PhD Fellowship strategic basic research

Elias Crum - 1S27825N

Table Of Contents

PhD Fellowship strategic basic research

GENERAL

Enter the English title of your research proposal.

PErsoNal Genome QUery IN clinical practice (PENGQUIN)

53 / 240

Enter the Dutch title of your research proposal.

PErsooNlijk Genoom beVRaging IN klinische praktijken (PENGVIN)

62 / 240

Complete the abstract of your research proposal - English version.

Medical care is becoming increasingly personalized through the use of patient genetic information. At present, data useful for clinical care, including genomic data, are commonly diffuse, organized arbitrarily, and confined within in data silos.
Coupled with stringent privacy regulations, high storage costs, and severely limited data sharing infrastructure, the scalability of personalized clinical strategies is considerably constrained.

My proposed Ph.D. aims to improve the connectivity and shareability of genomic data storage(s), while preserving data privacy, to decrease the costs of patient genome data use in clinical practice. I propose a citizen-centric storage framework that integrates insights from various semantic web domains research into a novel, practical solution for clinical use.

Specifically, I will (a) store patient geomic data using Solid pods, (b) represent personal genome sequence data using RDF as Linked Data, (c) apply policies to stored data, and (d) optimize link traversal queries for data retrieval using indexing methods.

Through this research, I aim to provide a comprehensive framework, complemented by a user-friendly web application, to facilitate practical usage of my proposed framework in clinical settings. Ultimately, this PhD seeks to provide the proof-of-concept for an innovative, streamlined product for clinical genome data storage, paving the way for scalable personalized medicine.

1442 / 1500

Complete the abstract of your research proposal - Dutch version.

Medische zorg wordt steeds persoonlijker door het gebruik van genetische informatie van patiënten. Op dit moment zijn klinische gegevens, waaronder genomische data, vaak verspreid, willekeurig georganiseerd en opgeslagen in datasilo's. Samen met strenge privacyregels, hoge opslagkosten en beperkte infrastructuur voor gegevensdeling, beperkt de schaalbaarheid van gepersonaliseerde klinische strategieën.
Mijn voorgestelde doctoraat streeft ernaar de deelbaarheid van de opslag van genomische gegevens te verbeteren, terwijl de gegevensprivacy behouden blijft, om de kosten van het gebruik van patiëntgenoomgegevens in de klinische praktijk te verlagen. Dit door de implementatie van een op burgers gericht opslagkader dat inzichten uit verschillende onderzoeksgebieden van het semantische web integreert tot een praktische oplossing voor klinisch gebruik.
Concreet zal ik (a) patiëntgenomische gegevens opslaan via Solid-pods, (b) persoonlijke genoomsequentiegegevens voorstellen met behulp van RDF als gelinkte gegevens, (c) beleid toepassen op opgeslagen gegevens, en (d) innovatieve link-traversiequeries ontwikkelen voor gegevensherstel.
Met dit onderzoek beoog ik een allesomvattend kader te bieden, aangevuld met een gebruiksvriendelijke webtoepassing, om praktische acceptatie in klinische settings te vergemakkelijken. Uiteindelijk streeft dit doctoraat naar een proof-of-concept voor de opslag van klinische genomische gegevens, en de weg vrij te maken voor gepersonaliseerde geneeskunde.

1499 / 1500

Select up to five scientific disciplines that best characterize the proposed research.

Enter up to three English free-text keywords or concepts that best characterize the proposed research.

These keywords allow reviewers to quickly understand the broad scope of your proposal.

Solid

Federated Querying

Data Privacy

Enter up to three Dutch free-text keywords or concepts that best characterize the proposed research.

These keywords allow reviewers to quickly understand the broad scope of your proposal.

Solid

Gefedereerd bevragen

Gegevensprivacy

Position your proposal in terms of economic finality.

Ultimately (medium- to long-term), the proposed strategic research may lead to added value for one or more specific company(ies), or for a sector or group of enterprises. The application potential may as well be expressed in terms of socio-economic benefits, related to the Flemish transition areas and priorities in science, technology and innovation. You can highlight multiple options simultaneously but you need to select at least one. In the first two cases you have to specify which companies or sectors are targeted. Furthermore, you can select up to 2 transition areas, each with an associated priority. It is also possible to choose two priorities under the same transition area. What you mention in this section should be referred to, elaborated and explained in the Project description, section 'strategic dimension': Hence, do not just drop company names here, and be specific in referring to sectors (be more precise than e.g. 'manufacturing' or 'space industry').

