Knowledge multiplies when shared — when calling things by their right name: improving the validation and exchange of genetic data in research and diagnostics

Fokkema, I.F.A.C.

CHAPTER 10

General discussion

Ivo F.A.C. Fokkema

10.1 Abstract

DNA variant databases have transformed research and clinical diagnostics by enabling broad data sharing. However, maintaining these databases remains challenging due to funding limitations and reliance on short-term projects. This chapter examines the evolution, challenges, and sustainability of DNA variant databases, with a particular focus on the Leiden Open Variation Database (LOVD).

The added value of DNA variant databases largely depends on the data quality. Yet, the validation and standardization of genetic variant nomenclature remains challenging. The Human Genome Variation Society (HGVS) Nomenclature remains the standard, yet errors persist in scientific literature and clinical reports, hindering findability and accuracy in diagnostics. Collaboration with publishers and automated validation tools could improve standardization.

Errors in variant descriptions in the scientific literature also hinder the uptake of published data into variant databases. Furthermore, although patients are generally willing to share their genetic data, laboratories often refrain due to time constraints, funding issues, and concerns about intellectual property. Efforts such as automated manuscript validation and structured data submission aim to increase contributions to public databases.

Finally, this discussion explores future approaches to data sharing, transitioning from centralization to federated systems using FAIR (Findable, Accessible, Interoperable, Reusable) principles. The LOVD team has taken initial steps toward FAIR compliance, though technical barriers remain. Future work will focus on improving automation and interoperability to ensure sustainable, secure, and efficient variant data sharing.

10.2 DNA variant databases: evolution, challenges, and sustainability

DNA variant databases have revolutionized both research and clinical diagnostics by providing direct access to variant data compiled by thousands of researchers and clinicians around the world (see **Chapters 3 and 4**). Three decades ago, in addition to a select number of centralized databases hosted at large institutions that provided mostly general variant information, numerous smaller, researcher-driven databases emerged, offering richly detailed patient and disease-specific information. The release of the Leiden Open Variation Database (LOVD) software greatly accelerated the development of this data-sharing landscape by greatly simplifying and standardizing the setting up of databases (see **Chapter 4**). A mere decade after its initial release, nearly 150 LOVD instances were active worldwide, spanning diverse genes, phenotypes, and diseases.

However, maintaining these database systems proved challenging, and over the years, the number of LOVD instances started to decrease. Some databases were deliberately merged into larger, actively maintained initiatives — such as older LOVD2 databases transitioning into the LOVD3 shared instance at the Leiden University Medical Center (LUMC) — improving reliability and reducing fragmentation. Regrettably, other valuable databases have become permanently lost due to technical or organizational failures. A notable example is the Brazilian Initiative on Precision Medicine (BIPMed),[1] where multiple large LOVD3 databases, each containing close to one million unique variants, were lost when critical personnel departed without transition plans, resulting in irrecoverable loss of platform access and data. As of March 2025, the number of publicly accessible LOVD instances has dwindled to 19, many of which remain poorly maintained or run outdated software.

A primary cause of this sustainability challenge is insufficient long-term funding and dedicated IT expertise.[2] Database projects often begin within limited-term grants but rarely include measures for post-project support. Likewise, funding providers rarely demand to collaborate with existing databases but continue funding the creation of new databases. Furthermore, skilled technical staff at hosting institutions are frequently unavailable for prolonged support, leaving databases without necessary maintenance or updating. The sustainability dilemma directly ties into funding models: free public access greatly increases short-term utility but poses financial and sustainability concerns, whereas paid or restricted access provides funding for maintenance and development but restricts usage to financially well-funded laboratories in developed countries. The shared LOVD database at the LUMC has sought a compromise by publicly displaying curated data but allowing submitters to select licensing conditions, empowering them to choose whether their data can be commercially licensed in aggregate data downloads sold to commercial diagnostic laboratories, funding the LOVD project. At the same time, the LOVD team has been forced into a difficult battle against unauthorized

automated data harvesting and web scraping, behavior increasingly exhibited by known and unknown users.

