PErsoNal Genome QUery IN healthcare and clinical practice (PENGQUIN)

Elias Crum[1,2][0009−0005−3991−754X]

[1] IDLab, Department of Electronics and Information Systems, Ghent University - imec
[2] Flemish institute for Technological Research (VITO)

Abstract. Medical care is in the process of becoming increasingly personalized through the use of patient genetic information. At present, data useful for clinical care, including genetic data, is commonly diffuse, organized arbitrarily, and stored in data silos. Thus, unstructured organization, high costs for data storage and generation, and tight privacy restrictions pose serious challenges to scaling personalized clinical strategies. I propose an early stage Ph.D. that aims to improve the connectedness and shareability of genomic data storage(s), while preserving data privacy, to decrease the costs of using patient genome data in clinical practice. In this pursuit, I will integrate various domains of semantic web research into a novel, holistic framework designed for use in clinical practice. Specifically, I will (a) store patient data using Solid pods, (b) represent personal genome sequence data in RDF as Linked Data, (c) attach policies to stored data, and (d) query data through link traversal queries.

Keywords: Solid · Linked Data · Querying · Genomic Data Sharing

1 Introduction / Motivation

As our understanding of genomics deepens, the role of Personal Genome Sequencing (PGS) in healthcare is expanding. At the time of writing, there are multiple domains of clinical practice where patient PGS data is now used to inform medical decision making, including drug development [13], cancer diagnosis and treatment [17], and rare genetic disease identification and treatment [15]. How this integration is deployed varies by clinical domain, but improved outcomes have generally been observed [16]. Despite great promise, barriers to adoption remain [22]. One major challenge is presented by the digital representation, storage, and access to the PGS data that underlies clinical usage. With my Ph.D., I aim to address the challenges presented by PGS data usage in clinical practice by leveraging decentralized storage, data representation, and querying technologies.

PGS data storage and sharing. PGS data are expensive to both generate and store. The average human genome is slightly over 3 billion base pairs in length (3 Gbp). During a whole genome sequencing workflow, various sequence

formats that offer different sets of information are produced [6]. Of these, Variant Call Format (VCF) files [10] serve as the state-of-the-art for most clinical genomic applications. VCF files are typically between 100-1000s MBs within computer memory and represent around 3 million nucleotide positions of an individual's PGS. For these genomic data, there are also significant privacy considerations. Unfortunately, the relationship between privacy and cost is largely antagonistic within the current technological framework. Increases in privacy protections often lead to increases in data siloing and increased costs for both patient and provider. For instance, if a patient moves hospitals, it is common for PGS data and genomic tests to be regenerated and indefinitely stored in that new location because of the lack of data sharing between institutions. One strategy for reducing costs is through increased data sharing and data discoverability between hospital systems. Another increase in efficiency could be achieved by greater connectivity of a single patient's data, such as previous test results, PGS data, medical history, etc. Increasing patient data connectedness and accessibility for authorized users is not a trivial problem while also maintaining privacy. It is at this crossroads that the current state-of-the-art in clinical data storage technology is largely incapable of achieving both goals.

A Solid solution. 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 [8]. Within the current system, a citizen-centric model is difficult to implement due to technological challenges presented by centralized databases. The Solid protocol [9], 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, 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.

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.

2 State of the Art

Current clinical data storage. Most clinically relevant health data, including PGS data, is stored using an institution-centric approach characterized by

the hospital or hospital system isolating its stored data into one or more centralized or cloud-based databases that are governed and maintained solely by the owner institution or a contracted organization [19]. The predominant organization of these databases is a relational structure, although alternative non-relational methods like NoSQL or RDF have become slightly more popular in recent years [14]. Due to technological limitations imposed by relational database structure, maintaining stored data privacy results in strict data accessibility policies that severely restrict data sharing potential. With the enlarged threat of hacking, phishing, and login credential compromisation that is only increasing [4], health care institutions have taken steps to enact tighter restrictions on data access and increase cyber security budgets. Collectively, these state-of-the-art approaches to PGS data storage inhibit the scalability of PGS data usage in clinical practice due to data siloing and resulting high costs.

