PHD FELLOWSHIP STRATEGIC BASIC RESEARCH
PROJECT OUTLINE (MAX. 12 pages)

Rationale and positioning with regard to the state-of-the-art

Decentralized Landscape Data decentralization initiatives [1–3] are working to reduce the data siloing caused by data centralization on the Web. A leading decentralized storage strategy is the use of personal data vaults. The Solid protocol in particular offers user-moderated access controls, data linking in and across vaults using the Resource Description Framework [4], represented as triples with universal semantics, built on Linked Data principles [5], and information extraction via querying using the SPARQL query language [6].

Personal Genome Sequencing in Healthcare Around the same time that the World Wide Web was being established, DNA sequencing technologies were just starting to be applied to the human genome [7]. At the time of writing, there are now multiple domains of clinical practice where patient personal genome sequence (PGS) data are now used to inform medical decision making. Examples include in drug development [8], cancer diagnosis and treatment [9], and rare genetic disease identification and treatment [10]. How this integration is deployed varies by clinical domain, but improved outcomes have generally been observed [11]. Despite great promise presented by various use cases, barriers to widespread adoption remain [12].

One major barrier to scalability is presented by the costs of data generation and storage[c]. The average human genome is slightly over 3 billion base pairs in length and during a whole genome sequencing workflow, various sequence formats that offer different sets of information are produced [**bagger_whole_2024**]. Of these, Variant Call Format (VCF) files [13] serve as the state-of-the-art for most clinical genomic applications and are typically between 100-1000s MB (0.1-1 GB) within computer memory.

Another is the computational costs of regenerating results and sequences because of little to no data sharing potential in the current system.

The costs of producing and maintaining these data are additioanlly increased by the privacy protections needed for PGS data[GDPR]. With the enlarged threat of hacking, phishing, and login credential compromisation that seems to only be increasing [14], hospitals and institutions are forced to enact tighter regulation over data access within their institution, severely limit outside institution access, and increase their cyber security budget to handle security audits.

PGS data sharing in academic research. In the realm of academic research, the development of infrastructure that allows for sharing of genome data between institutions, creating federated centralized databases, has gained traction recently. Initiatives such as GA4GH Beacons [15] and others are building infrastructure for this between institution data sharing. Despite this step towards increased sharing and cost reduction, advancements in state-of-the-art infrastructure and standards are not directly translatable to clinical practice.

Decentralized PGS data storage A possible solution to the challenges faced is through reorganization of how data is stored and discovered. The citizen-centric model places the patient at the center, and is not an entirely novel concept [**brands_patient-centered_2022**]. Within the current system, a citizen-centric model is difficult to implement due to technological challenges presented by centralized databases. The Solid protocol [16], a decentralized data storage approach, is composed of specifications more conducive to construction of a citizen-centrica data storage strategy for clinical data. Specifically, Solid offers the ability to granularize data privacy, allow authorized data access over the web, and represent stored data as Linked Data, all features that can work to remove some of the antagonism between cost reduction and privacy preservation. In recent years, there have been initiatives for representing biological data as RDF [17], specifically extending into clinical biology recently [18]. While there is little research into the benefits of representing genomics data as RDF,

past experiments have shown that linked data integration into clinical practice results in improved outcomes [**farinelli_linked_2015**].

Furthermore, using Solid pods for data storage also makes it possible for non-linked data stored in the pod, such as test result files, to be linked to RDF data, improving data connectivity. As of yet, decentralized storage technologies have not meaningfully expanded to use in clinical practice. but if they did, things like data sharing, reduced data duplication, increased data privacy controls, could contribute to the PGS cost reductions and improved scalability. On the other side of the coin, the size and sensitive nature of PGS data provide a relatively unstudied frontier of decentalized web technology.

Link Traversal Query Processing (LTQP) To make sense of linked genomic and clinical data, approaches to parsing and querying that data must also be investigated, especially to encourage greater data discoverability and usage in clinical practice. Recent work has established that the querying of Linked Data in decentralized environments is possible [19], but these results were obtained with assumptions different than those presented by patient genome pods. Here, querying will be performed over a potentially large number of data pods containing large amounts of linked data, a situation not extensively investigated. In this context, it is likely that existing LTQP algorithms and approaches will require innovation.

It is documented how many personal data vaults of small amounts of data can be queried, but little has been done to investigate how many data vaults of large amounts of relatively similar data could be queried.

