FDPcrawleR: A Lightweight R Framework for Auditing FAIR Data Points and FAIR Virtual Platforms

Kristina Vodorezova[1,*,†], Alberto Cámara[2,†], Nirupama Benis[1], Andra Waagmeester[1], Mark D. Wilkinson[2] and Ronald Cornet[1]

[1]*Department of Medical Informatics, Reusable Health Data group, Amsterdam Public Health Research Institute, Methodology Digital Health, Location AMC Meibergdreef 9, 1105 AZ Amsterdam, Netherlands*

[2]*Departamento de Biotecnología-Biología Vegetal, Escuela Técnica Superior de Ingeniería Agronómica, Alimentaria y de Biosistemas, Centro de Biotecnología y Genómica de Plantas (CBGP). Universidad Politécnica de Madrid (UPM) - Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria-CSIC (INIA-CSIC). Pozuelo de Alarcón (Madrid), Madrid, Spain*

Abstract

Rare disease research is hindered by the fragmentation of data across resources, making it essential to expose well-structured and interoperable metadata. FAIR Data Points (FDPs) offer a mechanism for publishing FAIR machine-readable (meta)data; however, in practice, the usability of FDPs for data federation and content-discovery depends on the quality of the metadata. To address this issue, this work presents an automated method for metadata completeness check across FDPs based on the FDP Index utilized by ERDERA Virtual Platform. The analysis reveals substantial omissions in metadata population, with only a minority of FDPs containing metadata elements that reference a URL for data access. Such gaps directly hinder meaningful federated discovery and are particularly problematic in the rare disease context, where dispersed and scarce datasets benefit from federation in terms of improved findability and reuse. These results highlight the need for enhanced metadata stewardship and FDP validation workflows. Overall, the paper presents a metadata completeness-check dashboard that helps strengthen FAIR metadata quality and supports more effective discovery across federated rare disease data platforms.

Keywords

Metadata, FAIR, FAIR Data Point, Data federation, Federated discovery, Data stewardship

1. Introduction

Rare diseases affect up to 30 million people in the EU [1]. Although the overall burden is large, each distinct rare disease has very low prevalence (less than 5 per 10000 people) [2]. According to the European Rare Diseases Research Alliance (ERDERA) [3], there are around 7000 rare diseases, with fewer than 5% of these conditions having approved therapy. Rare disease research constantly faces data challenges [4]. Rare disease data is inherently sensitive, as required by health data regulations, and is characterized by small cohort sizes, geographically dispersed knowledge, and considerable institutional fragmentation. Safe integration of such type of data can be yielded by a data federation approach where source data is mapped into an overarching metadata schema while remaining at the original storage site [5].

Metadata (data about data) is structured information that describes data content, context, provenance and accessibility terms, enhancing data for easier findability and further reuse [6]. Therefore, metadata plays an important role in managing rare disease data as it is essential for enabling federated discovery of rare disease data across distributed infrastructures.

The FAIR guiding principles [7] highlight the significance of machines in a data-rich research environment. Rich metadata that follows FAIR principles allows machines running automation workflows to determine which datasets exist, how they are described and how they can be accessed, thus making

17th International SWAT4HCLS Conference, March 23–26, Amsterdam, The Netherlands

*Corresponding author.

†These authors contributed equally.

k.vodorezova@amsterdamumc.nl (K. Vodorezova)

0000-0003-3518-2572 (K. Vodorezova); 0000-0001-5613-9704 (A. Cámara); 0000-0002-2101-6154 (N. Benis); 0000-0001-9773-4008 (A. Waagmeester); 0000-0001-6960-357X (M. D. Wilkinson); 0000-0002-1704-5980 (R. Cornet)

metadata machine-actionable. One solution for publishing FAIR metadata is a FAIR Data Point (FDP) [8]. An FDP is a Web service that publishes metadata in the form of Resource Description Framework (RDF) triples [9], a W3C standard for modelling linked data as a graph following DCAT-2 Vocabulary [10] as a schema and the Linked Data Platform (LDP) [11] to structure the hierarchy of the records.

