PHD FELLOWSHIP STRATEGIC BASIC RESEARCH
PROJECT OUTLINE (MAX. 12 A4 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 (RDF) [4], represented as triples with universal semantics, built on Linked Data principles [5], and information extraction via querying using the SPARQL query language [6]. Implementations of Solid for industry use are emerging [7], but other challenges presented by data querying infrastructure and governance remain.

Personal Genome Sequencing in Healthcare Around the same time that the World Wide Web was being established, DNA sequencing technologies were just being applied to the human genome [8]. 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 [9], cancer diagnosis and treatment [10], and rare genetic disease identification and treatment [11]. How this integration is deployed varies by clinical domain, but improved outcomes have generally been observed [12]. Despite great promise presented by various use cases, barriers to widespread adoption remain [13].

One major barrier to scalability is presented by the costs of PGS data generation and storage [14]. 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 [15]. Of these, Variant Call Format (VCF) files [16] serve as the state-of-the-art for most clinical genomic applications and are **typically between 100-1000s MBs** within computer memory. Because of their large size, there are existing state-of-the-art methods for compressing [17] and parsing [18] these files using indexes.

The costs of producing and maintaining these data are further increased by the **privacy protections needed for PGS data** [19]. With the enlarged threat of hacking, phishing, and login credential compromisation that is only increasing [20], health care institutions have taken steps to enact tighter restrictions on data access and increase cyber security budgets. Because of this data siloing, there is **little to no data sharing between health care institutions**. If a patient moves hospitals, it is common for PGS data and genomic tests to be regenerated and indefinetly stored in that new location because of the lack of data sharing infrastructure. To reduce costs and improve the scalability of PGS usage in clinical practice, alternative methods of data storage and privacy protection that allow for improved data sharing are merited. *I will establish user-friendly methods for the management and protection of stored PGS data while also making it possible to share that data without compromising its privacy.*

Decentralized PGS data storage The **citizen-centric model** places the patient at the center of their data, and is not an entirely novel concept [21]. Within the current system, a citizen-centric model is difficult to implement due to technological challenges presented by centralized, relational databases. The **Solid protocol** [1], a decentralized data storage approach, is composed of specifications **more conducive to construction of a citizen-centric data storage strategy** for clinical data. Specifically, Solid offers the ability to **granularize data privacy policies**, allow **authorized data access over the web**, and represent stored data as **Linked Data** [5], 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 [22], specifically extending into clinical biology recently [23]. While there is little research into the benefits of representing genomic data as RDF, past studies have shown that **linked data integration into clinical practice results in improved outcomes** [24]. Furthermore, using Solid Pods for data storage also makes it possible for non-linked data, such as test result files, to be linked to RDF data, improving data connectivity. As of yet, **decentralized storage technologies** have not meaningfully been used in clinical practice. If implemented, they could provide improved data sharing, reduced data duplication, and increased data privacy controls that could **contribute to clinical PGS cost reductions and improved scalability**. Adaptation of Solid decentralized technology to clinical genomics does not come without challenges. Serious challenges are posed by the large size of genomic data, the interoperability of future storage technologies with current applications and tools, and the lack of existing infrastructure for implementation of a decentralized storage framework. To drive innovation in the field of health data storage, *I will create and implement workflows for creating, hosting, and using Solid Pods for the storage of patient genomic and clinically relevant data.*

Decentralized Data Querying A challenge of Solid personal data vaults is that they do not intrinsically have access to computation. Thus, **to read data within a Solid pod, a query engine is needed for accessing and parsing stored data**. Query engine implementations within Solid are an ongoing area of research, and one framework for research-based implementions is provided by Comunica [25].

There are also **algorithmic challenges presented by querying genomic data vaults**. The processing of queries can be characterized by two stages, the planning and the execution. The traditional strategy followed a sequential optimize-then-execute approach, where query planning is done first based on pre-existing dataset statistics then the produced plan guides execution. For cases where there are not dataset statistics available, adaptive processing has been presented as a solution, where planning and execution are recursively performed throughout the querying process [26].