Companies ***(optional)***

My project is nested with the larger goals of VITO NV's Digital Precision Health group to develop tools and technologies that encourage increasingly citizen-centric health care.

177 / 240

Sectors ***(optional)***

My project targets the interation between heath data management, an individual's electronic health record, and the growing network of clinical and commercial applications for which these data will act as inputs.

211 / 240

Tick off the transition areas and their science, technology and innovation priorities. ***(optional)***

Transition Area Group	Transition Area
Health & Well-Being	Central electronic health record
Digital society	Next generation networks

PERSONAL DATA

Explain any career breaks.

Explain possible gaps in your CV in the input field below. Make sure your current position and previous appointments are well listed in the e-portal 'Personal details' section ('Posts / Career'). If you have interrupted your academic career at any given point for at least three months (maternity leave, parental leave, full-time sickness leave, unconventional career paths such as leave because of activities in industry or other non-academic sectors, …) provide details about this below (reason, start/end date). This will allow the reviewers to fairly assess your career stage.

My career as a student and researcher was briefly interrupted after the completion of my Master's program in fall 2022, before I began my Ph.D. at UGent in fall 2023.
During this year I was not formally enrolled in an academic program, but I was not really away from academia.
During this time I was employed as an adjunct professor of biology at Loyola University Chicago in Chicago, USA.
For a half of that year, I was also employed part time as a medical scribe in the emergency room of a university hospital in Chicago.
In this year I was also finishing off some research project loose ends that were not fully completed for my Master's thesis defence, including a primary author peer-reviewed journal publication.
While I was sufficiently busy during my year off, I also took significant time to reflect and subsequently made some influential decisions about my future.
The most crucial being my decision to pursue a Ph.D. instead of medical school as well as my decision that my Ph.D. candidacy would be in Flanders at Ghent University.

1045 / 3000

STUDY RESULTS (ACADEMIC EDUCATION)

This section will be used by the evaluators to assess your potential as a PhD researcher, based on your past academic trajectory.

Study narrative.

Show how your academic study trajectory has formed the ideal preparation for doing a PhD, in general and specifically on the topic of the proposed project. Where appropriate, refer to your grades of relevant courses, percentiles or relative ranking or other study results. You may also highlight specific programs or courses you took. If applicable, include additional information on your personal situation where you believe this may have affected your study results and should need to be taken into consideration during the evaluation.

My academic career has been both highly challenging and rewarding.
As is exemplified by my bachelor's and master's academic record, as well as my graduation with honors from my Master's program, I am a highly capable student, adept at excelling in an academically challenging environment, and keen to learn.
The courses that I have followed over my journey toward a Ph.D. have taught me skills, knowledge, and strategies to succeed at such an endeavor.
Specifically, within biology, I have excelled in Genomics, Molecular Biology, and Proteomics courses, receiving a perfect score in all of them, which has provided me with the conceptual and scientific foundations for pursuing the genomic components of my Ph.D.
Within bioinformatics, I followed courses such as computation biology, quantitative bioinformatics, and bioinformatics, completing all with a perfect score, where I learned about the computational, statistical, and data science aspects of quantitative biological research.
For foundational skills and knowledge in computer science, classes such as data structures, discrete structures, computational biology, and object-oriented programming, taught me about the architecture, concepts, and strategies leveraged in computational science research.
These courses also taught me coding proficiency in Python, Java, SQL, Bash, as well as other crucial software development fundamentals.
In my master's coursework, I followed classes such as responsible conduct for research, bioinformatics research design, and scientific writing which helped prepare me for the intangible aspects of Ph.D. research crucial to researcher success.
For my proposed Ph.D., my academic excellence in biology/bioinformatics, computer science, and research logistics courses has well prepared me for success while also providing me with experience and confidence that I can problem solve, learn challenging new knowledge domains, and communicate my findings with fellow researchers and scientific lay people alike.

2005 / 3000

Relative positioning of your study results.