To facilitate data discovery and federation in rare disease research, the ERDERA Virtual Platform (VP) was built with the FDP as the central solution. This platform was realized within the ERDERA project, building on solid foundations of the European Joint Programme on Rare Diseases (EJP RD) and aiming to advance rare disease research by harnessing the potential of health and research data. The purpose of the VP is to enable federated discovery and data visiting across rare disease resources for ERDERA partners by navigating their metadata and providing access to the already FAIRified rare disease data [12]. The VP Portal [13] is an interface to the VP, which allows for finding relevant rare disease data sources based on metadata of interest represented as ontology terms, like rare disease name, OrphaCodes [14], ICD-10 identifiers or gene name and symbol. Under the hood of the VP Portal, there is the VP Network, a network of rare disease relevant FDPs, indexed in the VP Index [15] (the FDP index where the VP looks for FDPs) and configured to be discoverable by the VP.

In the VP, two types of discovery are possible, namely, resource discovery and content discovery. The former supports metadata-based resource discovery, where a resource publishes DCAT-based machine-interpretable descriptions following the EJP RD VP metadata schema [16]. Resource discovery allows answering questions in the VP Portal that contain the metadata provided by the aforementioned metadata schema, like: "What are the URLs of the available catalogs in the VP?". The second type of connection enables content-based discovery, where a resource can answer privacy-preserving queries that return aggregated answers (yes/no or counts) through Beacon v4 endpoints [17] without exposing underlying

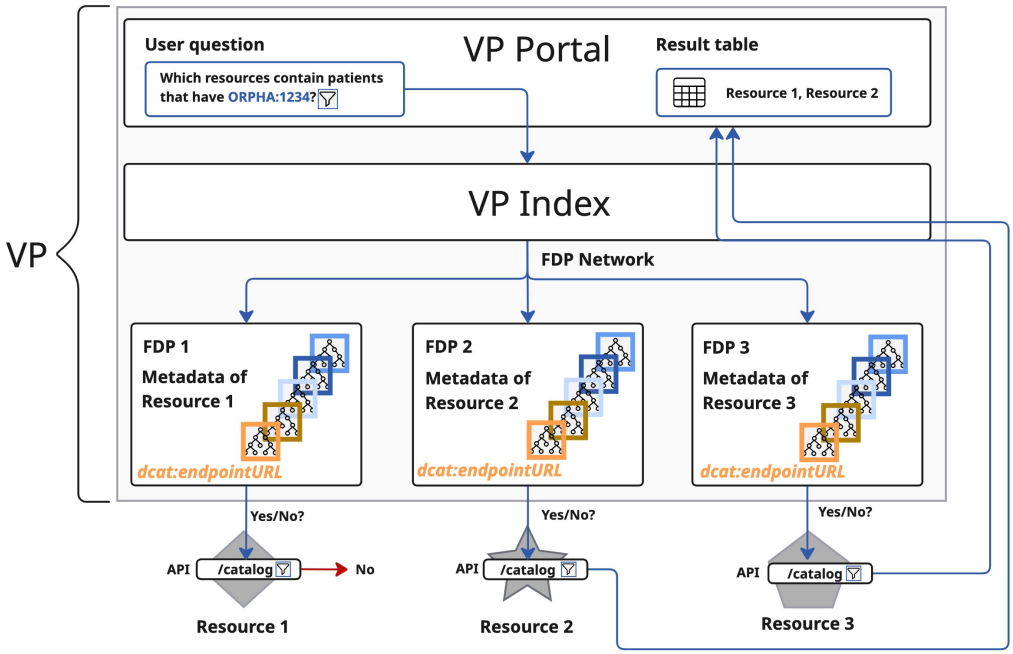

Figure 1: Federated discovery workflow in the ERDERA Virtual Platform (VP). Figure 1 illustrates how the VP enables federated content discovery of rare disease resources using an ontology term as a search filter. A user submits a query through the VP Portal, which is forwarded to the VP Index. The index routes the query to all indexed FDPs that expose metadata describing an actionable access endpoint via dcat:accessURL. For each FDP, the VP follows this URL and proceeds with the corresponding API provided locally on top of the resource data. For content discovery, the /catalogs endpoint is queried to determine whether the resource has the requested ontology term. Resources that return a positive API response are aggregated by the VP Index and presented in the results table. A similar workflow applies to count-based queries, but for the /individuals endpoint.

