- Status of development: Normative
- Further Information: http://varnomen.hgvs.org

5.1.5.8. HPO-ORDO ontological module (HOOM)

HOOM is a module that qualifies the relationship between a clinical entity and phenotypic abnormalities according to its frequency of occurrence in the disease population and further qualifiers such as "pathognomonic sign" and "diagnostic criterion". It is based on Orphanet knowledge base of annotations of rare diseases by its phenotypes, using Human Phenotype Ontology (HPO)

- Status of development: Normative
- Further Information:

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

5.1.5.9. Human Phenotype Ontology (HPO)

The Human Phenotype Ontology (HPO) provides a standardized vocabulary of phenotypic abnormalities encountered in human disease. It is an IRDiRC Recognized Resource and is part of the CDE for rare disease registration (see 5.1.4.1). The HPO is currently being developed using the medical literature, Orphanet, DECIPHER, and OMIM.

- Status of development: Normative
- Further Information: https://hpo.jax.org/app/

5.1.5.10. International Classification of Diseases (ICD)

International Classification of Diseases developed and maintained by the World Health Organisation (WHO), integrated in the WHO's Family of International Classifications (WHO-FIC). ICD is historically used for statistical reports on mortality and morbidity and is the leading classification used in health information systems across the world in its different versions and national adaptations. Current ICD international versions in use are ICD-10 and ICD-11.

- Status of development: Normative
- Further Information: https://icd.who.int

5.1.5.11. International Classification of Functioning, Disability and Health (ICF)

Part of the WHO' Family of International Classifications (WHO-FIC), ICF is the WHO framework for measuring health and disability at both individual and population levels. It is recommended in the CDE for rare disease registration (see 5.1.4.1).

- Status of development: Normative
- Further Information: https://www.who.int/standards/classifications/international-classification-of-functioning-disability-and-health

5.1.5.12. International Classification of Diseases for Oncology (ICD-O)

Part of the WHO' Family of International Classifications (WHO-FIC), ICD-O is a multi-axial classification of the site, morphology, behaviour, and grading of neoplasms. It is used principally in tumour or cancer registries for coding the site (topography) and the histology (morphology) of neoplasms, usually obtained from a pathology report. Current version is ICD-O-v3.

- Status of development: Normative
- Further Information: https://www.who.int/standards/classifications/other-classifications/international-classification-of-diseases-for-oncology

5.1.5.13. Informed Consent Ontology (ICO)

Informed Consent Ontology (ICO) is a community-based ontology in the domain of informed consent. It is an OBO library ontology and developed by following the OBO Foundry principles.

- Status of development: Normative
- Further Information:
  - https://obofoundry.org/ontology/ico.html
  - https://bioportal.bioontology.org/ontologies/ICO

5.1.5.14. LOINC

Logical Observation Identifiers Names and Codes. The international standard for identifying health measurements, observations, and documents.

- Status of development: Normative
- Further Information: https://loinc.org/

5.1.5.15. MedDRA - Medical Dictionary for Regulatory Activities Terminology

Rich and highly specific standardised medical terminology to facilitate sharing of regulatory information internationally for medical products used by humans

- Status of development: Normative
- Further Information: https://www.meddra.org/

5.1.5.16. MeSH - Medical Subject Headings

MeSH is a comprehensive controlled vocabulary for the purpose of indexing journal articles and books in the life science.

- Status of development: Normative
- Further Information: https://meshb.nlm.nih.gov/

5.1.5.17. NCIT - National Cancer Institute Thesaurus

A vocabulary for clinical care, translational and basic research, and public information and administrative activities.

- Status of development: Normative
- Further Information:
  - https://ncithesaurus.nci.nih.gov/ncitbrowser/
  - https://bioportal.bioontology.org/ontologies/NCIT

5.1.5.18. Online Mendelian Inheritance in Man (OMIM)

Online Mendelian Inheritance in Man is an online catalogue of human genes and genetic disorders. It is an IRDiRC Recognized Resource.