At present, the current state-of-the-art methods for data storage are centralized in nature, following an institution-centric model. This data storage strategy has posed great challenges to the scaling of personalized medicine, the use of patient genomic information to inform clinical decision-making.

Project Motivation. Despite there being no real solutions to the current antagonism between privacy and cost reduction for PGS data usage in healthcare, there is also a conspicuous gap in the current scientific discourse around the development and implementation of a proposed solution. This gap underscores the necessity of my Ph.D. **I aim to improve the connectedness and shareability of genomic data storage(s), while preserving data privacy, through the integration of various domains of semantic web research into a novel, holistic framework designed for use in clinical practice.** My Ph.D. will also aim to demonstrate the limitations of current state-of-the-art semantic web technologies in this novel application domain with the intention of driving innovation and discovering future research pursuits.

Scientific research objective(s)

My proposed research endeavors to fuse cutting edge semantic web technologies with decentralization technologies into a novel proof-of-concept PGS data storage and sharing framework for use in clinical practice. To realize such a framework, I will apply the technologies of five distinct areas of active research to a data ecosystem to which they have not been designed for. This ambition frames the central research question I aim to answer: Can combining the Solid specifications for data storage with other compatable cutting edge innovations for data policy, linking, and querying be instantiated and deployed as a framework that provides clear advantages over the existing PGS data storage protocols in health care?

The core research question can be decomposed into four specific research questions. First, can the decentralized storage protocol Solid, with which there have been few implementations of large data storage, and no published implementations of clinical genomic data storage, offer suitable storage infrastructure for this novel data. I hypothesize that the Solid protocol will be able to store clinical genomic data storage.

Second, to offer usage advantages over existing systems, can the representation of PGS data using the Resource Description Framework (RDF) as Linked Data be accomplished? A further aim

within this objective is exploring if storage of PGS data as RDF using Header Dictionary Triples (HDT) format [20] provides similar levels of usability of genomic data while optimizing storage efficiency. With the motivation of optimizing efficiency, I will also investigate the use of a bi-directional mapping index for the conversion between VCF and RDF.

Third, because of the sensitive nature of PGS data, are Solid data policy specifications for securing the privacy of PGS data while also allowing for increased sharability for authenticated requests sufficient? Additional investigation into how pre-computed privacy-preserving summaries [21] of genomic data within patient pod ...

Fourth, for the stored genomic data and linkages to be usable in clinical practice, a querying method is necessary. In a citizen-centric clinical data storage implmenetation, there are potentially thousands of heterogeneous, large sources to be queried over. It is an established challenge for federated decentralized querying in such environments, informing the question: can querying over these sources be achieved through the use of Link Traversal Query Processing algorithms that are modified to utilize pre-computed or on-the-fly generated indexes?

Together, these objectives will serve as the components of an operational framework. The framework, once produced, will be compared to existing strategies for storing and sharing PGS data to assess the efficacy of transitioning toward product production and specific clinical use case adaptation. The proposed scientific approach also aims to test the application of numerous fields of semantic web research to a clinical knowledge domain. In the process, providing insight into how many large data vaults informs future areas of innovation in decentralized web research.

3. Research methodology and work plan

My research plan consists of three component objectives, representative of three core functionalities of my proposed framework. First, I focus on the foundational infrastructure for data storage and formatting for the framework. Second, I focus on framework data privacy policies for granular, flexible data policy enforcement. Third, I integrate querying functionality to the data storage framework using a query engine approach and modified LTQP algorithmic approach to allow for data discoverability. After each component, I elaborate on its risks. Then, I present my work plan.

Component 1: Storage and formatting PGS data in a citizen-centric architecture

The foundation of my proposed framework is the data storage infrastructure. To implement a citizen-centric storage of PGS data, a decentralized storage strategy offers a maleable platform.

Task 1.1: Storing PGS data in Solid data vaults

The decentralized storage technology I chose to use is Solid, because of its growing populatiry[c] and its support of specifications for data sharing over the web and granular provacy controls...