data by combining Beacon facets (e.g., sex, age, diagnosis). Hayn [18] proposed a VP architecture for content-discovery requests such as e.g. counting patients across VP resources, answering "How many patients have a diagnosis matching this OrphaCode?" in the VP Portal.

When a user performs a search on the VP Portal, the queries containing ontology terms are sent to Beacon API services. There are 2 types of Beacon endpoints: the first type (/catalogs endpoint) enables querying the summary metadata by facets provided according to the VP API specification [19]. The workflow for routing a user search via /catalogs endpoint is shown in Figure 1. The second type of Beacon endpoint (/individuals endpoint) enables querying privacy-preserving aggregations of data by another independent set of facets. In the end of EJP RD, it was noticed that there was a lack of patient data made available and the adoption of both types of endpoints (/catalogs and /individuals) was limited.

In this paper, we discuss the metadata architecture of FDPs indexed by the VP Index. We introduce a metadata completeness check script for assessing the metadata content of each FDP, together with a visualization that illustrates the current level of metadata population across the VP Index. We are interested in actual accessibility of resources for content-based discovery using the present metadata.

2. Methods

FDP metadata is based on the DCAT-2 vocabulary and follows a hierarchical structure defined by the LDP. In this structure, the RDF representation of an FDP contains URLs to all included catalogs, and each catalog, in turn, is described by RDF that contains URLs to its datasets. Therefore, FDPs follow a hierarchical metadata model: FDP -> Catalog -> Dataset -> Distribution -> Data service. FDP uses hierarchical DCAT-structured records, allowing one catalog to contain many datasets, one dataset to have many distributions, and one distribution to have many data services. Important concept to highlight: the DCAT-2 vocabulary defines its dataset entity as an abstract representation of data, whereas distribution is defined as a specific representation of a dataset. There are two kinds of data services: those that depend on a dataset, and those that exist independently of a specific dataset (i.e., analytic tools). The former represents the way to access the contents of a dataset in the representation defined in the distribution by providing a URL either to a downloadable file or to an API endpoint (e.g. Beacon or SPARQL [20]) in the form of DCAT property dcat:endpointURL.

We have developed an R-script that checks the completeness of DCAT-structured metadata of all valid FDPs in the VP Index. Note that in the VP Index, there are five categories of FDPs: Active (valid FDPs that have been registered in the index, and have properly sent a ping to the index in the last 7 days to indicate they are still operating), Inactive (valid FDPs that have been registered in the index, but have failed to properly send a ping to the index in the last 7 days – often due to misconfigured settings), Unreachable (FDPs whose URL cannot be resolved – this often happens because of expired security certificates), Invalid (FPDs whose metadata doesn't follow the FDP reference implementation), and Unknown (FDPs that do not fall into any of the previous categories). Only Active and Inactive FDPs are considered valid, and therefore are the only ones taken into account for the purposes of this paper. Although the contents of inactive FDPs are temporarily unavailable to the VP, we assessed whether their level of metadata population would allow them to participate in content discovery once they are correctly reconfigured to send a ping to the index.

Figure 2 details how the script implements a web-crawler for FDPs, using the VP Index URL as an entry point and DCAT-2 and FDP Ontology (FDP-O) terms to navigate in the hierarchical metadata structure. Its final goal is to quantify how many FDPs have metadata at different levels of the FDP metadata hierarchy (technically, a partonomy) iteratively going from the VP Index to FDPs, from FDPs to Catalogs, from Catalogs to Datasets, from Datasets to Distributions, and finally from Distributions to Data services. The results of the analysis are visualized as a Sankey diagram (Figure 3). The code repository is available on GitHub [21].