SPARQL endpoints [27], large centralized well-indexed databases, represent examples of environments where query planning and execution are guided by indexes to improve performance [28]. Federated querying algorithms [29,30] build on those approaches for querying over multiple, but not a large number of, SPARQL endpoints.

For querying decentralized ecosystems, challenges are presented by the larger number of sources and inconsistency of indexes or statistics about those sources available prior to execution. Due to these constraints, recent work presenting Link Traversal Query Processing (LTQP) algorithms has established new approaches for federated querying in these decenralized environments [31]. For additional efficiency and specialized use cases, the use of aggregators that store summaries representing data types or other statistics about sources has also been proposed [32,33].

The established LTQP algorithms assume conditions where *many sources contain small amounts of data*, which is different than those presented by patient genome pods. In the case of genomic data pods, 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 query planning strategies will require innovation.** *I will develop new LTQP algorithms that are capable of querying deventralized genomic data vaults through the use of genome indexes and data vault summaries.*

Towards clinical implementation Despite there being no real solutions to the current antagonism between privacy and cost reduction for PGS data usage in health care, 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 cutting edge of semantic web research in the domains of **data storage, policies, and querying**. *I will accomplish this by creating a novel, holistic framework that is implemented as a web application, complete with guides and documentation to foster usability, for use in clinical practice.* Through this framework, I will also demonstrate the limitations of current state-of-the-art semantic web technologies in this novel application domain to drive innovation and uncover future research pursuits.

Scientific research objective(s)

My proposed research endeavors to combine cutting edge decentralized storage technology with semantic data representation and federated querying technologies into a novel proof-of-concept PGS data storage and querying framework for use in clinical practice. To create such a framework, I will combine technologies from different distinct areas of semantic web research and apply them to a data ecosystem that poses novel challenges. 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 in 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 more specific research questions. *First, can the decentralized storage protocol Solid* [1] *offer suitable infrastructure for PGS data?* I hypothesize that the Solid protocol will be able to store clinical genomic data. Further, I aim to establish Solid also offers usage advantages over existing systems through the representation of **PGS data using RDF as Linked Data** [34]. A further aim within this objective is exploring if storage of PGS data as RDF using **Header Dictionary Triples (HDT) format** [35] provides similar levels of usability of genomic data with significantly **decreased storage costs**. To ensure the widest range of connection capabilities while optimizing efficiency, I will also investigate the use of a **bi-directional mapping index** for the conversion between native genomic VCF files and RDF representations. For these aims, my background knowledge of genomic file anatomies and how data is semantically represented in genomics file formats will be valuable, specifically when optimizing indexing and format conversion strategies.

Second, because of the sensitive nature of PGS data, *do the specifications provided by Solid provide for adequate control of PGS data privacy while also allowing for increased authorized sharability?* I aim to demonstrate possible configurations of Solid data vault privacy policies as well as offer a functionality within a **web application for the alteration of these policies** by authorized users. I am uniquely situated for assessing these privacy policies because of my past experience working in a clinical setting alongside physicians as well as my understanding of different sets of US and EU regulations mediating data privacy requirements.

Third, 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 implementation, there could be potentially thousands of large sources to be queried over. Performant federated decentralized querying in such environments is an established challenge [36]. Therefore, *can querying over these sources be achieved through the use of LTQP algorithms?* **I anticipate that current generalized LTQP algorithms will not be able to perfrom well over large genomic data**. Thus, I will investigate two strategies for improving the performance of query processing. (A) I will investigate the **use of summaries** of patient data vault contents, stored outside of data vaults in aggregators, to modify query planning strategies in LTQP algorithms. Because of the privacy considerations inherint with patient data vault contents, I will follow presviously descibed theoretical methods to maintain data privacy in these summaries [33]. (B) I will develop LTQP algorithms that **integrate genomic data indexes** for within-vault querying. I have experience with index guided genomic data parsing which will help inform developing and optimizing query processing algorithms to incorporate these guides.