*Provide the following information for the master's degree on the basis of which the application is submitted (see PhD fellowship regulations article 7): the **overall result** you obtained for this **master's** degree**, expressed as a percentage;** your relative ranking within your study group expressed as the **percentile** (referring to your university study group) or rank.*

*• If you have not yet obtained a master's diploma, please enter the study results and percentile related to the relevant **bachelor's** diploma.*

*• Regarding diplomas from **non-Flemish universities**, either a percentile score (if available), or at least your rank within your study group should be provided. If neither of these data is available, use the text field at the bottom to provide qualitative information on all your study results.*

• Master-after-Master diplomas are not taken into account for percentile/rank information (but may be discussed above in the 'study narrative' section).

• More information on providing relative ranking information can be found on the programme webpages.

***Please select** the relevant diploma for percentile/rank information.*

Date	University	Degree	Grade	Country
13/08/2022	Loyola University Chicago	Bioinformatics	MSc	US

Upload the declaration on your percentile or rank within study group.

Note that this document is mandatory and an essential part of your application. However, exceptionally and when duly justified this document can be submitted within reasonable time after the submission deadline.

Attachment [Upload the declaration on your percentile or rank within study group.] [signed_class_rank-combined.pdf] has been added below in the report.

Enter the study results of your diploma.

Enter the global percentage -up to 2 decimal places e.g. xxx,xx

100,00

Provide details about the positioning of your grade based on the percentile or study group ranking.

Ranking study group

Ranking within the study group.

1

Number of students within study group.

13

Text field to provide additional information on your study results (the global percentage, percentile, rank in study group). *(optional)*

If you were not able to provide a global percentage and/or positioning in the study group, you can use this text field to present in a qualitative way the relative positioning of all your study results compared to your peers. You can also use this text field to provide additional information on your academic study

results (Bachelor, Master, Advanced Master, ...), i.e. detailed course scores can be added or, if you have not yet obtained your master, you can marks obtained in the first master year All evidence on study results should be uploaded in 'Personal Details/studies' section.

My positioning in my Masters class is reflected above.
My Bachelor's graduating class did not record class rankings but there were 2,934 bachelor's level graduates and I recieved a 3.991/4.0 cumulative bachelor's GPA. Thus, I finished with 99.775% in my bachelor's program. For comparability, a minimum of a 2.0 GPA is needed to graduate and is considered a passing grade.
Also to note, I was admitted to the Alpha Sigma Nu honors society, from which only the top 10% each academic class is invited to apply.

508 / 2000

MOTIVATION AND COMPETENCES

This section will be used by the evaluators to assess your potential as a PhD researcher, based on your motivation, acquired scientific competences and scientific mindset.

Write a motivation statement.

Elaborate on your motivation and research interests to pursue an individual PhD trajectory. Elaborate also on how your scientific background and competences will allow you to start the PhD project, and to grow into a strategically thinking and innovation- oriented expert. Provide a clear and substantiated overview on the skills you have already developed, and on the competences yet to be acquired and how you will acquire them.

Curiosity is a cornerstone of my identity.
As a student, I quickly fell in love with learning as a way to exercise my innate curiosity.
Entering my bachelor studies, I was determined to become a physician because of my love of biology and personal history with serious health issues as a child.
As a bachelor's student, I chose to pursue an interdisciplinary field that moved me towards my goals of becoming a physician while also allowing me to learn about a field I knew nothing about -- computer science.
Quickly I fell in love with genomics, specifically the complexity presented by genomics data, and the computational heuristics developed to learn about and leverage such data.
Gradually, I shifted my goals from medicine to pursuing computational medically motivated research.
During my master's thesis project, I developed a strong foundation of research skills and genomics domain knowledge.
Specifically, I because well acquainted with digital human genomics data formats and uses, compression, indexing, and parsing strategies for these complex and large data, and computational competencies in data processing, coding, and algorithm application.

Ultimately, my curiosity about data representation and data parsing optimization that began during my master's project drove me to begin my Ph.D. at Ghent University studying how representing genomic data using semantic web and decentralized storage technologies could improve clinical data flows.
I am new to the field of semantic web research, as well as the many sub-domains that focus on aspects such as data policies and governance, semantic representations of data, and Web technologies.