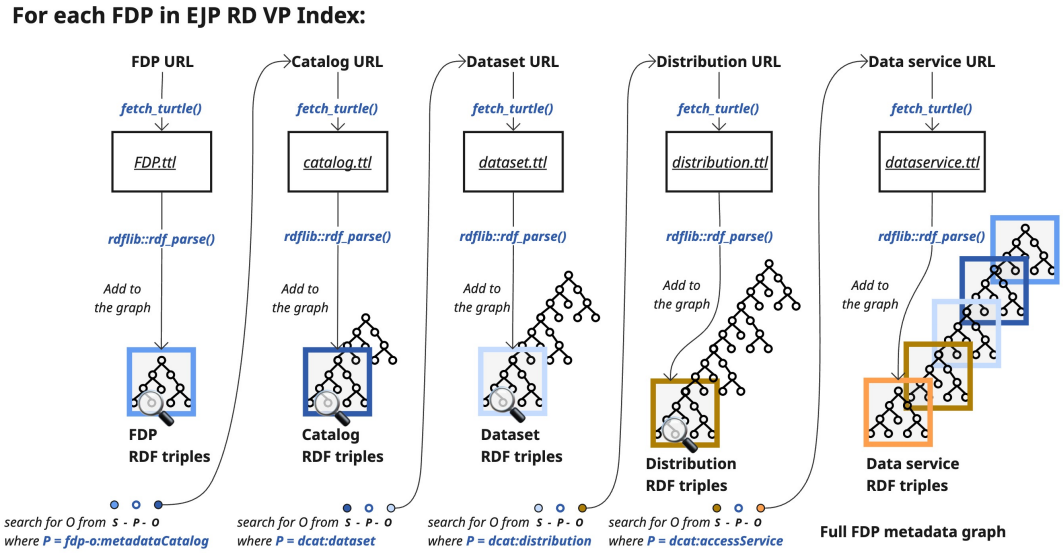

Figure 2: Workflow for building the full FDP metadata graph. The script iteratively retrieves and parses Turtle metadata from each FAIR Data Point, then follows RDF links to catalogs (fdp-o:metadataCatalog), datasets (dcat:dataset), distributions (dcat:distribution), and data services (dcat:accessService). At each step, the extracted triples are added to an RDF graph, resulting in a complete representation of all metadata layers for every FDP in the VP Index.

3. Results

Figure 3 summarizes metadata completeness across the FDPs indexed in the VP Index, distinguishing between active and inactive endpoints. Out of 25 FDPs identified, 16 are active and 9 are inactive at the time of analysis. Out of the 16 active FDPs, 10 FDPs expose some catalog metadata, while another 6 FDPs contain no catalog metadata and therefore cannot be used for federated discovery. Among the 10 active FDPs with catalog metadata, 7 provide datasets, of which 4 FDPs have distribution, and only 2 of them expose data service metadata.

A similar pattern is observed among 9 inactive FDPs, where metadata drop-off progresses at each successive metadata level, leaving no inactive FDPs populated from the FDP to the data service level.

The diagram highlights where FDPs drop out of the FDP metadata hierarchy and shows that 23 out of 25 FDPs in the VP Index are not yet fully populated to support content discovery. Most importantly, metadata at the data service level is required for actionable access to data or Beacon API, otherwise searches by ontology terms in the VP Portal for resource discovery are incomplete.

4. Discussion

To enable queries for content discovery, we recommend ensuring that an FDP has populated (meta)data on all levels of the hierarchy, from top-level FDP to data service. The VP discovers Beacon services only through data service metadata; therefore, only FDPs populated with metadata up to the data service level expose actionable access endpoints, while the majority of FDPs (23 out of 25) remain invisible (Figure 3), not only for content discovery queries, but even for metadata-based resource discovery queries. The Beacon /catalogs endpoint also needs to access the Beacon URL, which is located on the level of data service. Therefore, responses to VP Portal searches for content discovery queries do not reflect the full information about rare disease resources connected to the VP.