Here, I will test the viability of Solid data pods for patient PGS data storage, thus, testing my hypothesis that Solid can support PGS data storage. Using my experience with genomic data types and file representations, I will assemble a test dataset composed of publicly available genome files [**noauthor_platinum_n**
These files will be used as representative ”patient” PGS data for all future experimentation. I will also create server-hosted Solid pods using the Community Solid Server (CSS) implementation of Solid [**css**]. Each pod will be a storage container for a single individual's PGS data. I will upload a single PGS file, a VCF file, into one ”patient's” pod to test basic functionality of a Solid pod for hosting large genomic data. The result of this task will be the **development and documentation of a workflow for creating, hosting, and uploading PGS data into patient solid pods**.

Task 1.2: Converting PGS data as Linked Data using RDF

To campitalize on data storage efficiency and future application potential, I will convert PGS data from VCF to RDF allowing for linking of other medically relevant data to patient genomic data within the patient's pod and outside of it. This aim will address well documented current challenges in medicial record utilization relating to scatteredness of pertinent clinical information[c]. To convert PGS data from VCF to RDF, we will investigate a format translation process using the SPHN RDF ontology [18]. For this translation process, we will experiment with different approaches, such as a bidirectional mapping index, for efficient reversal of conversion to ensure the conversion process can be reversed for connecting to existing clinical workflows that request VCF format inputs. Direct conversion between VCF and RDF will be evaluated in terms of computational overhead, conversion time, and memory usage. The same evaluations will be performed use the generation and use of an intermediate mapping index file. These comparisons will be documented in a formal benchmarking study.

Data that is serialized as RDF can be represented in a number of formats in computer memory [c]. To minimize the storage costs of large PGS data, I will utilize HDT format to compress the PGS data while retaining the ability to query and index it [c].

I then intend to demonstrate the linking of part of a patient's genome to (A) other data within the patient's pod, (B) data in a public database outside of a patient's pod, and (C) data from another patient's pod. The power of linking the VCF data to other clinically relevant data will be during querying, which will be performed in Component 3.

Risks

The main risk of storing PGS data in Solid data vaults (Task 1.1) is related to the size of PGS data.

The main risk of converting PGS data to Linked Data using RDF (Task 1.2) is that an ontology that semantically represents PGS data is required for this conversion. To mitigate this risk, I will only focus

on converting VCF data to RDF, which will make use of the publicly available SPHN RDF ontology. If this ontology is insufficient, I will work with members of our group with experience in ontology definition to create my own ontology for the conversion process which may take some time but will allow for progression to later work packages.

Component 2: PGS data privacy policies

Task 2.1: Granular, flexible data policies

To effectively store sensitive information, privacy protections are of paramount importance.

I will experiment with the design and implementation of multiple levels of authorization as well as methods that allow for dynamic control over data discoverability, read/write access, and data access consent requests within a patient's Solid pod, made possible by the Solid specification[c]. I will develop and test three functionalities for privacy modifications. (1) registration of a pod to an individual patient, (2) submission of a request to access stored data from a data requester, the notification of the patient, and the consent or denial by the patient, and (3) permission revoking capabilities as well as an opt-in option to share their data with researchers. All of these methods will be integrated into the framework's web application. To utilize these methods, various levels of access to pod read and write privileges will be created to fill the needs and roles of participants of a PGS clinical workflow. Attaching differing levels of authorization to data will be assessed by creating various profiles that reflect clinical roles and access levels and attempting to access data via user-mediated, application requesting, and querying approaches.

Task 2.2: Privacy-preserving data vault summaries

Protecting the privacy of patient data includes restricting the number of vulnerable access points of that data exist. For large numbers of pods aggregators can be utilized to improve query speed and efficiency by limiting the query to only relevant data sources. For private genomic data, aggregators accessible outside of patient data vaults would infringe on the privacy of PGS data stored within patient data vaults. By using a strategy that incorporates ...

Risks

Privacy is a tricky subject especially in terms of governance

If these secure summaries cannot be ...

Component 3: Querying PGS data over one and many data vaults

The problem space presented by citizen-centric data vaults is novel for federated decentralized environemnt querying. Established techniques for federated SPARQL querying over a *small number of large sources* have been documented [**hibiscus**,**tpf**,**sparql_adaptive_anapsid**]. Techniques for federated SPARQL querying over decentralized environments with a *large number of small data sources* have also been documented[c]. Personal genomic data vaults for clinical practice present a third querying landscape consisting of a *large number of large data sources*. One advantage presented by the PGS data stored in the present situation is that the data is formatted in a homogeneous way allowing for summaries to be used for improving query performance[c]. Concretely, I will assess and improve algorithms for query execution guided by PGS data indexes, and assess the performance of these algorithms compared to current PGS parsing methods...?