I also bring with me expertise in genomics data structure, parsing strategies, and algorithmic approaches to biological problem solving.
As is exemplified by my graduation with honors from my Master's program, I am a highly capable student, adept at excelling in an intellectually challenging research environment, and keen to learn.
I will fill the gaps of my current lack of formal knowledge of semantic web technologies and concepts through mentoring and interactions with members of my research group, the Knowledge on the Web Scale (KNoWS) group within IDLab, at Ghent University, as this is a domain of research at which they excel.
I am also applying to attend a semantic web summer school (ISWS) this coming summer, looking to attend conference tutorials, as well as potentially following select masters level university courses that can provide me specific domain-specific knowledge.

2551 / 3000

Scientific activities, experiences and achievements.

In this input field you can further elaborate on first steps as a (potential) innovation-oriented scientist. List relevant activities, experiences and achievements that may be relevant for assessing your potential to start a PhD. For mobility and awards, other dedicated input fields are available below.

- *For (ongoing or finished) **master thesis** or equivalent (as well as dissertation advanced master): mention title, promotor, research group and host institution. If the thesis is related to your PhD topic, also specify initial objective, methodology used and (intermediate) results.*
- *For (PhD) **research** already started, specify initial objective, methodology used and (intermediate) results. If applicable mention (up to 5) **publications and other achievements**. Mind, do mention for each achievement item (publications and other achievements) **your share** and its nature, and those of other significant partners in the workload.*
  For publications: list all authors, title of publication and journal name (without abbreviations) with volume, start/end page and year. Mention whether the publication was peer reviewed or not. For book publications, give all necessary bibliographic information (author(s) or editor(s), book title, publisher, place, year, number of pages).
  Make sure your complete publication list is up to date in the e-portal 'Personal details' section ("Publications").
  For other achievements: provide a short description, when it was undertaken and finalised and list all the relevant participants involved in it.
- *List any other distinct **research output** that does not fit in the bibliographic publication list and that is meaningful in a broad sense with respect to this fellowship application. It may be constituted by a data base, surveys, a technical diagram, software, objects (maquettes, prototypes…), any other type of activity or output you consider to be relevant. Date the output where appropriate.*
- *Mention any relevant, past or concretely planned, experiences (internships, presentations, collaborations, …)*

I have successfully defended a master's thesis in Bioinformatics. Project Title: CATALOGING AND ENGINEERING TEMPERATE COLIPHAGES OF THE HUMAN
URINARY MICROBIOME; Promoter: Dr. Catherine Putonti; Putonti Lab research group; Loyola University Chicago.

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.
I have presented a poster at the Semantic Web Applications and Tools 4 Health Care and Life Sciences 2024 conference that proposes the use of Solid data vaults for storing genomic and health data.
I composed the accepted 2-page poster abstract, produced the physical poster, and presented the poster at the conference.
I was the contributing author for a consortium poster also presented at the Semantic Web Applications and Tools 4 Health Care and Life Sciences 2024 conference. I helped provide input for the 2-page poster abstract, helped contribute to the composition of the physical poster, and was one of the consortium members that presented the poster at the conference.
Within WP1, I have successfully assembled the test dataset and set up CSS pod instances.
I have also successfully upload VCF files into these pods signalling completion of WP1.

I was the primary author of the publication titled: Coliphages of the human urinary microbiota
Authors: Elias Crum, Zubia Merchant, Adriana Ene, Taylor Miller-Ensminger, Genevieve Johnson, Alan J Wolfe, Catherine Putonti
Journal: Plos one; Volume 18; Issue 4; Pages e0283930
Peer Reviewed: yes
Date Published: 2023/4/13

I was the contributing author of the publication titled: Genome Investigation of Urinary Gardnerella Strains and Their Relationship to Isolates of the Vaginal Microbiota.
Authors: Catherine Putonti, Krystal Thomas-White, Elias Crum, Evann E Hilt, Travis K Price, Alan J Wolfe
Journal: Msphere; Volume 6; Issue 3; Pages e00154-21
Peer Reviewed: yes
Date Published: 2021/6/30

I have presented two posters at conferences during my research stay at Loyola University Chicago, both at the St. Albert Day intra-university research symposium held annually for biological research performed by undergraduate, graduate, and medical students.
These poster were not published due to the inter-university aspect of research.

2384 / 3000

Specify earlier mobility (research stays) in other organizations.