Additionally, the previously mentioned issues regarding the reluctance of resources to expose real patient data via APIs also pose a significant impediment to the adoption of content discovery queries, as metadata is not enough for retrieving aggregated answers to queries.

The R script presented in this work provides data stewards with an automated method for checking

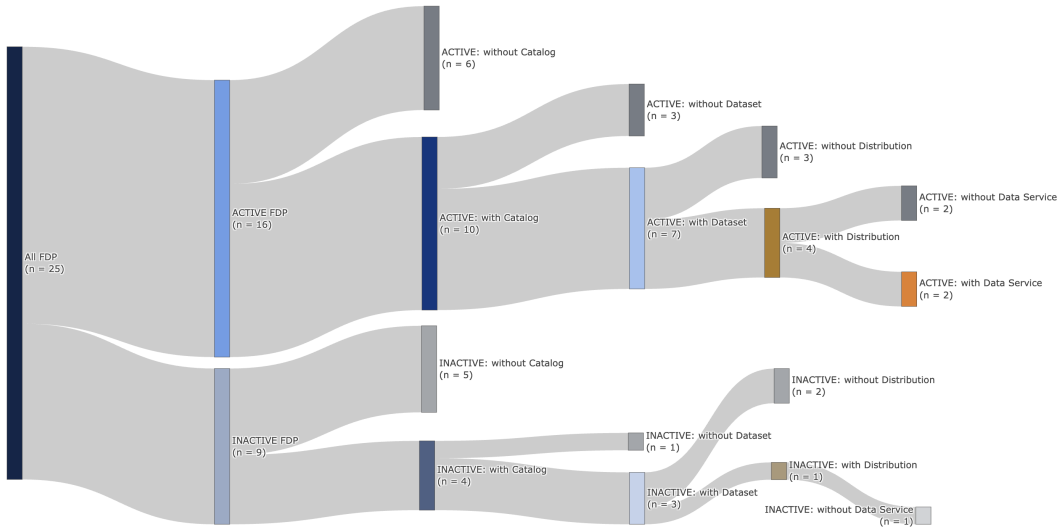

Figure 3: Metadata completeness across the VP Index. Sankey diagram showing how 25 FDPs (16 active and 9 inactive) progress through the FAIR metadata hierarchy (Catalog → Dataset → Distribution → Data Service). Only 2 active FDPs expose data service metadata, indicating that most nodes are not yet ready for federated content discovery.

where the VP Index metadata information is missing or incomplete in practice and which FDPs will not be contributing to the results of the search on the VP Portal. When paired with a FAIR evaluator like [22], this can provide significant information to pinpoint which FDPs are correctly populated with metadata and assess the issues of those that do not properly display on the VP.

Since the current FDP specification [23] defines the metadata architecture but does not require each level to be populated, many resources remain structurally valid yet functionally invisible. One solution is to take advantage of the already existing connection between Catalog and Data service (dcat:service) to expose Beacon access at the catalog level, enabling the VP to return richer results for content discovery based on ontology terms even when distributions are missing. Although this will not enable content discovery queries, it would increase the visibility of rare disease resources that cannot yet or will not publish distributions. However, this approach would be beneficial only if accompanied by catalog and datasets annotation efforts. At the same time, the FDP specification does not have metadata completeness within its requirements, meaning that metadata completeness validation must instead be performed by data stewards.

Beyond diagnosing individual FDPs, the script also reproduces the full RDF graph of all the resources registered in the VP Index. This graph can be further queried in SPARQL by data stewards for more in-depth analysis of metadata quality. Furthermore, the script can be modified to perform the same analysis for other FDP indexes.

5. Future work

The metadata completeness check FDPcrawleR script highlights where FDPs metadata is incomplete but cannot determine whether the absence is intentional. The absence of metadata is not always caused by a problem during the onboarding process (e.g. improper DCAT properties) or FDP creation, it can sometimes be intentional. Examples include early stage-deployments that register only a single catalog, sensitive datasets where distributions are intentionally not shared, or placeholder FDPs used for test purposes. Data stewards therefore need complementary processes to interpret metadata gaps and ensure real onboarding problems get fixed first. Since the FDPcrawleR script also reconstructs the full RDF graph of all the metadata contained in the VP Index, advanced SPARQL queries can support fine-grained analysis of metadata patterns. Future extensions may also incorporate retrieval and inspection of

Beacon responses, enabling content analysis if more FDPs populate their distribution and data service metadata.