Together, these components will be **combined into an operational framework** in the fourth component. The driving question being, *can these three different groups of features be combined into a cohesive web application and deployed together?* 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, insight into how unique challenges introduced by clinical constraints will provide future areas of research, both applied and fundamental.

Research methodology and work plan

My research plan consists of three component objectives, representative of three core functionalities of my proposed framework. A final fourth component will be the unification of the three functional components into a web application for deploying the 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. Last, 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 increase the efficiency of data storage and usage, a citizen-centered data storage approach will be attempted. This organizational strategy is not feasable in centralized databases given current technologies, thus, I will utilize a decentralized storage approach. Of the decentralized storage initiaves [1–3], Solid was chosen because it is not social network specific, is growing in popularity, and has specifications useful for privacy and data sharing infrastructure.

Task 1.1: Storing PGS data in Solid data vaults

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 [37,38]. 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 [39]. The use of the CSS for Solid pod hosting for research purposes is state-of-the-art, but there have been no published experiments documenting the use of CSS pod instances for storing PGS data, which are much larger than in past Solid experimentation. 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 data vaults**.

Task 1.2: Representing PGS data as Linked Data using RDF

To improve 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 a 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 [40]. To convert PGS data from VCF to RDF, we will investigate a format translation process using the SPHN RDF ontology [41]. For this translation process, I will experiment with using a bi-directional mapping index for efficient reversal of conversion to ensure connection 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. Evaluations detailing the use of an intermediate mapping index file will also be done and compared to direct VCF to RDF conversion. These comparisons will be documented in a formal benchmarking study. Because representation of VCF files in RDF has not been heavily studied, these will be the first published experiments of their kind.

Data that is serialized as RDF can be represented in a number of formats in computer memory [42]. 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 [35]. This approach has not be applied to genomic data before.

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. While Linked Data is state-of-the-art, these concepts have not yet been applied to clinical genomic data.

Risks

The main risk of storing PGS data in Solid data vaults (Task 1.1) is related to the size of PGS data. I will have access to servers at UGhent and VITO NV where implementations of experimental Solid pods will not face size limitations that interfere with project progress.

The main risk of converting PGS data to Linked Data using RDF (Task 1.2) is that this conversion requires an ontology. The ontology offers semantic information about the PGS data being stored, thus the specific semantic standards used will be important for future discoverability through querying. To make produced RDF data as universally applicable as possible, I will only focus on converting VCF data to RDF, and this will be done using the publicly available SPHN RDF ontology [23]. If this ontology is insufficient, I will work with members of the IDLab at UGhent with experience in ontology definition to create my own ontology for the conversion process.

Component 2: PGS data privacy policies

Task 2: Implementing data policies

A large advantage of the Solid decentralized data storage protocol over current institution-centric methods of data storage is the more flexible methods of creating, modifying, and enforcing data access policies. 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. 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. Assigning the above permissions within Solid is an open area of research and there are currently state-of-the-art protocols implemented in the CSS that allow their implementation [39]. The described access schema has not been attempted in the presented level of detail for clinical genomic data.

Risks

If the above proposed schema for privacy policies cannot be achieved, a simplier and more generalized schema will be devised and implemented. Privacy is a nuanced subject especially in terms of governance concerning sensitive data. I aim to show the possibilities preseted by Solid in this framework, not dictate suggestions for its deployable implementations.

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

The problem space presented by citizen-centric genomic data vaults is novel for federated decentralized environment querying. Established techniques for federated SPARQL querying over a *small number of large sources* have been documented [29,30]. Techniques for federated SPARQL querying over decentralized environments with a *large number of small data sources* have also been documented [31]. 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 proposed manner is that assumptions about the data contained within patient data vaults can be made based on the conserved function of vaults. This homogeneity of data will be leveraged to inform query algorithm planning. Concretely, **I will assess and improve LTQP algorithms for genomic data vaults**. Novel improvements to existing algorithms will be attempted by using within patient vault PGS data indexes and outside patient vault privacy preserving summaries. Benchmarking studies will assess the performance of these algorithms among various possible clinical usecases.