Indicate the research stays which have already been undertaken, prior to this project. If applicable, motivate shortly the added value of each stay to the project. Include details on the organization, type of organization, country, contact person, start/end date, function/activities.

Research stay 1: Stritch School of Medicine, Wolfe Lab, Summer 2019
Country: USA; Contact Person: Dr. Michael J. Wolfe; Start May 2019 / End August 2019; Research assistant

This project was both microbiological and computational in nature.
I became familiar with the methods, practices, and procedures of scientific researchers as well as distinct data collection and reporting techniques.
I learned how to communicate my work with colleagues, extract meaning from data, and engage with colleagues to problem solve and provide constructive criticism in weekly meetings with researchers and clinicians.
I was challenged by learning to fully understand the complex scientific underpinnings of my project and then by communicating those concepts effectively to diverse audiences.
For a final presentation, I gave a 30 minute presentation to around 80 principal investigators, post-doctoral researchers, and graduate students.

Research stay 2: Loyola University Chicago, Putonti Lab, Fall 2019-Spring 2023
Country: USA; Contact Person: Dr. Catherine Putonti; Start Sept 2019 / End April 2023; Independent researcher.

As a student researcher in Dr. Putonti's lab at Loyola, I perform independent work focusing on bacteriophage and bacteria of the urinary microbiome at both bachelor and master levels.
My research involved both computational analysis of bacterial and viral genomes as well as molecular biologically-based lab techniques.
Specifically, I worked to determine the genomic contributions to a host range shift in the 3 bacteriophage, worked to characterized bacterial and viral phylogenic diversity based on genomic and genetic similarity analyses, and engineered bacteriophages to selectively kill specific strains of bacteria.
Actively performing research has taught me that very few experiments will work as planned -- problem-solving, flexibility, and persistence are prerequisites to discovery and success, and when planning, organization and details are crucial.

1989 / 2000

Specify concrete mobility plans (research stays) within the FWO fellowship.

‚Indicate the research stays which are planned within the FWO fellowship. Motivate shortly the added value of each stay for the project. Include details on the organization, type of organization, country, contact person, start/end date, function/activities. See Programme Regulations Art. 4 §2

I do not currently plan to pursue any research stays at external institutions during the duration of my Ph.D.

109 / 2000

List any scientific awards.

List prizes and awards, (e.g. best master thesis…). Specify the awarding body, title, date, amount and theme.

I was awarded a Mulcahy research scholarship by Loyola University Chicago for the 2020-2021 school year in November 2021 which was for the amount of $3000.

I was awarded an institutional scholarship by Loyola University Chicago for my Master's program (2021-2022 academic year) for a total of $20000 in spring 2020.

I was inducted to the Jesuit honors society Alpha Sigma Nu in October 2019 based on distinction in scholarship, service, and faith.

451 / 600

HOST INSTITUTION - PROMOTOR

This part of the application form provides info on host institutions and (co-)promotors of your research. There are 3 levels where data can be filled in.

1. ***As a FWO PhD researcher, you must be affiliated to a main Flemish host institution*. You must refer to a (main) promotor in this institution.***

   ** Eligible main host institutions are: Universities in the Flemish Community.*

   Select a main Flemish host institution (Art. 4§1 of the FWO regulations) from the pick list, and name a main promotor. The main promotor will be invited by FWO to submit a recommendation letter. Co-promotors will receive a notification by FWO.

   (Optional) You can name a co-promotor, affiliated to the same main host institution.

2. ***(Optional) In case of a collaboration with a Flemish or Federal scientific institution, where the research is carried out, (Regulations Art 4§1), the co-hosting organization and co-promotor should be named. It should be mentioned on level 2.***

   Select an organization from the pick list, and name a co-promotor. If needed you can name another co-promotor affiliated to this organization.*

   ** If the organization is not mentioned on the pick list, select 'other' and name the organization FWO will consider whether this organization fulfills the requirements to act as a co-hosting institute.*

3. ***(Optional) In case another co-promotor oversees your PhD project. Mention the organization he/she is affiliated to, and the corresponding co-promotor. It should be mentioned on level 3.***

1. Main Flemish host institution

Main Flemish host institution

Ghent University (UGent)

Promotor

Eligibility main promotor: check Art. 10§2 of the regulations

The (main) promotor will be invited by FWO to submit a recommendation statement on the PhD fellowship application.