As a follow-up to this work, we plan to enhance the capabilities of this tool by creating a real-time online FDP monitoring system composed of FDPCrawleR and a metadata validator. This system would allow data stewards to check not only the presence of metadata at different FDP levels, but also to ensure that FDPs comply with the VP requirements. Therefore, the system's uses would be two-fold: keeping track of already registered FDPs by providing notifications when updates cause validation issues and testing new FDPs before they become registered in the VP.

6. Conclusion

This work examined the metadata architecture of the FDPs indexed in the VP Index and assessed their level of metadata population using the FDPcrawlerR script. The visualization of metadata completeness shows that the contents of 23 FDPs in the VP Index (out of 25) do not expose metadata in a way that makes their resources accessible to API-based content discovery on the VP Portal. With the focus not only on the existence of metadata but also on its actual accessibility for federated discovery, this analysis provides insight into how FDP metadata is currently exposed within the VP network. This result indicates the governance changes between EJP-RD and ERDERA and the evolution of standards between the two projects. It also highlights the importance of tools such as FDPCrawleR, as they support ongoing data stewards' efforts to identify and address issues that pose a barrier to proper and efficient use of FDPs in the VP.

The FDPcrawleR script for FDP metadata completeness checks provides essential support to data stewards by automating the verification of metadata for both newly onboarded and already connected resources. Improving VP metadata completeness directly increases the discoverability and reusability of rare disease resources, enabling more effective federated discovery across the European rare disease ecosystem. Automated metadata completeness checks across the FDP network facilitate a transition from simple metadata aggregation to safe, aggregated federated content discovery on the VP. Verification of the quality of FDPs metadata promotes data findability and accessibility across ERDERA's partners and contributes to advancing rare disease research.

References

[1] S. Nguengang Wakap, D. M. Lambert, A. Olry, C. Rodwell, C. Gueydan, V. Lanneau, D. Murphy, Y. Le Cam, A. Rath, Estimating cumulative point prevalence of rare diseases: analysis of the Orphanet database, European Journal of Human Genetics 28 (2020) 165–173. doi:10.1038/s41431-019-0508-0.

[2] European Parliament, EUR-Lex - 31999D1295 - EN, 1999. URL: https://eur-lex.europa.eu/eli/dec/1999/1295/oj/eng.

[3] European Rare Diseases Research Alliance (ERDERA), Homepage, 2025. URL: https://erdera.org/.

[4] N. Denton, M. Molloy, S. Charleston, C. Lipset, J. Hirsch, A. E. Mulberg, P. Howard, E. D. Marsh, Data silos are undermining drug development and failing rare disease patients, Orphanet Journal of Rare Diseases 16 (2021) 161. doi:10.1186/s13023-021-01806-4.

[5] Z. Gu, F. Corcoglioniti, D. Lanti, A. Mosca, G. Xiao, J. Xiong, D. Calvanese, A systematic overview of data federation systems, Semantic Web 15 (2024) 107–165. doi:10.3233/SW-223201.

[6] H. Ulrich, A.-K. Kock-Schoppenhauer, N. Deppenwiese, R. Gött, J. Kern, M. Lablans, R. W. Majeed, M. R. Stöhr, J. Stausberg, J. Varghese, M. Dugas, J. Ingenerf, Understanding the Nature of Metadata: Systematic Review, Journal of Medical Internet Research 24 (2022) e25440. doi:10.2196/25440.