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. To query these linkages, I will utilize the knowledge graph query language SPARQL [6]. Query execution requires a source for computation which is not currently provided by the Solid pods themselves. I will implement an instance of Comunica [25] to perform the queries. Because Comunica is open-source and intended for research purposes, it will enable experimentation with modified query processing algorithms.

For actual benchmarks, **I will assemble a suit of represetative patient genomic data vaults**. I will first assemble a test set of 10 patient data vaults to test basic functionalities, then two additional test sets of 100 patient data vaults and 1000 patient data vaults to represent real-world use cases where it is assumed that many pateints'data can be stored by my framework. Query functionality will be evaluated using query execution time and computational load metrics as well as a query results assessment. Query results will additionally establish the functionality of data linkages (Task 1.2). LTQP algorithms are an active area of active research, but most of the work done has been with generalized algorithms and datastores with small amounts of data [31]. I aim to adapt this querying approach to the specific domain of genomic and health data which has not been attempted before.

Benchmarking will be **initially performed for existing LTQP algorithms** [31]. Initially, success will be determined by queries that return correct results verified by a truth-set. In a clinical setting, time constraints are not as important as accuracy and reliability of results, motivating our primary assessment criteria. Because excessive query times may decrease the usefulness of such a tool for physicians in clinical practice, I will also assess the time it takes for queries to execute and attempt to minimize this time in the following tasks.

Task 3.2: Privacy-preserving data vault summaries

It is **unlikely that existing LTQP algorithms perform well** for genomic data vaults due to their large size and low connectivity to other patient data vaults. Therefore, I will experiment with the generation of **data vault summaries, stored in aggregators, that do not compromise the privacy of patient data**. To implement these privacy preserving summaries, I will first assemble within-data vault summaries with designated access controls. These summaries may be generated using some of the data from genomic bi-directional mapping indexes (Task 1.2). Then, one or multiple aggregator(s) will assemble multi-vault summaries that are stored outside the data vaults using a summary combination algorithm [33]. These summaries are intended to be used by modified LTQP algorithms (Task 3.3) to improve query processing time and efficiency. The described privacy preserving summaries have only been proposed in theory and the proposed implementation will be the first of its kind.

Task 3.3: Algorithm incorporation of PGS data indexes and data vault summaries

For the optimization of LTQP algorithm performance over PGS data vaults, I will look to improve existing algorithms by **incorporating previously generated aggregator summaries (Task 3.2) and genome data indexes (Task 1.2)**. I hypothesize the summaries will be specifically useful for query planning because they can help scope the number of data vaults that need to be queried among other benefits. Alongside the use of summaries, I also will experiment with the use of within data vault indexes for the querying of genomic data specifically. There are well established VCF file parsing tools that allow for highly performant parsing of VCF data via the use of an indexing strategy [18]. I indend to implement a similar strategy by using a pre-computed genome index (Task 1.2) to improve the performance of genome-specific queries. Combining the two together, a dynamic algorithmic approach that allows for summaries and genome indexes to inform query rewriting at differnt points during query compilation and execution will be examined.

The algorithms described above will be benchmarked using the same benchmarking set and evaluation criteria as generic LTQP algorithms (Task 3.1). I will additionally benchmark the query times of a single data vault PGS to existing tools VCF files parsing tools on the bases of speed, computational load, and result correctness.

The federated querting algorithms utilizing indexes and summaries proposed are novel in nature and have not been developed before.

Risks

Task 3.1 presents the risk that PGS data are too large for LTQP algorithms to execute over. If these issues arise, I will assess the execution of simpler queries as well as the execution of multiple, sequential queries over a subset of the genome data vaults.

Another risk is associated with the generation of secure summaries in Task 3.2. If these summaries cannot be created in a way that adheres to the privacy demands of PGS data, I will investigate the use of other, more secure methods. Another option that may be explored is the creation of aggregator data vaults, with privacy protections, where summaries can be stored to improve privacy protections.