Over the recent years, automated access to the databases has dramatically increased. Five years ago, the LOVD worldwide variant search API, which allows queries across all LOVD databases from one central endpoint, recorded around 125 million hits annually. Traffic surged to a staggering 881 million hits in 2024, driven by increased incorporation of our API into automated diagnostic pipelines used in diagnostic laboratories and large-scale genomic analysis workflows. While most API traffic is non-disruptive, rising numbers of commercial entities attempting to copy the complete LOVD database contents have presented substantial operational challenges. Initially limited primarily to web search providers such as Google, Apple, and Microsoft, web-scraping activity now includes many AI-focused companies such as Amazon, Bytedance, OpenAI, and Facebook. On numerous occasions, massive simultaneous web requests coming from countless unique IP addresses produced server overload, severely impacting LOVD availability or even disabling the database completely, functioning in effect like a distributed denial-of-service (DDoS) attack. The LOVD team eventually implemented stringent anti-scraping defenses, which is not always easy given the sheer number of IP addresses used. In particular, Facebook even attempted to hide its identity to bypass our restrictions when it was blocked completely for a while due to abuse of the LOVD servers.

Maintaining the shared LOVD at the LUMC involves substantial costs, including server upkeep, web scraping protection, traffic management, and user support, amounting to approximately €25,000 per year. Additionally, the minimum development effort required to fix bugs and implement minor feature updates costs an estimated €30,000 annually. However, keeping the software up to date with emerging developments, user demands, addressing increasing traffic demands, and significantly enhancing submission and curation features would require at least three times that amount. This brings the total annual cost of running LOVD to approximately €115,000. Notably, this figure excludes data curation costs, currently managed by scarce volunteers. If funding would be sought for a dedicated data curator, an additional €40,000 – €80,000 will be required.

Despite its widespread use (tens of thousands of unique visitors per month, tens of millions of monthly page views and API requests, and more than 1500 citations to the LOVD papers), securing sustainable funding for LOVD has remained elusive for two decades. A 2021 donation campaign targeting database users yielded less than €200, underscoring the reluctance of clinical labs and research groups to invest in a resource they, and through them, patients, rely on daily. This reluctance has forced us to take a semi-closed approach, restricting free access to large-scale data downloads. While the lack of full data downloads is often criticized, it is the only way to ensure the survival of the world's largest collection of gene

variant databases. Currently, commercial diagnostic laboratories willing to pay for data access provide less than half of the funds required to maintain the databases. Until LOVD's primary users start contributing to its sustainability, our efforts must focus on improving commercial data exports and securing support from diagnostic labs willing to invest in LOVD's continued existence and improvement.

10.3 Variant naming, validation, and standardization

Even with extensive resources in place, the value of these largely depends on the quality of the data. The most important factor impacting the data quality is the unequivocal description of a variant, making it impossible to confuse it with another variant through the use of a common standard for describing variants. The Human Genome Variation Society (HGVS) Nomenclature, initially developed nearly 25 years ago, is the de facto standard for the naming of genetic variants on DNA, RNA, and protein levels in databases and clinical reports (see **Chapter 5**). However, studies consistently show that errors and inconsistencies in variant naming persist in most publicly available genetic studies and clinical reports, negatively impacting data findability, diagnostic throughput, and accuracy (**Chapters 6 and 7**).

Even with multiple tools available for validating variant descriptions, there may be several reasons why these remain unused:

- Authors continue to use legacy names for variants that are unclear to others outside of their specific field of research (e.g., cancer research or pharmacogenetics) and do not realize that increasingly automated diagnostics is not able to interpret their findings;
- Authors may not be aware that the validation software exists or do not realize their descriptions may be invalid;
- Authors focus on getting their manuscript published and have no incentive to get variant descriptions reported correctly and unambiguously;
- Variant validation tools can be quite strict in the input they allow and require the input to be mostly already correctly formatted, which may lead to authors having trouble validating their variant descriptions.