- Status of development: Normative
- Further Information: https://omim.org

5.1.5.19. Orphanet nomenclature of rare diseases (ORPHAcodes) and Orphanet ontology of rare diseases (ORDO)

Orphanet nomenclature of rare diseases (ORPHAcodes) and its enriched ontological format Orphanet Rare Diseases Ontology (IRDiRC Recognized Resource). Orphanet nomenclature is a standardised vocabulary allowing semantic annotation of rare disease diagnosis and is part of the CDE for rare disease registration (see 5.1.4.1). ORDO includes alignments between ORPHAcodes and other medical terminologies (ICD-10, OMIM, MedDRA, UMLS, MeSH), gene-disease relationships, disease epidemiological data by geographical location.
The Orphanet nomenclature of rare diseases and its alignments with other terminologies files, including SNOMED CT, are also released in XML and JSON formats in Orphadata.org, recognised ELIXIR Core Data Resource. A dedicated API and a Data Visualisation tool are also available.
Orphanet nomenclature and ORDO are multilingual resources.

- Status of development: Normative
- Further Information:
  - www.orphadata.org
  - http://bioportal.bioontology.org/ontologies/ORDO

5.1.5.20. PROV Ontology

It provides a set of classes, properties, and restrictions that can be used to represent and interchange provenance information generated in different systems and under different contexts

- Status of development: Normative
- Further Information: https://www.w3.org/TR/prov-o/

5.1.5.21. Semanticscience Integrated Ontology (SIO)

The Semanticscience Integrated Ontology (SIO) is an ontology to facilitate biomedical knowledge discovery. SIO features a simple upper level comprised of essential types and relations for the rich description of arbitrary (real, hypothesized, virtual, fictional) objects, processes and their attributes. SIO specifies simple design patterns to describe and associate qualities, capabilities, functions, quantities, and informational entities including textual, geometrical, and mathematical entities, and provides specific extensions in the domains of chemistry, biology, biochemistry, and bioinformatics.

- Status of development: Normative

- Further Information:
  - https://github.com/MaastrichtU-IDS/semanticscience
  - https://jbiomedsem.biomedcentral.com/articles/10.1186/2041-1480-5-14

5.1.5.22. SNOMED CT

SNOMED Clinical Terms is a systematically organized computer processable collection of medical terms providing codes, terms, synonyms and definitions used in clinical documentation and reporting. SNOMED CT is considered to be the most comprehensive, multilingual clinical healthcare terminology in the world. The primary purpose of SNOMED CT is to encode the meanings that are used in health information and to support the effective clinical recording of data with the aim of improving patient care. SNOMED CT provides the core general terminology for electronic health records. SNOMED CT comprehensive coverage includes clinical findings, symptoms, diagnoses, procedures, body structures, organisms and other aetiologies, substances, pharmaceuticals, devices and specimens.

- Status of development: Normative
- Further Information: https://www.snomed.org/

5.1.6. Standards on data discovery

5.1.6.1. Automatable Data Discovery and Access Matrix (ADA-M)

Automatable Data Discovery and Access Matrix (GA4GH Standard) Comprehensive information model that provides the basis for producing structured metadata "Profiles" of data access regulatory conditions (e.g., Informed consent information in a machine-readable format).

- Status of development: Normative
- Further Information: https://github.com/ga4gh/ADA-M

5.1.6.2. Beacon-1

The Beacon protocol defines an open standard for genomics data discovery, developed by members of the Global Alliance for Genomics & Health.

- Status of development: Normative
- Further Information: https://github.com/ga4gh-beacon/specification

5.1.6.3. Beacon-2 API

GA4GH API specification for querying records on attributes relating to genotype, phenotype, sample, demographics, or (gen)omics data. Validated extensions enable interaction with many other areas of standardisation in GA4GH. It is one way that federated queries may be structured. It is one way that federated queries may be structured.

- Status of development: Trial use

- Further Information: https://beacon-project.io/

5.1.6.4. Bioschemas

Bioschemas aims to improve the Findability on the Web of life sciences resources such as datasets, software, and training materials. It does this by encouraging people in the life sciences to use Schema.org markup in their websites so that they are indexable by search engines and other services.