Lastly, there is a risk that I cannot devise solutions to incorporating indexes and/or summaries in LTQP algorithms and/or these algorithm modification do not improve performance enough to be usable in Task 3.3. I will initially investigate how imposing limits on query complexity and reducing the number of data vaults that are included in the possible query space could improve performance. If there are still issues, I will investigate the implementation of algorithms utilized by centralized SPARQL endpoints that are known to be able to query large sets of data.

Component 4: Ph.D. Finalization

Task 4.1: Framework consolidation and deployment

The three components will be combined into a functional framework. The framework will include a **web application that offers a central location for accessing the functionalities discussed above**. The framework will be deployed and demonstrated as it could be used in clinical practice.

Task 4.2: Ph.D. dissertation composition and defense

The findings and results will be packaged into a **Ph.D. dissertation and defense**.

Risks

The main concern is that some of the functionalities or components of previous work packages will not be able to be integrated together into a single web application. If necessary, different applications will be created to accommodate any components that do not fit into the planned unified web application.

Work plan

My project consists of 4 work packages that correspond with the componenets presented above.

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: Representing PGS data as Linked Data using RDF	*6 months*
WP2: PGS data privacy policies	***8 months***
- Task 2: Implementing data policies	*8 months*
WP3: Querying PGS data over one and many data vaults	***18 months***
- Task 3.1: Link Traversal Query Processing algorithm benchmarking	*6 months*
- Task 3.2: Privacy-preserving data vault summaries	*6 months*
- Task 3.3: Algorithm incorporation of PGS data indexes and data vault summaries	*12 months*
WP4: Ph.D. Finalization	***12 months***
- Task 4.1: Framework consolidation and deployment	*6 months*
- Task 4.2: Ph.D. dissertation composition and defense	*9 months*

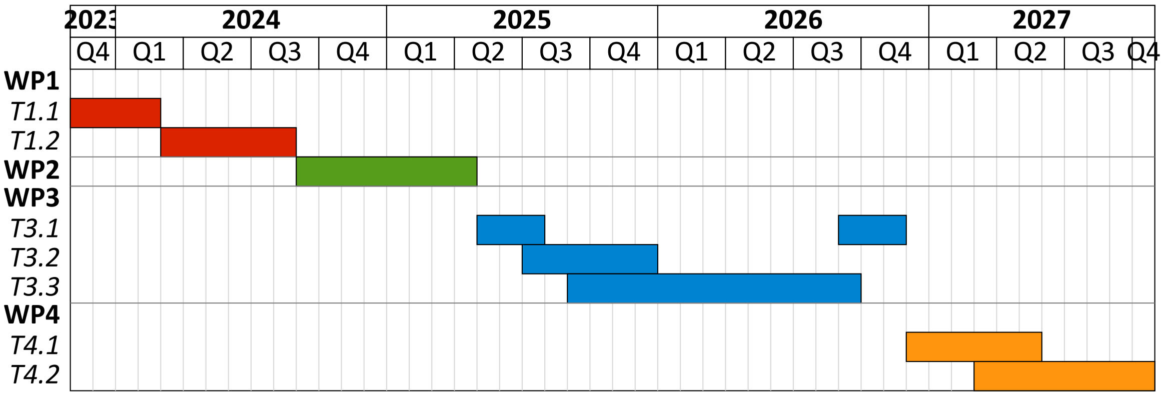
A Gantt chart details when tasks will be undertaken. The primary work packages of my Ph.D. are WP1 and WP3. WP2 is a supporting work package and WP4 will be completed by combining all other work packages. WP1 and WP3 are dependent, meaning that WP1 must be partially completed before WP3 can begin. Thus, 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 using the Solid protocol (Task 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 (Task 3.1). To optimize query performance, I will experiment with the production of privacy preserving summaries stored in aggregators outside of data vaults (Task 3.2). Then, I will develop modified LTQP algorithms that use genomic indexes and data vault summaries to improve query efficiency (Task 3.3).