In case of collaboration with other research units in the same or other host organizations, co-promotors should be mentioned. These will receive a notification by FWO. They will not be invited to submit a recommendation statement.

Title	*anonymized*
First name	Ruben
Last name	Verborgh
Date of birth ***(optional)***	*anonymized*
Current occupation	Professor
Employment rate	*anonymized*
Email	*anonymized*
Research unit	IDLab
Street and number	*anonymized*
City	*anonymized*

Co-promotor(s) ***(optional)***

You may specify one or more co-promotors.

Title	*anonymized*
First name	Ruben
Last name	Taelman
Date of birth ***(optional)***	*anonymized*
Current occupation	Post Doctoral Researcher
Employment rate	*anonymized*
Email	*anonymized*
Research unit	IDLab
Street and number	*anonymized*
City	*anonymized*

2. Other host institution(s) – Flemish or federal

If you will carry out your research in another host institution (Flemish or federal) according to Art 4 §1 of the regulations, please click "Add" to select an institution in the drop-down menu. If the institution is not mentioned in the picklist, select 'Other' and name the organization. FWO will consider whether this organization fulfills the requirements to act as a co-hosting institute.

Other Flemish- or federal host institution

Flemish Institute for Technological Research (VITO)

Co-promotor(s)

Title	*anonymized*
First name	Gökhan
Last name	Ertaylan
Date of birth ***(optional)***	*anonymized*
Current occupation	Research Lead - Research and Development at VITO NV
Employment rate	*anonymized*
Email	*anonymized*
Research unit	Digital Precision Health
Street and number	*anonymized*
City	*anonymized*

Title	*anonymized*
First name	Bart
Last name	Buelens
Date of birth ***(optional)***	*anonymized*
Current occupation	Head of Data Science at VITO
Employment rate	*anonymized*
Email	*anonymized*
Research unit	Data Science
Street and number	*anonymized*
City	*anonymized*

3. Other organization(s)

PROJECT

Project description.

The project description should be structured following the template provided by FWO. The sequence of the different topics should be followed exactly as provided in the original template. The total project outline has a maximum of 12 A4 pages (Font Calibri 11, single line spacing, original template margins ...) herein included all tables, graphs, illustrations, etc.

Attachment [Project description.] [FWO_Project_Description_EDC.pdf] has been added below in the report.

OTHER FUNDING

Have the content of this proposal and at least the main part of the proposed research actions, be it with literally the same text or in a varied form, already been submitted before AND was it funded or is the funding decision still pending (applications that finally did not result in funding should not be mentioned)?

No

Enter any additional remarks and the decision date(s) of pending funding decision(s) mentioned above.

- *You are encouraged to use this field as an opportunity to point out potential overlap, complementarity, added value of current funding applied for or already obtained, ... related to the applications mentioned above.*
- *There can be good reason for applying or already having applied for funding at FWO or elsewhere. It is however important that the panel understands how pending applications for funding or obtained funding mentioned above relate to the current application.*

State 'NA' if not applicable.

NA

2 / 1000

PROJECT POSITIONING AND EMBEDDING

Explain how this project fits into the research activities of the involved host institution(s).

Elaborate on the positioning and embedding of your project in the research group(s), its scientific as well as strategic ambitions. If applicable, also position your own previous and current research to the proposed PhD fellowship project.

The IDLab, and specifically the KNoWS group at Ghent University, works in multiple semantic web research domains that can be integrated into decentralized storage technologies such as Solid. These fields include data privacy and governance, decentralized federated querying, and big data semantic representations.
My project touches on each of these domains and aims to integrate state-of-the-art solutions from each into a collective framework.
Further, my Ph.D. explores situations currently unexplored but promising for future research within our group, especially related to large data storage in Solid pods and querying of that data.
Within UGent's SolidLab project, my Ph.D. will also provide a practical implementation of Solid technology to a biomedical knowledge domain that allows for novel boundary exploration.

The Digital Precision Health group at VITO aims to develop and implement products that improve the scalability of personalized medicine.
My project is embedded in this mission by aiming to demonstrate how a decentralized storage structure on which other applications, tools, and software can be built and integrated.
In this aim, I will also combine state-of-the-art semantic web technologies into a user-friendly application that allows for non-experts, like physicians or patients, to utilize the advantages of these

technologies.
In this way, my project will demonstrate a path toward product production that encourages future clinical implementation.