- Status of development: Normative
- Further Information: https://bioschemas.org/

5.1.6.5. VP index API

A draft of the EJP-RD - 'Central catalogue directory' component. It implements a REST API that can be used to fetch, add, and remove catalogue addresses via HTTP requests.

- Status of development: Draft
- Further Information: https://github.com/ejp-rd-vp/query-builder-catalogue-directory

5.1.7. Standards on data exchange mechanisms

5.1.7.1. DOIP - Digital Object Interface Protocol

The Digital Object Interface Protocol (DOIP) is a simple, but powerful conceptual protocol for software applications ("clients") to interact with "services" which could be either the digital objects or the information systems that manage those digital objects.

- Status of development: Normative
- Further Information: https://www.dona.net/doipv1doc

5.1.7.2. FHIR - Fast Healthcare Interoperability Resources

Standard for exchanging healthcare information electronically.

- Status of development: Normative
- Further Information: https://www.hl7.org/fhir/overview.html

5.1.7.3. mzML/mzIdentML

ProteomeXchange Consortium was established to provide globally coordinated standard data submission and dissemination pipelines involving the main proteomics repositories, and to encourage open data policies in the field

- Status of development: Normative

- Further Information:
  - http://www.psidev.info/mzidentml
  - www.proteomexchange.org

5.1.7.4. PhenoPackets

This standard (ISO/ WD 4454) enables the exchange of clinical phenotype related information between information systems within EJP RD and with other project (such as Solve-RD), not least to/from ERN registries, RD-NEXUS discovery tools, Linked Data Platforms, and GPAP. Its scope includes data on individuals and family/pedigree information. It is compatible with ontologies such as the HPO, ORDO and OMIM and with genetic data. It is a GA4GH approved standard.
A semantic model of PhenoPackets has been developed within EJP RD.

- Status of development: Normative

- Further Information: http://phenopackets.org

5.1.7.5. REST

Representational State Transfer (Design Guide)

- Status of development: Normative

- Further Information: https://en.wikipedia.org/wiki/Representational_state_transfer

5.1.8. Standards on security, authentication, and authorisation

5.1.8.1. GA4GH Passport

A standard for a global federated data sharing network that allows the searching, and subsequent -optional- processing of the results in a cloud environment.

- Status of development: Normative

- Further Information: https://github.com/ga4gh-duri/ga4gh-duri.github.io/blob/master/researcher_ids/ga4gh_passport_v1.md

5.1.8.2. GA4GH Visa

An assertion from a Passport Visa Assertion Source organization that is bound to a Passport Visa Identity and signed by a Passport Visa Issuer service whose signature is verifiable via its public key.

- Status of development: Normative

- Further Information: https://github.com/ga4gh-duri/ga4gh-duri.github.io/blob/master/researcher_ids/ga4gh_passport_v1.md#passport-visa

5.1.8.3. Open ID Connect

OpenID Connect is the third generation of OpenID technology. It is an authentication layer on top of the OAuth 2.0 authorization framework.[82] It allows computing clients to verify the identity of an end user based on the authentication performed by an authorization server, as well as to obtain the basic profile information about the end user in an interoperable and REST-like manner. In technical terms, OpenID Connect specifies a RESTful HTTP API, using JSON as a data format.

- Status of development: Normative
- Further Information: https://en.wikipedia.org/wiki/OpenID#OpenID_Connect_(OIDC)

5.1.8.4. SAML 2.0

Security Assertion Markup Language 2.0 (SAML 2.0) is a version of the SAML standard for exchanging authentication and authorization identities between security domains. SAML 2.0 is an XML-based protocol that uses security tokens containing assertions to pass information about a principal (usually an end user) between a SAML authority, named an Identity Provider, and a SAML consumer, named a Service Provider. SAML 2.0 enables web-based, cross-domain single sign-on (SSO), which helps reduce the administrative overhead of distributing multiple authentication tokens to the user.

- Status of development: Normative
- Further Information: https://en.wikipedia.org/wiki/SAML_2.0

5.2. Tools

The following tools represent examples of tools that were agreed to be used within the VP. However, this list is not intended to be complete, since any tool that fulfils requirements specified in the VIPS can be included in the VP.