PGS data sharing in academic research. In academic research, the development of infrastructure that allows for sharing of genome data between institutions, creating federated centralized databases, can be observed in initiatives such as GA4GH Beacons [20]. 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.

Solid, RDF, and Linked Data in clinical practice. The Solid specifications have been shown to provide competent infrastructure for preserving the privacy of sensitive stored data [11]. Further, it has been shown that additional safeguards can be imposed using privacy policies represented in RDF triples, allowing more granular discoverability controls [7].

In recent years, there have been initiatives for representing biological data as RDF [21], specifically extending into clinical biology recently [24]. Past experiments have shown that general linked data integration into clinical practice results in improved outcomes [12]. There is also precedent for investigation into the representation of genomic information as RDF [18][24]. Solid pods could provide the infrastructure for such data storage strategies, while also preserving possibilities for non-linked data stored in the pod, such as test result files, to be linked to genomic data, improving data connectivity.

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 [23], 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.

3 Problem Statement and Contributions

My Ph.D. is situated to provide a proof-of-concept PGS data storage and querying framework for use in clinical practice. In this pursuit, my core research ques-

tion is: *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 advantages over the existing PGS data storage protocols in health care?*

To address this central question, I will investigate the more specific questions: A) Can the decentralized storage protocol Solid [9] offer suitable infrastructure for PGS data? B) Do the specifications provided by Solid provide for adequate control of PGS data privacy while also allowing for increased authorized sharability? C) Can querying over these sources be achieved through the use of modified LTQP algorithms? D) Can these features be combined into a cohesive web application and deployed together?

My hypothesis is that such a framework can be developed and would offer unique advantages over the existing state-of-the-art institution-centric PGS data storage solutions.

The following five objectives will be undertaken to test my hypotheses: (1) Solid Pod PGS Data Storage, (2) Genomic data as Linked Data, (3) Data policies, (4) Data querying, (5) Framework deployment.

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. Explicitly, an approach to how decentralized storage specifications can be applied to sensitive medical data storage, how genomic and medical data can be represented and queried as Linked Data, how existing Linked Data querying algorithms perform over genomic and health data, how granularized data policies impact querying and linking data in a medical context, and how the combination of these semantic technologies could provide an improvement over existing state-of-the-art clinical PGS data storage and usage strategies, are specific questions my Ph.D. aims to answer.

4 Research Methodology and Approach

My Ph.D. will be split into four component work packages (WPs) with a fifth work package where the components will be unified into a cohesive framework with an accompanying web application. This workflow is reflected in Fig. 1.

4.1 Work Package 1: Storing and publishing personal genomic data in a decentralized environment

I will test the viability of Solid data pods as storage infrastructure for patient PGS data, thus, testing my hypothesis that Solid can support PGS data storage.

A test dataset will be constructed using publicly available Illumina platinum genome files [3]. These files will be used as representative "patient" PGS data for experimentation.