I will unify the three framework components into a web application (Task 4.1) that will be deployed as a demonstration complete with documentation and user guides. Throughout the previous work packages, I aim to publish incremental findings in high-impact journals, such as the Semantic Web Journal and Journal of Healthcare Informatics Research, and conferences, such as the International Semantic Web Conference and the European Conference on Computational Biology. The completed Ph.D. will be combined into a dissertation and defended in Fall 2027 (Task 4.2).



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 [43], but over that same time the cost of storing and maintaining the privacy of that data has not kept pace [44]. Thus, unintended barriers to scaling current clinical genomic workflows as well as to researching new workflows have been observed.

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. More generally, here I propose a way to separate genome data hosting and sharing from its applications to encourage more cost efficiency. Because most data is predominantly 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.

An organization 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 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) [45] initiative and The European Health Data Space (EHDS) that aim 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 [46] is another Flemish initiative exploring the development product-level applications on top of citizen-centric data storage in Solid pods. To be compliant with, and hopefully to compliment, these initiatives, all data stored in my framework will be formatted according to FAIR data principles [47].

In the short-term, the project is being developed to be integrated into ongoing research and product development at VITO in the department of Digital Precision Health. One specific use case of my framework is its integration into a pharmacogenomics tool for the identification of documented adverse drug reactions from a VCF genomic input file. My framework would provide the file storage infrastructure for inputs and outputs as well as potentially other functionalities, such as representation fo the output as data linked to the area of the patients genome an adverse reaction is found. Additionally, my framework has potential to store or link data used for other known and widely-used clinical genomics workflows such as for NIPT and rare genetic disease screening.

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] Athumi. URL: *https://athumi.be/* (visited on 02/26/2024).

[8] 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.

[9] 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.

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

[11] 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.

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

[13] 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.

[14] G. A. Alarcón Garavito, T. Moniz, N. Déom, F. Redin, A. Pichini, and C. Vindrola-Padros. The implementation of large-scale genomic screening or diagnostic programmes: a rapid evidence review. *Eur J Hum Genet*, 31(3):282–295, March 2023.

[15] F. O. Bagger, L. Borgwardt, A. S. Jespersen, A. R. Hansen, B. Bertelsen, M. Kodama, and F. C. Nielsen. Whole genome sequencing in clinical practice. *BMC Medical Genomics*, 17(1):39, January 29, 2024.

[16] 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. (Visited on 07/27/2023).

[17] R. Wertenbroek, S. Rubinacci, I. Xenarios, Y. Thoma, and O. Delaneau. XSI—a genotype compression tool for compressive genomics in large biobanks. *Bioinformatics*, 38(15):3778–3784, June 2022.

[18] L. Yang, S. Jiang, B. Jiang, D. J. Liu, and X. Zhan. Seqminer2: an efficient tool to query and retrieve genotypes for statistical genetics analyses from biobank scale sequence dataset. *Bioinformatics*, 36(19):4951–4954, August 5, 2020.

[19] Regulation (EU) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/EC (general data protection regulation), May 4, 2016.

[20] 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).

[21] M. R. Brands, S. C. Gouw, M. Beestrum, R. M. Cronin, K. Fijnvandraat, and S. M. Badawy. Patient-centered digital health records and their effects on health outcomes: systematic review. *J Med Internet Res*, 24(12):e43086, December 22, 2022.

[22] . The SIB swiss institute of bioinformatics semantic web of data. *Nucleic Acids Research*, 52:D44–D51, D1, January 5, 2024.

[23] 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. (Visited on 01/22/2024).

[24] F. Farinelli, M. Barcellos de Almeida, and Y. Linhares de Souza. Linked health data: how linked data can help provide better health decisions. *Stud Health Technol Inform*, 216:1122, 2015.

[25] Comunica – a knowledge graph querying framework. Comunica – A knowledge graph querying framework. URL: *https://comunica.dev/* (visited on 02/12/2024).

[26] A. Deshpande, Z. Ives, and V. Raman. Adaptive query processing. *Foundations and Trends®*, 2007.

[27] L. Feigenbaum, G. Todd Williams, K. Grant Clark, and E. Torres. SPARQL 1.1 Protocol. Rec. W3C, March 2013. URL: *https://www.w3.org/TR/2013/REC-sparql11-protocol-20130321/*.