Task 3.1: Link Traversal Query Processing algorithm benchmarking

This work package will establish a querying mechanism for data in the patient Solid pods that takes into account patient pod data, user permissions, and data linkages. I hypothesize that a querying functionality that utilizes a query engine computational strategy will be able to query over patient Solid pods and return query results.

I will executing queries across PGS data contained in patient pod(s) through the use of the query

language SPARQL [**noauthor_sparql_nodate**]. Query execution requires a source for computation which is not currently provided by the Solid pods themselves. I will investigate the use of a query engine, such as that offered by Comunica [**comunica**], to perform the queries apart from the data stores.

For PGS data querying, I will benchmark and potentially build upon the link traversal query processing (LTQP) paradigm [19], which has been shown to be an effective method for querying within Solid.

Query engine functionality will be evaluated using query execution time and computational load metrics as well as query results assessment. Query results will establish the functionality of data linkages from WP2.

Benchmarking will be done for existing LTQP algorithms and altered query algorithms that utilize genomic index files and results will be compared. Ideally, success will be determined by queries that return correct results in under 10 minutes for users and potentially longer for applications. In a clinical setting, time constraints are not as important as accuracy and reliability of results although excessive query times decrease the usefulness of such a tool for physicians in clinical practice.

LTQP algorithms are an active area of active research, but most of the work done has been with generalized algorithms and ,amy datastores with small amounts of data. I aim to adapt this querying approach to the specific domain of genomic and health data which has not been attempted before.

Task 3.2: Algorithm incorporation of PGS data indexes and data vault summaries

I will look to innovate and improve performance by combining existing algorithms with strategies that leverage the unique structure of PGS data such as the use of pre-computed indexes, like the one generated for RDF-VCF conversion, as a guide for faster query processing.

Risks

benchmarks cannot be run because PGS data is too big

incorporating indexes in LTQP algos doesnt improve performance enough to be usable...

Component 3: Querying PGS data over one and many data vaults

Component consolidation and framework deployment

In an effort to improve data flows for research purposes, I intend to connect the proposed framework to the international Beacon initiative [15] to increase the availability of genomic data for researchers. In this aim we will investigate the necessary requirements and infrastructure necessary to connect patient Solid pods, containing PGS data, as beacon endpoints that can be discoverable and queried via the Beacon API. The connection of a decentralized, citizen-centric storage framework to the Beacon network is novel in nature as all other existing endpoints are institution-centric relational databases maintained by hospitals or research institutions.

All other functionalities will also be packaged into a web application with supporting documentation for final deployment and exhibition of how such a framework could function in clinical practice. This framework would be the first of its kind.

Beacon API connection will be evaluated on functionality and integrate all previous work package components. Similarly, evaluation of the web application from which a user can interact with the framework will also be based on functionality.

Work plan

My project consists of 4 work packages. This will be bundled into a Ph.D. dissertation for a final thesis defense. Below Gantt-chart of the Ph.D. project on a quarterly basis. Each work package is split up into different tasks with a dedicated amount of time allocated to it. This will allow for a good time and project management.

WP1: Storage and formatting PGS data in a citizen-centric architecture ***10 months***

- Task 1.1: Storing PGS data in Solid data vaults *4 months*
- Task 1.2: Converting PGS data as Linked Data using RDF *6 months*

WP2: PGS data privacy policies ***11 months***

- Task 2.1: Granular, flexible data policies *6 months*
- Task 2.2: Privacy-preserving data vault summaries *5 months*

WP3: Querying PGS data over one and many data vaults ***9 months***

- Task 3.1: Link Traversal Query Processing algorithm benchmarking *5 months*
- Task 3.2: Algorithm incorporation of PGS data indexes and data vault summaries *4 months*

WP4: Querying PGS data over one and many data vaults ***6 months***

I will undertake work packages and tasks as shown below in a Gantt chart. The primary work packages of my Ph.D. are WP1 and WP3. WP2 is a suporting work package and WP4 will be completed by combining all other work packages. WP1 and WP3 are dependent, meaning that WP1 must be partially completed before WP3 can begin. Thus, these work packages are planned sequentially, which provides me with knowledge of how data storage architecture could be leveraged for WP3.