5.2.1. Authentication and Authorization Infrastructure

5.2.1.1. LifeScience AAI

LifeScience Authentication and Authorisation Infrastructure

- Status of development: trial use
- Further Information: https://elixir-europe.org/about-us/commissioned-services/identity-access

5.2.2. Pseudonymisation

5.2.2.1. EUPID

European Unified Patient Identifier (EU Standard) International unique global identifier systems for patients (EUPID) - an EJP-RD supported method for encrypting patient IDs, to help protect patient identity whilst still being able to track and connect patient records (available via ERDRI). Allows to count patients among European registries avoiding double counts.

- Status of Development: Trial use
- Further Information: https://eupid.eu/

5.2.3. Resource FAIRness assessment service

5.2.3.1. Data Model Alignment Service

The CDE semantic model was built to represent, in 'ontologised' linked data form, the CDEs defined by JRC for RD registries. The aforementioned data standards are commonly used for health data and some of them are already adopted by ERN registries. Mapping of the CDE semantic model to/from the 3 data standards (CDISC, ODHSI OMOP and FHIR) is therefore needed. Therefore, a data model alignment service is envisioned comprising a mapping table for CDE terms and scripts for transforming data into the standards, and vice versa. The CDE model uses ontological terms from various ontologies to represent the CDEs. These ontological terms are to be mapped to the ontologies or terminologies used in the 3 standards, presented in a mapping table.

- Status of development: Draft
- Further Information: see VIPS

5.2.3.2. FAIR Data Point (FDP)

FAIR Metadata Endpoint Serves as a promoter of your registry, increases Findability at machine readable level. Accessibility rules to your dataset are also available in the FDP. At the moment, it works at metadata level.

- Status of development: Normative
- Further Information: https://fairdatapoint.readthedocs.io/en/latest/index.html

5.2.4. Consent and Use Conditions

5.2.4.1. CCE / DUC Profile Creation Tool [ULEIC]

An online form that enables users to created DUC structured 'Profiles' for CCE elements, to support the testing and adoption of these standards. The Leicester version of this tool is a standalone service that does not require (but does permit) users to save a copy of the created Profile on the server, with creation date and version. These can be retrieved later and further edited. Download formats include JSON, CSV and TXT. Users can only view and access Profiles that they themselves created.

- Status of development: Trial use
- Further Information: https://duc.le.ac.uk/

5.2.4.2. CCE / DUC Profile Creation Tool [UMCG]

An online form that enables users to created DUC structured 'Profiles' for CCE elements, to support the testing and adoption of these standards. The UMCG version of this tool is integrated into the MOLGENIS platform, and thereby provides a local service for groups whose registry is built on the MOLGENIS platform. Access requires login to their MOLGENIS account

- Status of development: Trial use
- Further Information: https://irdirc.molgeniscloud.org/menu/main/home

5.2.5. Record linkage services

5.2.5.1. EUPID

See above (Tools / Pseudonymization / EUPID).

5.2.6. Resource/Data/Sample Discovery and Access

The following list of resource/data/sample discovery and access tools represents examples that are not intended to be complete. However, these tools represent references that can help by other tool providers to link their tools to the VP. Additionally, they might be used as hubs to the VP for other resource/data/sample providers that prefer to connect to those tools than to link with the VP directly. Detailed lists of resources are described in Del. 11.8 and 11.18.

5.2.6.1. bio.tools

bio.tools is an open source, open data registry of biological and biomedical software descriptions used to help researchers find, understand, utilise and cite the resources they need in their day-to-day work.
Resources in bio.tools include everything from simple command-line tools and scripts, through to databases and complex, multi-functional analysis workflows. Resources are described in a rigorous semantics and syntax, providing end-users with the convenience of concise, consistent, and therefore comparable information.
An entry in bio.tools is assigned a human-readable unique identifier which provides a persistent reference to the resource even after the resource is no longer accessible. In this way bio.tools preserves information and ensures citations are always available.