[28] M. Schmidt, M. Meier, and G. Lausen. Foundations of SPARQL query optimization. In *Proceedings of the 13th International Conference on Database Theory*, pages 4–33. ACM, 2010.

[29] M. Saleem and A.-C. N. Ngomo. Hibiscus: hypergraph-based source selection for SPARQL endpoint federation. In *European semantic web conference*, pages 176–191. Springer, 2014.

[30] R. Verborgh, M. Vander Sande, O. Hartig, J. Van Herwegen, L. De Vocht, B. De Meester, G. Haesendonck, and P. Colpaert. Triple pattern fragments: a low-cost knowledge graph interface for the web. *Journal of Web Semantics*, 37:184–206, 2016.

[31] R. Taelman and R. Verborgh. Evaluation of link traversal query execution over decentralized environments with structural assumptions, 2023. Publisher: arXiv Version Number: 1.

[32] M. Vandenbrande. Aggregators to realize scalable querying across decentralized data sources. In *International semantic web conference*, 2023. URL: *https://api.semanticscholar.org/CorpusID:265531325*.

[33] 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).

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

[35] J. D. Fernández, M. A. Martínez-Prieto, C. Gutiérrez, A. Polleres, and M. Arias. Binary rdf representation for publication and exchange (hdt). *Web Semantics: Science, Services and Agents on the World Wide Web*, 19:22–41, 2013. URL: *http://www.websemanticsjournal.org/index.php/ps/article/view/328*.

[36] M.-H. Dang, J. Aimonier-Davat, P. Molli, O. Hartig, H. Skaf-Molli, and Y. Le Crom. FedShop: a benchmark for testing the scalability of SPARQL federation engines. In *The Semantic Web – ISWC 2023: 22nd International Semantic Web Conference, Athens, Greece, November 6–10, 2023, Proceedings, Part II*, pages 285–301, Berlin, Heidelberg. Springer-Verlag, November 6, 2023.

[37] Platinum genomes. URL: *https://emea.illumina.com/platinumgenomes.html* (visited on 02/12/2024).

[38] 1000 genomes | a deep catalog of human genetic variation. URL: *https://www.internationalgenome.org/home* (visited on 02/23/2024).

[39] CommunitySolidServer/CommunitySolidServer: an open and modular implementation of the solid specifications. URL: *https://github.com/CommunitySolidServer/CommunitySolidServer* (visited on 02/12/2024).

[40] R. Pastorino, C. De Vito, G. Migliara, K. Glocker, I. Binenbaum, W. Ricciardi, and S. Boccia. Benefits and challenges of big data in healthcare: an overview of the european initiatives. *Eur J Public Health*, 29:23–27, Suppl 3, October 2019.

[41] 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.

[42] Resource description framework (RDF) serialization | LINCS. URL: *https://lincsproject.ca/docs/terms/resource-description-framework-serialization* (visited on 02/23/2024).

[43] K. A. Wetterstrand. The cost of sequencing a human genome. Genome.gov. November 1, 2021. URL: *https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost* (visited on 11/13/2023).

[44] J. C. McCallum. Historical cost of computer memory and storage. Our World in Data. 2022. URL: *https://ourworldindata.org/grapher/historical-cost-of-computer-memory-and-storage* (visited on 02/23/2024).

[45] M. Viceconti, M. De Vos, S. Mellone, and L. Geris. Position paper from the digital twins in healthcare to the virtual human twin: a moon-shot project for digital health research. *IEEE J Biomed Health Inform*, PP, October 11, 2023.

[46] We are. URL: *https://we-are-health.be/en* (visited on 02/21/2024).

[47] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. Silva Santos da, 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. G. Gray, P. Groth, C. Goble, J. S. Grethe, J. Heringa, P. A. C. ’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. Schaik van, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M. A. Swertz, M. Thompson, J. Lei van der, E. Mulligen van, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, and B. Mons. The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, 3(1):160018, March 15, 2016.