1479 / 2000

Position the project in a national and international context.

Mention specific research collaborations planned in the course of this project, if appropriate, mention larger projects, programmes or networks your proposal may be part of.

Solid data pods offer discoverable, privacy controlled interoperable storage for RDF data.
Because of this property, a collaborative project with the Swiss Institute of Bioinformatics and Institute of Organic Chemistry and Biochemistry of the Czech Academy of Sciences has been created to explore their potential as data stores for sensitive research data that can be queried in a federated manner along with existing large, public SPARQL endpoints to answer research questions.
This consortium has been awarded a CHIST-ERA Call ORD research grant for the years of 2023-2025.

My project is also designed to fit within the European Virtual Human Twin (EDITH) initiative.
I will be working to design a platform that could be integrated within and/or provide inspiration for storage and sharing strategies within the EDITH initiative.

My project, in representing genomic data as RDF, is also dedicated to encouraging greater data interoperability embodied by the European FAIR data initiative.

I aim to integrate my framework to connect to or align with Belgian Genome Biobank project and system architecture that is planned to be implemented in 2026.

1156 / 1200

Did you take the issues of gender/sex and diversity into account while designing your research plan (e.g. selection of human participants and/or animals in experiments, relevance of research questions and/or results with respect to gender differences, …)?

This issue will be taken into account during the evaluation as part of your research methodology and work plan.

Not applicable

Justification.

NA

2 / 1200

Did you or will you work with societal actors other than research partners in the whole or parts of the research process (from design of the application up to the execution of the research)?

'Societal actors' consist of all kinds of groups in society (like patients and/or their organizations, other citizens, firms, …) involved in or connected to the research in one way or another. There is no limitation to what kind of partners in society possibly can be included, nor is involving societal partners an obligation: whether such an involvement could be relevant or not is left to the judgment of the applicants of the research proposal. Take into account, however, that the evaluators may find that collaboration with societal actors is recommendable or even necessary; you may anticipate this by clarifying your position in the designated textbox. Please be aware that this question on societal actors does not concern science communication or valorization.

This issue will be taken into account during the evaluation as part of your research methodology and work plan.

Not applicable

Justification.

NA

SCIENCE COMMUNICATION

Indicate how the results of the proposed research will be communicated to a non-expert audience.

FWO encourages its fellows to disseminate the results of their research widely and valorise them where possible.

My project is designed to produce a proof-of-concept framework that can be used by non-experts. This may be the most attractive aspect of this project, in that it is designed to be for non-expert audiences.
Thus, my largest communication goal will be implementation of the framework, making that implementation findable and usable over normal internet browsers, and demonstration of how the web application can be used.
This will hopefully be possible for both those within and outside of the scientific community through internet availability.
It should be said that all usable implementations of my project will only contain publicly available genome data, example patient data that only representational, and no way to actually input or discover real personal patient data.

I will also host a personal website that communicates intermediate results, methodologies, and specific implementations.
I plan to do this through regular blog posts, video and written tutorials, and github-hosted software advertisements.
Interaction with more general lay audiences will be attempted through X and Mastadon.

I will devote the most time and effort to blog posts and their advertisement.

1186 / 1200

PEER REVIEW

INTERNAL PEER REVIEW

There are 24 thematic SB-panels. More info on these panels and their specific scopes can be found here. You should select the panel that fits best with your research project, in terms of research methodology (rather than the application field).

Specify the expert panel.

SBWT5B - Informatics and data communication

Motivate your choice of expert panel.

Carefully read the scientific scope of the selected expert panel and motivate why your application fits the scope of this panel - i.e. why this panel has the most appropriate expertise to evaluate your proposal.

My project can be described as the unification of cutting-edge semantic web technologies into a functional framework that provides the digital infrastructure for improvement of data flows in a biomedical knowledge domain.
Thus, my project can be described as applied web and information system engineering covered by the SBWT5B - Informatics and data communication panel. This project is strategic in nature because of its goal to improve the state-of-the-art in clinical genomics data handling.