- Status of Development: Normative
- Further information: https://bio.tools

5.2.6.2. CDE in a box

CDE in a box is a tool suite for generating, storing, and publishing common data elements (CDEs) according to the CDE semantic model. The suite performs a data uplifting activity - it takes CSV files as inputs and generates CDE model-compliant semantic RDF files and stores them in a secure triplestore. CDE in a box is composed of three major components: a triplestore to store ontologised version of the CDE dataset and its metadata, FAIR Data Point software to publish the metadata of the CDE dataset, and a transformation service to convert CDEs that are provided to the CDE in a box as CSV files. The components used in the CDE in a box are dockized. Users can easily deploy these components on servers that have the docker engine installed.

- Status of Development: Normative
- Further Information: https://github.com/ejp-rd-vp/cde-in-box

5.2.6.3. The European Genome-Phenome Archive (EGA)

The European Genome-phenome Archive is a resource for long term secure archiving of all types of potentially identifiable genetic, phenotypic, and clinical data resulting from biomedical research projects. Its mission is to foster hosted data reuse, enable reproducibility, and accelerate biomedical and translational research in line with the FAIR principles. Launched in 2008, the EGA has grown quickly, currently archiving over 4,500 studies from nearly one thousand institutions. The EGA operates a distributed data access model in which requests are made to the data controller, not to the EGA, therefore, the submitter keeps control on who has access to the data and under which conditions. Given the size and value of data hosted, the EGA is constantly improving its value chain, that is, how the EGA can contribute to enhancing the value of human health data by facilitating its submission, discovery, access, and distribution, as well as leading the design and implementation of standards and methods necessary to deliver the value chain. The EGA has become a key GA4GH Driver Project, leading multiple development efforts and implementing new standards and tools, and has been appointed as an ELIXIR Core Data Resource.

- Status of Development: Normative

- Further Information: https://ega-archive.org/

5.2.6.4. HPSCreg

The human pluripotent stem cell registry (hPSCreg) was founded in 2007 by the European Commission to monitor and inform the EC about the state of human embryonic stem cell (hESC) research carried out in the EU. Since its inception, hPSCreg has expanded to include human induced pluripotent stem cells (hiPSC), which share many of the same biological properties of hESCs, but due to their somatic cell origin, are not associated with the ethical issues of embryo destruction to generate hESC lines. The Registry collects data on the ethical provenance of biological material used to derived hPSC lines, information on how the hPSC lines have been derived and genetically engineered (if applicable), as well as characteristics demonstrating the pluripotency of the hPSC lines. Data on stem cell lines are collected according to stem cell community standards (e.g., standard cell line nomenclature, characterisation of pluripotency) and the use of relevant ontologies to richly annotate the dataset with metadata (e.g., Disease Ontology, Orphanet Rare Disease Ontology, Cell Ontology, and others provided by EMBL-EBI Ontology Lookup Service). In the last few years, the resource has been increasingly FAIRified in its interactions with other European projects such as Cellosaurus, EBiSC, and FAIRplus. hPSCreg issues cell line certificates attesting to the ethical provenance and biological properties of the hPSC lines. These certificates are recognized by the EC to fulfil ethical requirements for EC-funded research. The Registry contains over 3700 hPSC lines from more than 30 countries, of which approximately 25% originate from donors with rare disease. To further track hPSC lines into their translational applications, hPSCreg established a clinical study database containing only trials that use hPSC cells or their derived cell types in interventional therapeutic applications. As of November 2021, the clinical study database had 87 clinical studies, of which 35% were used to treat rare disease (by ICD-10 codes) listed in Orphanet. hPSCreg is an International Rare Diseases Research Consortium (IRDiRC) recognized resource.

- Status of Development: Normative
- Further information: https://hpscreg.eu/

5.2.6.5. INFRAFRONTIER

INFRAFRONTIER is the European Research Infrastructure for the generation, phenotyping, archiving and distribution of model mammalian genomes. The core services of INFRAFRONTIER comprise the systemic phenotyping of mouse mutants in the participating mouse clinics, and the archiving and distribution of mouse mutant lines by the European Mouse Mutant Archive (EMMA). INFRAFRONTIER aids in rare disease research by providing access to nearly 1800 mouse strains (via EMMA) that are related to over 1400 distinct rare diseases. In addition, INFRAFRONTIER provides specialized services such as the generation of germ-free mice (axenic service) and training in state-of-the-art cryopreservation and phenotyping technologies. The INFRAFRONTIER/EMMA DB has been selected as a FAIRified data resource by the international FAIRSharing Consortium.