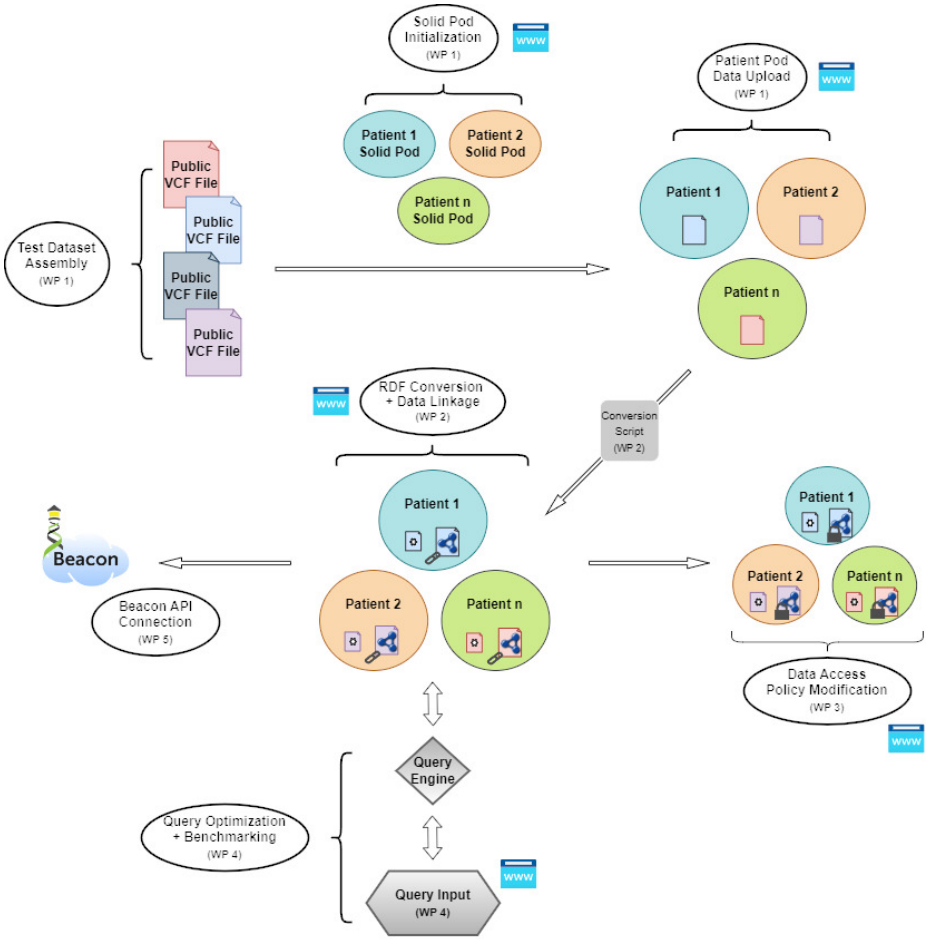

Fig. 1. PENGQUIN Ph.D. workflow. White circles represent milestones to be achieved in respective work packages (WP) that are denoted in parentheses. The small blue "WWW" box seen next to some objectives represents integration of that item into the framework's web application. WP 1 will achieve the assembly of a VCF file test dataset, the initialization of simulation Patient Solid Pods, and the upload of simulation patient VCF files into that patient's Solid pod. WP 2 will convert VCF file data into a serialized RDF representation and link parts of that VCF data to other simulated medical data. WP 3 will establish data access policies for patient Solid pods and specific data stored within them. WP 4 will experiment with querying approaches for discovering and accessing data stored in patient Solid pods through the use of user provided queries and the use of a query engine. WP 5 will establish the connection of the patient's Solid pod to the Beacon network.

I will also create server-hosted Solid pods using the Community Solid Server (CSS) implementation of Solid [1]. Each pod will be a storage container for a single individual's PGS data. I will upload a single VCF file, into one "patient's" pod to test basic functionality of a Solid pod for hosting large patient genomic data. 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 Solid pods for storing PGS data, which are much larger than the datasets used in past Solid experimentation. The main risk of storing PGS data in Solid data vaults is related to the size of PGS data which will be addressed through organizational large-scale storage access.

4.2 Work Package 2: Storing PGS data in RDF and as Linked Data

I hypothesize that the conversion from VCF to RDF is possible, and the resulting RDF representation will allow for linking of other medically relevant data within the patient's pod and outside of it. The conversion process will be made reversible to enable connection to existing clinical workflows that request VCF format.

To convert PGS data from VCF to RDF, we will investigate a format translation process using the SPHN RDF ontology [24]. During this translation process, we will experiment with different approaches, such as a bidirectional mapping index, for efficient reversal of conversion.

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. Linkages will be added through triple insertions at different relevant points in the serialized VCF RDF file stored in the patient's Solid Pod. All data linkages will be made using files and triples that contain simulated data or publicly available data.

While Linked Data is state-of-the-art, these concepts have not yet been applied to clinical genomic data. The power of linking the VCF data to other clinically relevant data will be especially realized when these semantic links are discovered during querying, which will be investigated in WP5.

The main risk of converting PGS data to Linked Data using RDF is that this conversion requires an ontology. The conversion process will utilize the publicly available SPHN RDF ontology [24] and if this ontology is insufficient, I will work with members of the IDLab at UGhent with experience in ontology definition to make changes or additions where necessary.

4.3 Work Package 3: PGS data privacy policies

In this work package, 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. I hypothesize that various levels of authorization can be implemented and provide protections for maintaining the privacy of PGS data stored in Solid pods.