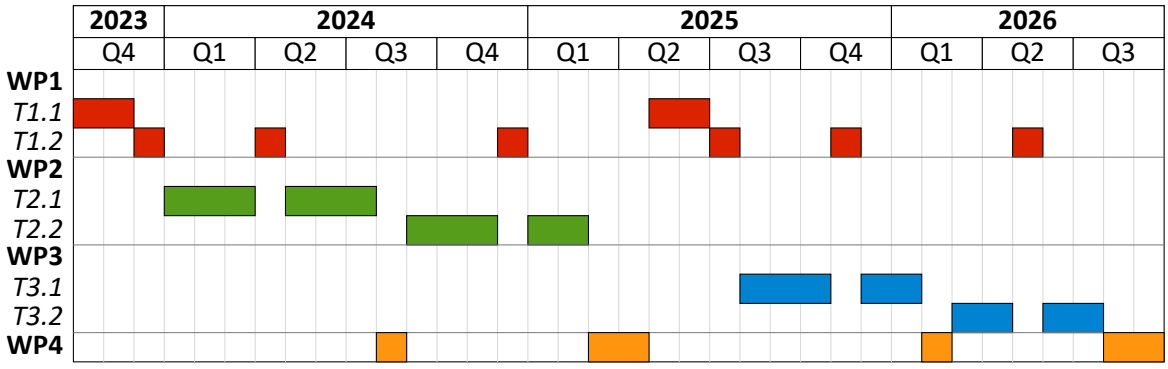
The focus of WP1 is on storing and formatting PGS data. The development of a workflow for setting up Solid data vaults and getting PGS data into those vaults is the first step (Task 1.1). Once the PGS data is stored, investigation into data format conversion and data representation will be performed (Task 1.2).

For WP2, I will apply privacy protecting policies to stored PGS data and summaries of that data. This task will utilize the Solid protocol for data policies within data vaults (Task 2.1). For summaries stored in aggregators outside of data vaults I will investigate the used of $_{(Task 2.2).}$

In WP3, I will build on the work of WP1 and assess the query performance of LTQP algorithms on genomic data vaults. I will start by benchmarking existing LTQP algorithms in a formal benchmarking study. Then, I will develop modified LTQP algorithms that use genomic indexes and data vault summaries to improve query efficiency.

I will unify the three framework components into a web application (WP4) that will be deployed as a demonstration complete with documentation and a live presentation component. Throughout the previous work packages, I aim to publish incremental findings in high-impact journals and conferences. For this, I will focus on journals such as the Semantic Web Journal, Journal of Healthcare Informatics Research, and the PLOS Digital Health Journal. Furthermore, I will target conferences such as the International Semantic Web Conference, the Semantic Web Applications and Tools 4 Health Care and Life Sciences conference, and the European Conference on Computational Biology.

Aside from disseminating to the scientific community, I will also community my research to the industrial community and wider public throughout the project (not included in the Gantt chart). Concretely, I will write blog posts, publish videos, and interact with the greater community via X.



Strategic dimension and application potential

My Ph.D. project is strongly motivated by the potential economic and societal gains presented by personalized medicine. For all existing and future applications of genetically-informed precision health, patient genome sequence data in some form will be required. In recent years, the cost of digital genome sequence generation has steadily decreased[c], but over that same time the cost of storing and maintaining the privacy of that data has not kept pace[c]. Thus, unintended barriers to scaling current clinical genomic workflows as well as to researching new workflows have been observed.

The the framework that will result from my Ph.D. will be uniquely positioned to compete with the current state-of-the-art intitiution-centric data storage systems due to the flexibility and cost efficiency it offers as a citizen-centric approach. The approach accomplishes this cost efficiency though infrastructure for privacy preserving patient genomic data sharing, data policy customization, and integrated data querying capabilities. Because most data is confined within a single health care institution, data sharing between institutions is an economic niche that is largely unfilled.

The private genomic service industry dominated by companies such as 23andMe, Ancestry.com, sequencing.com, and others establishes that genomics data generation and storage holds importance to consumers for various personal and medical reasons. At the same time, hospital systems exclusively store and maintain all patient PGS data that is used for clinical applications. There is notable nuance between these two sectors including different forms of genomic data being generated, stored, and used, differing legal oversight concerning commercial genomic data and health data, and formatting differences between the genomic data stored. Regardless, in our modern age of big data, data duplication due to data siloing, energy waste due to computational demands during data regeneration, and intrinsic security concerns for modern data storage techniques are major economic inefficiencies of the current system.