These problems highlight the importance of two complementary approaches for improvement. Firstly, persuading major scientific publishers to require standardized variant validation before manuscript acceptance is crucial (see Chapter 6). Secondly, by improving validation tools and automation, the time needed for the validation can be greatly reduced (see Chapter 7). Collaborative discussions involving our group, Elsevier, Nature Publishing Group, and the developers of the widely used VariantValidator software aim to address this by creating automated solutions that simultaneously extract and validate variant descriptions directly from submitted manuscripts. We believe this will greatly increase the quality of variant descriptions

in published papers. Ideally, once manuscripts are accepted for publication, variant data could be automatically deposited in gene variant databases and linked with the article's DOI, maximizing both immediate visibility and effective incorporation into diagnostic pipelines. Moreover, this collaboration has the potential to generate significant and sustainable funding for the LOVD project, making it one of our top priorities this year.

Even with improved validation software able to pick up common user errors and recognize rare types of variant descriptions, there is room for improvement within the HGVS Nomenclature itself. Chapter 7 highlights some of the areas within the HGVS Nomenclature lacking rules on implementation or prioritization, making it currently difficult or impossible to apply the HGVS Nomenclature rules for certain rare or complex variant descriptions. The interchangeability of variant types like deletion-insertions that can be described as a range of smaller changes on the same allele, variability in DNA repeat sequences that can be described as deletions or duplications, or large insertions that can be described in a range of different ways, highlight that currently, the HGVS Nomenclature does not fully protect against having several valid descriptions for the same sequence change. On the protein level, the HGVS Nomenclature mentions valid alternatives for describing frameshift variants, deliberately creating alternative notations.[3] As a direct consequence, only specialized software is currently able to fully and reliably compare variant descriptions to assess if they are equal. That is why the Global Alliance for Genomics and Health (GA4GH) developed the Variation Representation Specification (VRS) standard,[4] which assigns unique identifiers to fully normalized variant representations. While VRS cannot yet mirror HGVS's descriptive complexity, incorporating exhaustive normalization principles into a revised HGVS standard would solve the ambiguity problem. If this problem is not addressed, the only alternative that allows reliable matching is implementing specialized software in databases that normalize both the queried variant description and the database contents following the same algorithm. As this approach would make database queries significantly slower, implementing normalization rules as part of the HGVS Nomenclature is preferable.

10.4 Data sharing: achievements and remaining hurdles

The central aim of this thesis was to address key issues surrounding the collection, validation, management, sharing, and accessibility of genetic variant data, with a particular focus on automated solutions. This work has significantly advanced data-sharing practices by successfully implementing approaches to automate the collection, validation, and curation of genetic data from the different genome diagnostics laboratories across the Netherlands, and we described the process in detail, hoping to inspire others to follow this path (see **Chapter 8**). As a result of the mechanisms we established, to this day, numerous patients worldwide are obtaining their genetic diagnoses. Additionally, we enhanced and modernized

a unique and versatile data format that we originally developed in 2012 (see **Chapter 9**). It is able to effectively store complex genetic datasets, and our updates substantially improve the data format's interoperability with current bioinformatics tools and infrastructure.

Despite these technological advances and improvements to data-sharing infrastructures, a major bottleneck persists: only a very limited fraction of the genetic data generated globally is actually shared with the wider scientific and clinical community.[5] The scale of this issue is demonstrated when we examine the statistics of the variant submissions through LOVD. In 2024 alone, LOVD's worldwide query API processed over 881 million requests from users, representing an approximate minimum of the number of relevant genetic variants assessed in diagnostic laboratories that year. However, during the same year, only 65,205 variants were directly submitted to the LOVD shared installation. This implies that only a very small portion (using these numbers, 0.007%) of generated diagnostic variant data is actively being contributed to LOVD. This tremendous underreporting fundamentally limits our collective knowledge about genetic variations and significantly reduces the potential diagnostic utility of the generated data.