- Status of Development: Normative

- Further information: https://www.infrafrontier.eu

5.2.6.6. MOLGENIS

MOLGENIS is a generic, open source and free to use data platform for researchers to accelerate scientific collaborations and for bioinformaticians. MOLGENIS enables its users to quickly create FAIR database online to find, capture, exchange, manage and analyse a wide diversity of scientific data. MOLGENIS is fully customizable: data structure, user interface and layout can be fully changed and custom (bioinformatics) scripts can be plug-in. It provides Excel/CSV based option to configure the tables, column and relations, and then provides generic APIs and user interfaces to query, upload and download data using REST, CSV, Excel. Also, MOLGENIS provides options to extend functionality using 'apps', I.e., small JavaScript+html applications that provide a rich and specialised user experience. It is built on industry standard java, JavaScript, REST, PostgreSQL and elasticsearch.
In EJP-RD, MOLGENIS is being used for BBMRI-ERIC Directory, RD-connect sample database, Sandbox/RD3, and IRDIRC consent database described above. In the broader EJP-RD community, MOLGENIS is used for development of rare disease patient registries for Ithaca, SKIN, CRANIO and Genturis. Beyond EJP-RD, MOLGENIS is also used for multi-centre cohort study and real-world evidence catalogues. All these systems gain interoperability potential into EJP-RD by having federated AAI, FAIR data point and REST apis. This year, the MOLGENIS team has been piloting also semantic extensions that would further ease integration of MOLGENIS based resources into the virtual platform. MOLGENIS is available free for local installation (support via molgenis-support@umcg.nl) but is also provided as 'SaaS' (software as a service) to project partners in EJP-RD or for a fee for outside users.

- Status of Development: Normative

- Further information:
  - https://www.molgenis.org/
  - http://github.com/molgenis

5.2.6.7. RD-Connect GPAP

The RD-Connect GPAP is a sophisticated and user-friendly online analysis system for RD gene discovery and diagnosis. The RD-Connect GPAP is an IRDiRC recognized resource hosted at the CNAG-CRG. De-identified phenotypic data is collected using HPO, ORDO and OMIM ontologies through custom templates implemented through the RD-Connect GPAP-Phenostore module. Pseudonymized experiment data (exomes and genomes) and metadata are collected in the RD-Connect GPAP, and processed using a standardized analysis and annotation pipeline. Integrated genome-phenome results are made available to authorized users for prioritisation and interpretation of genomic variants in the RD-Connect GPAP. Raw genomic data is deposited at the EGA for long-term archive and controlled access.

- Status of Development: Normative

- Further information: https://platform.rd-connect.eu

5.2.6.8. Rare Disease Networked Exploration of the UnSeen (RD-NEXUS)

A fully featured, modular, secure platform that supports federated discovery networks, whilst being agnostic to the type of asset or data structures/models being interrogated to enable discovery queries. May be installed locally, or hosted elsewhere (e.g., at ULEIC), while all admin actions, data controllership, and datasets remain with the installation and the relevant PIs. Containerised versions available, optionally integrated with the CDE in a box software. Fully customisable in terms of query interfaces, user accounts and permissions, discoverable data sources, threshold counts for positive searches, response types per user (from simple links through to yes/no indicators, count responses, handoffs, and data provisioning) as well as the particular networks that each installation contributes to. It employs multiple query engines synchronously (SQL, ElasticSearch, Neo4J, BCFTools) and is compliant with all relevant global and EJP-RD standards. Can operate a proxy client to enable integration of triple-stores into federated discovery networks, and allows for interrogation of one or more datasets included in each RD-NEXUS installation (obfuscated if necessary) as well as connection to primary databases. Exceptionally powerful and flexible capabilities in terms of 'similarity' searching, with options to save, share and re-use queries.

- Status of Development: Normative
- Further Information: https://github.com/Cafe-Variome/RDNexus