A hypothetical company that, in coordination with policy makers and regulatory bodies, creates a scalable storage and data sharing infrastructure for genomic data, which could also grow to include all patient health data in time, stands to greatly increase the efficiency of PGS data usage in healthcare. Such efficiency increases could help lower patient costs for specialized genetic tests, remove data management and administration from hospitals, thereby reducing costs, and establish a new market within which economic growth could result.

My project presented above is designed to present a proof-of-concept framework, both providing and demonstrating the technological foundations for the storage of PGS data in Solid pods, the controlling of access to that data on a granular level, the ability for that data to be queried, and exhibiting the accessibility of the stored PGS data to users, web applications, and medical tools in formats that can be used by both those currently in use and applications developed in the future. Such a framework will provide the outline of necessary implementation considerations from a technological perspective while also highlighting strengths and weaknesses of such a system that may be influential in attempts at scaling such an infrastructure. My project is also being undertaken in parallel with the European Virtual Human Twin (EDITH) [**edith**] initiative that aims at evolving the way medical data is stored to be increasingly citizen/patient centric both within Flanders as well as the greater European Union. The WE ARE project [**weare**] is another Flemish initiative exploring ways in which citizen-centric data stprage and ownership.

In the short-term, the project is being developed to be integrated into ongoing research and product development at VITO Digital Precision Health. Products aimed at improving the way drug prescription is practiced by using a genetic screening tool that leverages documented genetic predispositions to drug ineffectiveness are currently being developed to be connected to my framework of Solid pod stored PGS data. Additionally, connection to other known and widely-used workflows such as for NIPT and rare genetic disease screening is a primary goal for my project. Genomic data interoperability is of utmost importance for clinical application and is therefore a cornerstone of my project.

Lastly, public perception is a crucial element to the economic growth of a product or sector. With personal data usage transparency as well as greater calls for digital data privacy protections becoming more important to the public, such considerations should also be priorities to how health data is managed. The existing system of genomic data storage for use in healthcare is prone to data leaks and heavily restricted patient transparency due to the central architecture of institution-centric data stores. With my proposed framework, patients would be more intimately connected to their data, potentially even having a say over to whom and what their data is visible. Such improved transparency, when paired with decreased risk of large-scale data leaks, is likely to be well-received by the general public. Such public support could help drive such a framework adoption to a larger scale such as nationally or even to be the standard for a system like the EU. This large scale goal, while nowhere near attainable in the near future, would present the greatest possible outcome for such a project and exhibit a somewhat unintuitive increase in greater genomic data privacy and shareability. In this scenario, there is also room for healthy competition within such a niche as various pod providers could offer hospital systems and educational institutions different rates for data storage and associated computation.

References

[1] E. Mansour, A. V. Sambra, S. Hawke, M. Zereba, S. Capadisli, A. Ghanem, A. Aboulnaga, and T. Berners-Lee. A demonstration of the Solid platform for social web applications. In *Proceedings of the 25th International Conference Companion on World Wide Web*, pages 223–226, 2016.

[2] M. Zignani, S. Gaito, and G. P. Rossi. Follow the "Mastodon": structure and evolution of a decentralized online social network. In *Twelfth International AAAI Conference on Web and Social Media*, 2018.

[3] T. Kuhn, R. Taelman, V. Emonet, H. Antonatos, S. Soiland-Reyes, and M. Dumontier. Semantic micro-contributions with decentralized nanopublication services. *PeerJ Computer Science*, March 2021.

[4] R. Cyganiak, D. Wood, and M. Lanthaler. RDF 1.1: Concepts and Abstract Syntax. Recommendation, W3C, February 2014. URL: *https://www.w3.org/TR/2014/REC-rdf11-concepts-20140225/*.

[5] T. Berners-Lee. Linked data. 2009. URL: *https://www.w3.org/DesignIssues/LinkedData.html*.

[6] S. Harris, A. Seaborne, and E. Prud'hommeaux. SPARQL 1.1 Query Language. Recommendation, W3C, March 2013. URL: *https://www.w3.org/TR/2013/REC-sparql11-query-20130321/*.