Importantly, the willingness of patients to contribute genetic data contrasts sharply with the reluctance or outright refusal observed among laboratories and research groups. When patients and families are given the opportunity to contribute their genetic data to help themselves and others, they overwhelmingly agree as long as their permission is sought.[6] The LOVD team has even received several direct submissions from patients. Yet, while we do wish to acknowledge those who responded positively and shared their data through the LOVD databases or otherwise, despite numerous requests and appeals to diagnostic laboratories, research consortia, and patient organizations, the overwhelming majority of responses have been negative or cautious at best.

Several responses were encountered repeatedly. Predominantly, a lack of time and funding was cited as the main reason for not sharing data. For diagnostic labs, a simple solution could be to make the relatively small costs of public data sharing an integral part of the costs of a clinical diagnosis. In research contexts, data is rarely collected in a format that can be submitted to databases directly, therefore requiring reformatting to a format suitable for database submission, an activity rarely included in the research funding. Furthermore, there are insufficient incentives for database submission. Specifically, while publishing findings in scientific papers generally increases the visibility and prestige of researchers and helps secure future funding, contributing directly to databases does not provide comparable rewards. An additional barrier highlighted by potential contributors involves concerns about data provenance and protecting intellectual property; datasets shared in open databases may be (illegally) copied and reshared without proper attribution to the original submitters.

Although LOVD includes licenses that prohibit unattributed re-use, the persistent attempts by unauthorized data scrapers to illegally duplicate LOVD's data emphasize a legitimate concern about control over datasets post-publication.

As a direct result of the above objections, if data is shared at all, it is often merely published in scientific literature rather than in databases. However, literature-based genetic information is inherently poorly suitable for broad-scale data sharing and downstream reuse. Searching the literature and extracting data from papers is labor-intensive and commonly yields alternative or incomplete variant descriptions. From 2019 to 2023, we undertook a project funded by the Foundation Fighting Blindness (FFB) in the United States, systematically extracting variant data from close to 5,000 scientific publications for incorporation into LOVD. Our experience revealed that most publications contained errors, inaccuracies, or omissions that required significant manual curation and validation. Moreover, 1.8% of the analyzed papers yielded no usable data at all, emphasizing serious flaws even in published datasets. Considering the FFB project alone consumed more than half a million US dollars in funding, data extraction and curation from published literature clearly remain expensive and inefficient processes. To resolve these inefficiencies, we are currently extending our collaboration with VariantValidator by developing enhanced manuscript validation pipelines that automatically structure variant data at manuscript acceptance, potentially even allowing immediate electronic submission to databases upon publication.

10.5 Fragmentation, then centralization, to federation: approaches for future data-sharing solutions

Initially, variant data was fragmented across a myriad of resources such as gene-specific databases, patient organizations, the scientific literature, and diverse online repositories with varying data formats. Gradually, however, the landscape evolved toward data centralization, driven largely by the mainstream adoption of LOVD by alternative variant database, the merging of older LOVD2 instances into a single LOVD3 instance, and the creation of centralized databases such as ClinVar, maintained by the U.S. National Center for Biotechnology Information (NCBI) (see **Chapter 3**). Centralized platforms greatly improved user-friendliness, as researchers and clinicians need to search fewer repositories. Nevertheless, centralization introduces its own limitations. Increased data volume is rapidly outpacing efforts by single institutions or organizations to store, manage, and curate all genetic data. Moreover, tightly centralized systems necessarily compromise the rich feature sets originally provided by independent, specialized databases, thus sacrificing essential specificity and functionality.