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 and filled to represent a real clinical PGS workflow.

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. The described access schema has not been attempted in the presented level of detail for clinical genomic data. If the above proposed schema for privacy policies cannot be achieved, a simpler and more generalized schema will be devised and implemented. Privacy is a nuanced subject, especially in terms of governance and here, I aim to show the possibilities presented by Solid in this framework, not dictate suggestions for its deployable implementations.

4.4 Work Package 4: Querying over PGS data in one and many pods

This work package will establish a querying mechanism for patient pod data that incorporates 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 execute queries across PGS data contained in patient pod(s) through the use of the query language SPARQL [5]. 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 [2], to perform the queries apart from the data stores.

For PGS data querying, I will benchmark and build upon the link traversal query processing (LTQP) paradigm [23], which has been shown to be an effective method for querying Linked Data between Solid Pods. A significant challenge is presented by the storage landscape of a large number of large data sources. I will look to innovate established LTQP algorithmic approaches by integrating 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.

I aim to adapt this querying approach to the specific domain of genomic and health data which has not been attempted before. There is a risk that I cannot devise solutions to incorporating indexes in LTQP algorithms and/or these algorithm modifications do not improve performance enough to be usable. If these challenges are encountered, I will first investigate imposing limits on query complexity and reducing the number of data vaults that are included in the possible query space, then, if still unsuccessful, look to investigate the implementation of algorithms utilized by centralized SPARQL endpoints that are known to be able to query large sets of data.

4.5 Work Package 5: Component consolidation and framework deployment

To improve data flows for research purposes, I intend to connect the proposed framework to the international Beacon initiative [20]. In this aim, I will investigate the establishment of 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. Any issues within this aim will be addressed through collaboration with creators of the Beacon initiative.

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.

5 Evaluation Plan

WP1. All tasks within WP1, including test dataset assembly, Solid pod creation and hosting using the CSS, and test data uploading to Solid pods will be evaluated only for functionality.

WP2. Direct conversion between VCF and RDF will be evaluated in terms of computational overhead, conversion time, and memory usage, both in the Solid pod and during conversion. The same evaluations will also be performed on the process when an intermediate mapping index file is used. Comparisons will be documented in a formal benchmarking study. Functionality of data linkage aims will be assessed by data querying in WP4.

WP3 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.

WP4 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.

WP5 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.

6 Preliminary Results

During the first months of my Ph.D., I have been composing a scoping review paper on the current landscape of clinical genomic data sharing. I plan to submit the review paper for publication in a peer-reviewed journal in the coming months. Within WP1, I have successfully assembled the test dataset and set up CSS pod instances. I have also successfully uploaded VCF files into these pods signalling completion of WP1.

7 Conclusions/Lessons Learned

With the emergence of patient genomic data as a tool for clinicians, establishing the infrastructure for patient genomic data sharing that maintains patient data privacy is an economic niche that is largely unfilled. My Ph.D. framework is poised to provide the outline of necessary technological implementation considerations while hopefully contributing to future product development. More generally, my Ph.D. is designed to assess the potential in applying cutting-edge semantic web research to modern clinical challenges with the hope of gaining insight into the efficacy of these applied implementations while identifying future research directions.

Acknowledgments. I acknowledge my Ph.D. promoters Ruben Taelman[1], Bart Buelens[2], Gokhan Ertaylan[2], and Ruben Verborgh[1] for their help and guidence. Funding provided from VITO NV (`UG_PhD_2303_contract`).

