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.

APPENDIX &

Summary
Samenvatting
Curriculum Vitae
List of publications

Summary

Genetic disorders and genetic predisposition to disease pose a significant burden on health-care systems worldwide. Yet, more than twenty years after the completion of the Human Genome Project, only a fraction of the genomic data generated globally is actively shared. When data are shared, this is often done using inefficient methods or incorrect variant descriptions, which in the worst cases render the data unusable. The work described in this thesis focuses on improving data sharing in genetics, in particular through gene variant databases, and by ensuring that data are represented using standards that enable unique and unambiguous variant descriptions.

The key subjects underlying this thesis are introduced in the first three chapters. **Chapter 1** outlines the scope of the thesis and provides an overview of all chapters, explaining their connections. **Chapter 2** introduces genetic variation, the potential consequences of variants, and the standards essential for describing, interpreting, and reporting genetic variants and phenotypes. **Chapter 3** discusses general and "focused" databases, outlining their primary aims and benefits.

Gene variant databases form the backbone of DNA-based diagnostics. They are the most reliable source of information on variants and their consequences, as they typically adhere to strict data standards and guidelines. These databases store detailed case-level data and usually focus on a specific gene or disease, curated by experts in the field. To enable researchers and clinical laboratories to build their own gene variant databases, we developed the free, open-source Leiden Open Variation Database (LOVD) software, which quickly became the most widely used tool for this purpose. After two successful major releases, we completely redesigned the software and released LOVD3 (described in **Chapter 4**). LOVD3 enabled larger and more diverse groups of curators to collaborate within a single instance. At the same time, advances in sequencing shifted the field from gene-based to genome-based variant descriptions, requiring a thorough redesign of the data model as well as all views and forms. LOVD3 is genome-centered and supports both summary variant data and detailed case-level data, including information on individuals, phenotypes, screenings, and variants. This flexibility, combined with Application Programming Interfaces (APIs) that enable direct interaction with the database, has created the world's largest gene variant database network. It supports queries across different LOVD instances, allowing users to quickly identify where relevant information on a variant is stored.

Combining information from different systems for diagnosing a patient in clinical diagnostics is only possible when these systems use the same standards to describe DNA variants and phenotypic data. Over the past two decades, the Human Genome Variation Society (HGVS)

Nomenclature has become the global standard for describing variants in DNA, RNA, and protein sequences. Yet even now, most variants reported in the literature are incorrectly or incompletely described. While this strengthens the role of databases as the most reliable source of variant information, it also hampers the integration of literature data into databases and ultimately slows diagnostics. In recent years, the HGVS Variant Nomenclature Committee (HVNC) has released several updates (described in **Chapter 5**) to improve the correct application of the nomenclature in the literature and clinical reports. The website was redesigned to improve the readability of the documentation, errors and ambiguities were removed, all changes were clearly documented, and a dedicated page was added listing software capable of validating and correcting variant descriptions. Community engagement was also strengthened by providing additional channels for users to ask questions and by facilitating methods for suggesting changes to the nomenclature.

Databases rely on specialized software to validate submitted variant descriptions. A direct way to improve descriptions in the literature is to integrate this validation software into the manuscript submission process. This requires collaboration between publishers and software developers, ensuring that manuscripts are thoroughly checked and corrected before publication. Moreover, such software could even automatically submit the variants described in a paper to relevant databases. The LOVD team collaborates with the developers of VariantValidator and has prepared their software for integration into publisher workflows (see **Chapter 6**).

However, to fully automate the recognition of variant descriptions in the literature, additional tools are required. Existing validation software assumes that descriptions are already largely compliant with the HGVS Nomenclature; however, many publications use legacy names or incomplete descriptions, which current tools cannot adequately recognize or correct. Furthermore, descriptions of rare variants or those that express a certain level of uncertainty are rarely supported. To address this, we developed the LOVD HGVS syntax checker (see **Chapter 7**), designed to recognize both valid and commonly encountered invalid variant descriptions. For invalid inputs, the software automatically generates suggested corrections and assigns a certainty score to each correction. By focusing on syntax, the LOVD HGVS syntax checker achieves unprecedented coverage of the HGVS Nomenclature. Since full sequence-level validation is still required, LOVD and VariantValidator were integrated, offering a comprehensive solution within a single platform. We are currently in active discussions with publishers to integrate these tools into their submission pipelines, which would revolutionize the quality of variant descriptions in the scientific literature.

While most research data are published in journals, the majority of genetic data originates in diagnostic laboratories. In the Netherlands alone, thousands of individuals are screened

annually for genetic disorders. Although diagnostic labs rarely share detailed patient-level data, Dutch labs routinely share variants they identify along with their classifications (pathogenic or not), providing valuable information. This data is processed through a centralized pipeline (described in **Chapter 8**), allowing laboratories to collaborate and share their findings. It is also processed and released through LOVD, making it available worldwide. Because the data are fully validated before publication, their quality is greatly improved, and both internal conflicts and discrepancies between laboratories are detected and reported automatically.

Sharing more than just variants and classifications, however, requires more advanced approaches. Databases such as LOVD can store complex case-level data, including phenotypes, laboratory results, multiple variants, and detailed information on the effect on one or more genes. Such datasets cannot be represented in simple tab-delimited text files but require structured formats. To support this, we updated and modernized the VarioML format, initially published in 2012, into a native JSON implementation (see **Chapter 9**). The new format is easier to integrate into modern systems, supports both patient- and variant-centered views, and has been implemented in LOVD3 to enable downloads and automated submissions. By employing an ontology-driven approach, VarioML also moves LOVD closer to compliance with the FAIR principles (Findable, Accessible, Interoperable, Reusable).

Despite their transformative role in research and diagnostics, DNA variant databases such as LOVD face persistent challenges (see **Chapter 10**). First, funding remains scarce, and several large databases have already been lost due to financial or staffing issues. Maintenance is further complicated by increased interest from large commercial entities, which can generate overwhelming traffic and occasionally bring down servers. Second, only a fraction of available data is submitted to databases, and more efforts are needed to unlock and share additional data. Third, adherence to standards remains limited in resources such as the literature and clinical reports, meaning vast amounts of non-standardized descriptions must be recognized and corrected, ideally in automated ways. In the near future, our efforts to address these issues will converge through collaborations with publishers to automate data validation at the pre-publication stage, followed by data submission after the manuscript has been accepted for publication. At the same time, we will continue to develop methods for automated interaction with databases, including federated queries across FAIR-compliant systems.

In conclusion, the work presented in this thesis has advanced the fields of gene variant databases, data validation, and data sharing, supporting genetic research and clinical diagnostics on a global scale.