Over the past few years, the push for centralization has gradually shifted toward a federated approach.[7] In 2010, we introduced the aforementioned worldwide query API that enabled

diagnostic pipelines to search all public LOVD databases within the LOVD network through a single endpoint. This was later complemented by an interface on LOVD.nl, allowing human users to perform the same searches. However, because this API relies on data cached on the central LOVD server, it does not constitute true federation, which requires data to remain at its original location. In 2014, the GA4GH launched the Beacon network, a system designed to facilitate federated queries.[8] Initially, it supported only a basic query, determining whether a given variant was present in a data source. Beacon automatically searches across approximately 80 different resources for matches. Unlike centralized caching, this federated approach ensures that data remains distributed, though it also results in longer response times as each resource processes the query independently. When a data source indicates that it contains information on a variant, users must then visit that resource to access the details. While the later release of Beacon v2 enabled more direct data sharing, fully realizing complex federated queries remains an ongoing challenge.

One of the most promising initiatives for enabling complex federated queries is the adoption of the FAIR Principles.[9] FAIR, an acronym for Findable, Accessible, Interoperable, and Reusable, aims to maximize data reusability by ensuring that computers can not only discover and access data but also interpret and interact with it. Additionally, FAIR principles require clear usage rights through the application of data licenses, specifying what a system can and cannot do with the data. An interesting concept inspired by FAIR is the FAIR Data Train, an approach where data analysis requests and responses are transported by a virtual "train" that moves between different "stations" (data and computation providers). At each station, the train interacts with the data provider, retrieving only the necessary information before continuing its journey. Computations are carried out on location or in secure computing environments. The original data remains with the provider, ensuring highly secure federated queries that comply with strict privacy regulations. The best-known implementation of this concept is the Personal Health Train (PHT),[10] though real-world deployments remain limited.

LOVD has taken its first steps toward implementing the FAIR principles, a process known as "FAIRification", by developing a FAIR Data Point (FDP, `fdp.lovd.nl`). This FDP stores metadata about LOVD and directs users to the actual data, laying the groundwork for FAIR compliance. Additionally, a preliminary FAIR data model for LOVD's contents has been developed (`github.com/LOVDnl/EJP-RD-LOVD-model`). However, fully integrating FAIR principles remains a challenging and slow process. As a relatively new framework, FAIR lacks sufficient comprehensive documentation, and its implementation often requires expertise in less common web standards. The implementation of machine actionability requires linked data, most commonly expressed in an RDF (Resource Description Framework) format. In most fields, including health sciences, data exchange typically relies on JavaScript Object Notation (JSON) as a format and JSON Schema for data model validation.[11] JSON-LD (JSON for Linked

Data) is a serialization of RDF in JSON format and is, therefore, a logical choice for an RDF format with an immediate broad support in the web developer community. However, developing the FDP and its FAIR data model required proficiency in RDF exchanged in the form of Terse RDF Triple Language (TTL, pronounced "Turtle") and Shape Expressions (ShEx), technologies that are less widely supported and can, therefore, be difficult to implement. While JSON and JSON Schema are natively supported in nearly all popular programming languages, TTL and ShEx libraries are harder to find and sometimes unreliable. As a result, few developers can implement FAIR principles without first learning these alternative technologies. We believe this technical barrier hinders a broader adoption of FAIR, and we recommend that groups working on FAIR standards focus their efforts on incorporating more broadly accepted worldwide technical standards into the FAIR domain.

Despite these challenges, we remain committed to continue FAIRifying LOVD in the future. By integrating FAIR principles directly into the software, we aim to streamline the FAIRification process for datasets containing genetic variants and patient data. Eventually, simply importing data into an LOVD instance will automatically ensure its FAIR compliance. For smaller, distributed datasets, such as those collected by patient organizations, using LOVD is a far more effective approach to FAIRification than manually processing each dataset individually, as well as a more constructive way to spend funding. LOVD's flexible design allows the software to be tailored to the specific needs of each dataset, while its well-established infrastructure enables seamless integration into diagnostic pipelines. At the same time, data providers retain full control by hosting their own data. This federated approach could be the solution that many data providers, concerned about losing control over their data, have been waiting for.

In the future, gene variant databases will continue to play a vital role in both research and clinical diagnostics. As federated queries become more widespread, the detailed case-level information stored in databases like LOVD will become increasingly accessible, offering valuable insights into the most complex genetic disorders. The work presented in this thesis has contributed to advancing the fields of gene variant databases, data validation, and data sharing, and supports genetic research and diagnostics on a global scale.