[7] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. Da Silva Santos, P. E. Bourne, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. T. Evelo, R. Finkers, A. Gonzalez-Beltran, A. J. Gray, P. Groth, C. Goble, J. S. Grethe, J. Heringa, P. A. 'T Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S. J. Lusher,

M. E. Martone, A. Mons, A. L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. Van Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M. A. Swertz, M. Thompson, J. Van Der Lei, E. Van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, B. Mons, The FAIR Guiding Principles for scientific data management and stewardship, Scientific Data 3 (2016) 160018. doi:10.1038/sdata.2016.18.

[8] L. O. B. Da Silva Santos, K. Burger, R. Kaliyaperumal, M. D. Wilkinson, FAIR Data Point: A FAIR-Oriented Approach for Metadata Publication, Data Intelligence 5 (2023) 163–183. doi:10.1162/dint_a_00160.

[9] World Wide Web Consortium (W3C), RDF 1.1 Concepts and Abstract Syntax, 2014. URL: https://www.w3.org/TR/rdf11-concepts/.

[10] World Wide Web Consortium (W3C), Data Catalog Vocabulary (DCAT) - Version 2, 2020. URL: https://www.w3.org/TR/vocab-dcat-2/.

[11] World Wide Web Consortium (W3C), Linked Data Platform 1.0, 2015. URL: https://www.w3.org/TR/ldp/.

[12] European Joint Programme on Rare Diseases (EJP RD), EJP RD Virtual Platform: Resources onboarding manual — EJP RD Onboarding Document documentation, 2024. URL: https://vp-onboarding-doc.readthedocs.io/en/latest/index.html.

[13] ERDERA, ERDERA Virtual Platform Portal, 2025. URL: https://vp.erdera.org/.

[14] A. Rath, A. Olry, F. Dhombres, M. M. Brandt, B. Urbero, S. Ayme, Representation of rare diseases in health information systems: The orphanet approach to serve a wide range of end users, Human Mutation 33 (2012) 803–808. doi:10.1002/humu.22078.

[15] European Joint Programme on Rare Diseases (EJP RD), EJP RD VP Index, 2025. URL: https://index.vp.ejprarediseases.org/.

[16] European Joint Programme on Rare Diseases (EJP RD), ejp-rd-vp/resource-metadata-schema, 2025. URL: https://github.com/ejp-rd-vp/resource-metadata-schema, original-date: 2019-05-01T08:56:28Z.

[17] Global Alliance for Genomics and Health (GA4GH), ga4gh-beacon/beacon-v2, 2025. URL: https://github.com/ga4gh-beacon/beacon-v2, original-date: 2022-02-22T08:53:58Z.

[18] D. Hayn, E. Sandner, A. Vengadeswaran, E.-A. Tătaru, M. Wilkinson, M. Hanauer, K. Kreiner, G. Schreier, Privacy-Preserving Linkage of Distributed Pseudonymised Datasets in a Virtual European Rare Disease Platform, in: J. Mantas, A. Hasman, G. Demiris, K. Saranto, M. Marschollek, T. N. Arvanitis, I. Ognjanović, A. Benis, P. Gallos, E. Zoulias, E. Andrikopoulou (Eds.), Studies in Health Technology and Informatics, IOS Press, 2024. doi:10.3233/SHTI240683.

[19] European Joint Programme on Rare Diseases (EJP RD), ejp-rd-vp/vp-api-specs, 2024. URL: https://github.com/ejp-rd-vp/vp-api-specs, original-date: 2021-08-11T12:03:27Z.

[20] World Wide Web Consortium (W3C), SPARQL 1.1 Query Language, 2013. URL: https://www.w3.org/TR/sparql11-query/.

[21] K. Vodorezova, vodor001/FDPcrawleR, 2025. URL: https://github.com/vodor001/FDPcrawleR, original-date: 2025-12-05T01:55:39Z.

[22] M. D. Wilkinson, The FAIR Maturity Evaluation Service, 2024. URL: https://fairsharing.github.io/FAIR-Evaluator-FrontEnd/.

[23] L. O. B. Da Silva Santos, K. Burger, R. Kaliyaperumal, FAIR Data Point Specifications, 2023. URL: https://specs.fairdatapoint.org/fdp-specs-v1.2.html.