[7] L. E. Hood, M. W. Hunkapiller, and L. M. Smith. Automated DNA sequencing and analysis of the human genome. *Genomics*, 1(3):201–212, November 1987.

[8] Y. K. Ko and J.-A. Gim. New drug development and clinical trial design by applying genomic information management. *Pharmaceutics*, 14(8):1539, July 24, 2022.

[9] H. L. McLeod. Cancer pharmacogenomics: early promise, but concerted effort needed. *Science*, 339(6127):1563–1566, March 29, 2013.

[10] E. Souche, S. Beltran, E. Brosens, J. W. Belmont, M. Fossum, O. Riess, C. Gilissen, A. Ardeshirdavani, G. Houge, M. Van Gijn, J. Clayton-Smith, M. Synofzik, N. De Leeuw, Z. C. Deans, Y. Dincer, S. H. Eck, S. Van Der Crabben, M. Balasubramanian, H. Graessner, M. Sturm, H. Firth, A. Ferlini, R. Nabbout, E. De Baere, T. Liehr, M. Macek, G. Matthijs, H. Scheffer, P. Bauer, H. G. Yntema, and M. M. Weiss. Recommendations for whole genome sequencing in diagnostics for rare diseases. *European Journal of Human Genetics*, 30(9):1017–1021, September 2022.

[11] S. Mathur and J. Sutton. Personalized medicine could transform healthcare. *Biomedical Reports*, 7(1):3–5, July 2017.

[12] D. Stefanicka-Wojtas and D. Kurpas. Barriers and facilitators to the implementation of personalised medicine across europe. *J Pers Med*, 13(2):203, January 23, 2023.

[13] P. Danecek, A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E. Handsaker, G. Lunter, G. T. Marth, S. T. Sherry, G. McVean, R. Durbin, and. The variant call format and VCFtools. *Bioinformatics*, 27(15):2156–2158, August 1, 2011. URL: *https://academic.oup.com/bioinformatics/article/27/15/2156/402296* (visited on 07/27/2023).

[14] Ransomware attacks on hospitals have changed | cybersecurity | center | AHA. URL: *https://www.aha.org/center/cybersecurity-and-risk-advisory-services/ransomware-attacks-hospitals-have-changed* (visited on 02/09/2024).

[15] J. Rambla, M. Baudis, R. Ariosa, T. Beck, L. A. Fromont, A. Navarro, R. Paloots, M. Rueda, G. Saunders, B. Singh, J. D. Spalding, J. Törnroos, C. Vasallo, C. D. Veal, and A. J. Brookes. Beacon v2 and beacon networks: a "lingua franca" for federated data discovery in biomedical genomics, and beyond. *Human Mutation*, 43(6):791–799, June 2022.

[16] S. Capadisli, T. Berners-Lee, R. Verborgh, and K. Kjernsmo. Solid protocol. URL: *https://solidproject.org/TR/protocol#introduction* (visited on 01/03/2024).

[17] . The SIB swiss institute of bioinformatics semantic web of data. *Nucleic Acids Research*, 52:D44–D51, D1, January 5, 2024. URL: *https://doi.org/10.1093/nar/gkad902* (visited on 02/09/2024).

[18] E. Van Der Horst, D. Unni, F. Kopmels, J. Armida, V. Touré, W. Franke, K. Crameri, E. Cirillo, and S. Österle. Bridging Clinical and Genomic Knowledge: An Extension of the SPHN RDF Schema for Seamless Integration and FAIRification of Omics Data. preprint, Medicine and Pharmacology, December 6, 2023. URL: *https://www.preprints.org/manuscript/202312.0373/v1* (visited on 01/22/2024).

[19] R. Taelman and R. Verborgh. Evaluation of link traversal query execution over decentralized environments with structural assumptions, 2023. URL: *https://arxiv.org/abs/2302.06933* (visited on 09/11/2023). Publisher: arXiv Version Number: 1.

[20] Technical Specification – RDF HDT. Technical report. URL: *https://www.rdfhdt.org/technical-specification/*.

[21] R. Taelman, S. Steyskal, and S. Kirrane. Towards querying in decentralized environments with privacy-preserving aggregation. (arXiv:2008.06265), August 14, 2020. arXiv: *2008.06265[cs]*. URL: *http://arxiv.org/abs/2008.06265* (visited on 02/21/2024).