10.6 References

[1] Cristiane S. Rocha, Rodrigo Secolin, Maíra R. Rodrigues, Benilton S. Carvalho, and Iscia Lopes-Cendes. *The Brazilian Initiative on Precision Medicine (BIPMed): fostering genomic data-sharing of underrepresented populations*. npj Genomic Medicine 2020; 5 (1) 42.

[2] Teresa K. Attwood, Bora Agit, and Lynda B.M. Ellis. *Longevity of Biological Databases*. EMBnet Journal 2015; 21 (0) 803.

[3] HGVS Variant Nomenclature Committee. *HGVS Nomenclature: Protein Frameshift*. URL: `https://hgvs-nomenclature.org/recommendations/protein/frameshift/` (visited on April 15, 2025).

[4] Alex H. Wagner, Lawrence Babb, Gil Alterovitz, Michael Baudis, Matthew Brush, Daniel L. Cameron, Melissa Cline, Malachi Griffith, Obi L. Griffith, Sarah E. Hunt, David Kreda, Jennifer M. Lee, Stephanie Li, Javier Lopez, Eric Moyer, Tristan Nelson, Ronak Y. Patel, Kevin Riehle, Peter N. Robinson, Shawn Rynearson, Helen Schuilenburg, Kirill Tsukanov, Brian Walsh, Melissa Konopko, Heidi L. Rehm, et al. *The GA4GH Variation Representation Specification: A computational framework for variation representation and federated identification*. Cell Genomics 2021; 1 (2).

[5] Nicola Milia, Alessandra Congiu, Paolo Anagnostou, Francesco Montinaro, Marco Capocasa, Emanuele Sanna, and Giovanni Destro Bisol. *Mine, yours, ours? Sharing data on human genetic variation*. PloS One 2012; 7 (6).

[6] Miranda E. Vidgen, Sid Kaladharan, Eva Malacova, Cameron Hurst, and Nicola Waddell. *Sharing genomic data from clinical testing with researchers: public survey of expectations of clinical genomic data management in Queensland, Australia*. BMC Medical Ethics 2020; 21 (1).

[7] Maria A. Rujano, Jan Willem Boiten, Christian Ohmann, Steve Canham, Sergio Contrino, Romain David, Jonathan Ewbank, Claudia Filippone, Claire Connellan, Ilse Custers, Rick van Nuland, Michaela Th. Mayrhofer, Petr Holub, Eva García Álvarez, Emmanuel Bacry, Nigel Hughes, Mallory A. Freeberg, Birgit Schaffhauser, Harald Wagener, Alex Sánchez-Pla, Guido Bertolini, and Maria Panagiotopoulou. *Sharing sensitive data in life sciences: an overview of centralized and federated approaches*. Briefings in Bioinformatics 2024; 25 (4).

[8] Global Alliance for Genomics and Health. *A federated ecosystem for sharing genomic, clinical data*. Science 2016; 352 (6291) 1278–1280.

[9] Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J.G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, et al. *The FAIR Guiding Principles for scientific data management and stewardship*. Scientific Data 2016; 3.

[10] Oya Beyan, Ananya Choudhury, Johan van Soest, Oliver Kohlbacher, Lukas Zimmermann, Holger Stenzhorn, Rezaul Karim, Michel Dumontier, Stefan Decker, Luiz Olavo Bonino Da Silva Santos, and Andre Dekker. *Distributed Analytics on Sensitive Medical Data: The Personal Health Train*. Data Intelligence 2020; 2 (1-2) 96–107.

[11] Jiwen Xin, Cyrus Afrasiabi, Sebastien Lelong, Julee Adesara, Ginger Tsueng, Andrew I. Su, and Chunlei Wu. *Cross-linking BioThings APIs through JSON-LD to facilitate knowledge exploration*. BMC Bioinformatics 2018; 19 (1).