References

1. CommunitySolidServer/CommunitySolidServer: An open and modular implementation of the solid specifications, https://github.com/CommunitySolidServer/CommunitySolidServer
2. Comunica – a knowledge graph querying framework, https://comunica.dev/
3. Platinum genomes, https://emea.illumina.com/platinumgenomes.html
4. Ransomware attacks on hospitals have changed | cybersecurity | center | AHA, https://www.aha.org/center/cybersecurity-and-risk-advisory-services/ransomware-attacks-hospitals-have-changed
5. SPARQL 1.1 query language, https://www.w3.org/TR/sparql11-query/
6. Bagger, F.O., Borgwardt, L., Jespersen, A.S., Hansen, A.R., Bertelsen, B., Kodama, M., Nielsen, F.C.: Whole genome sequencing in clinical practice **17**(1), 39. https://doi.org/10.1186/s12920-024-01795-w
7. Benaribi, F.I., Malki, M., Faraoun, K.M., Ouchani, S.: A SPARQL-based framework to preserve privacy of sensitive data on the semantic web **17**(3), 183–199. https://doi.org/10.1007/s11761-023-00368-6
8. Brands, M.R., Gouw, S.C., Beestrum, M., Cronin, R.M., Fijnvandraat, K., Badawy, S.M.: Patient-centered digital health records and their effects on health outcomes: Systematic review **24**(12), e43086. https://doi.org/10.2196/43086

9. Capadisli, S., Berners-Lee, T., Verborgh, R., Kjernsmo, K.: Solid protocol, https://solidproject.org/TR/protocol
10. Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., McVean, G., Durbin, R., 1000 Genomes Project Analysis Group: The variant call format and VCFtools **27**(15), 2156–2158. https://doi.org/10.1093/bioinformatics/btr330
11. Esposito, C., Hartig, O., Horne, R., Sun, C.: Assessing the solid protocol in relation to security & privacy obligations, http://arxiv.org/abs/2210.08270
12. Farinelli, F., Barcellos de Almeida, M., Linhares de Souza, Y.: Linked health data: how linked data can help provide better health decisions **216**, 1122
13. Ko, Y.K., Gim, J.A.: New drug development and clinical trial design by applying genomic information management **14**(8), 1539. https://doi.org/10.3390/pharmaceutics14081539
14. Kotsilieris, T.: An efficient agent based data management method of NoSQL environments for health care applications **9**(3), 322. https://doi.org/10.3390/healthcare9030322
15. Marwaha, S., Knowles, J.W., Ashley, E.A.: A guide for the diagnosis of rare and undiagnosed disease: beyond the exome **14**(1), 23. https://doi.org/10.1186/s13073-022-01026-w
16. Mathur, S., Sutton, J.: Personalized medicine could transform healthcare **7**(1), 3–5. https://doi.org/10.3892/br.2017.922
17. McLeod, H.L.: Cancer pharmacogenomics: Early promise, but concerted effort needed **339**(6127), 1563–1566. https://doi.org/10.1126/science.1234139
18. Prasanna, S., Rao, D., Simoes, E., Rao, P.: Scalable knowledge graph construction and inference on human genome variants. https://doi.org/10.48550/arXiv.2312.04423
19. Quantin, C., Jaquet-Chiffelle, D.O., Coatrieux, G., Benzenine, E., Auverlot, B., Allaert, F.A.: Medical record: systematic centralization versus secure on demand aggregation **11**, 18. https://doi.org/10.1186/1472-6947-11-18
20. Rambla, J., Baudis, M., Ariosa, R., Beck, T., Fromont, L.A., Navarro, A., Paloots, R., Rueda, M., Saunders, G., Singh, B., Spalding, J.D., Törnroos, J., Vasallo, C., Veal, C.D., Brookes, A.J.: Beacon v2 and beacon networks: A "lingua franca" for federated data discovery in biomedical genomics, and beyond **43**(6), 791–799. https://doi.org/10.1002/humu.24369
21. SIB Swiss Institute of Bioinformatics RDF Group Members: The SIB swiss institute of bioinformatics semantic web of data **52**, D44–D51. https://doi.org/10.1093/nar/gkad902
22. Stefanicka-Wojtas, D., Kurpas, D.: Barriers and facilitators to the implementation of personalised medicine across europe **13**(2), 203. https://doi.org/10.3390/jpm13020203
23. Taelman, R., Verborgh, R.: Evaluation of link traversal query execution over decentralized environments with structural assumptions . https://doi.org/10.48550/ARXIV.2302.06933
24. Van Der Horst, E., Unni, D., Kopmels, F., Armida, J., Touré, V., Franke, W., Crameri, K., Cirillo, E., Österle, S.: Bridging clinical and genomic knowledge: An extension of the SPHN RDF schema for seamless integration and FAIRification of omics data. https://doi.org/10.20944/preprints202312.0373.v1