498 / 500

ETHICS

FWO Ethics Table

*The table below lists questions about possible ethical aspects in research proposals. Please go through the main table and tick 'YES' for aspect(s) relevant to your proposal. Then **answer any related sub-questions by clicking on the appropriate ethical topic** that becomes listed under 'Ethical Issues'. You can return to the main table by*

clicking on 'Ethical issues'.

If you mark a 'yes' for the question, it follows that:

- ***For the questions marked with *:*** *the applicant is legally or on the basis of institutional regulations obliged to ask for an ethical approval at the competent ethics committee of the host institution. Please do take into account that even when there is no obligation with regard to the research itself, for the publication of the results an approval may still be necessary and that no retroactive ethics committee approvals are provided.*

  *If you have answered questions with an * positively, you must submit an ethics approval request with detailed documentation on e.g. study methodology, procedures, informed consent form, insurance, etc to the ethics committee* ***as soon as your application has been approved for funding.*** *Study-specific procedures cannot begin until this ethics approval has been formally given. Only if the approval relates to a work package planned at a later stage of the project, and if legislation allows, the host institution may decide to authorize the researcher to obtain ethical approval at a later stage, i.e. at the latest before the initiation of the relevant part of the research. Please keep in mind that this delayed application/permission is not possible for all research institutions. Also keep in mind that the ethics advisory procedure can take some time and that therefore you should submit your proposal to the ethics committee well in time.*

- ***For the questions that are not marked:*** *Perhaps no ethics approval may be needed for your research proposal. However, please do take into account that your host research institution might have a stricter policy towards ethics approval for certain research topics and methodology. Furthermore, even when there is no obligation with regard to the research itself, for the publication of the results an ethics approval may still be necessary. At any case, the applicant will have to reflect on those issues and take, if necessary, appropriate measures. If in doubt, it is advised to contact the supporting services of your host institution.*

For more information on each of the ethics issues and how to address them, check the FWO webpage on research ethics and the Guidelines on FWO's ethics checklist.

Ethical issues

Are you using human embryos and/or human embryonic stem cells in your study?

No

Does your research involve human subjects?

No

Do you use human cells and/or tissues in your research?

No

Does your study require the processing of personal data?

No

Does your research involve animal testing?

No

Does your research use genetic resources and/or associated traditional knowledge covered by Access and Benefit Sharing legislation and/or the Nagoya Protocol?

No

Does your research involve international collaboration with non-EU countries?

No

Could your research potentially harm the environment and/or the health and safety of people involved?

No

Could your research have dual-use or military applications?

No

Could your research be misused, compromise security and/or human rights?

No

Does your research involve artificial intelligence?

No

Are there any other ethical considerations that need to be taken into account?

No

I confirm that I have read all questions above and that there are no ethical issues concerning my research proposal.

Yes

DATA MANAGEMENT PLAN

Data management is an integral part of sound scientific research. It covers the description of data and metadata, their storage and long-term preservation, the designation of responsible persons, the handling of highly sensitive data, and the open access to and sharing of research data.
The FWO has made data management a key element of its policy for all support channels provided by the FWO. The FWO expects researchers to pay due attention to this dimension before, during and for at least five years after their research.
For background information on data management and the procedures regarding the Data Management Plan (DMP), which FWO expects from its applicants when applying for research funding, please see our website.
Please note that the answers to the questions below and the Data Management Plan should cover the full project, including all (inter)national partners involved in cross-institutional projects.

Describe the datatypes (surveys, sequences, manuscripts, objects ...) you will collect and/or generate and/or (re)use during your research project.

-Genome Data: VCF formatted whole genome sequence data downloaded from public genome repositories.
- Research articles: Final PDF, HTML, and LaTeX representations of papers submitted to conferences and journals, including all sources for figures.
- Software: All implementations and documentation of Solid storage, algorithms, techniques, and supporting tools for running experiments.
- Benchmarks: Queries and datasets for running experiments, with detailed documentation.
- Experimental results: Raw output from benchmarks in machine-readable formats from all used metrics.
- Presentations: Slides and sources for compiling them. If applicable, recordings of the presentation.

679 / 700

Specify in which way the following provisions are in place in order to preserve the data during and at least 5 years after the end of the research.

Motivate your answer.

- *Designation of responsible person (If already designated, please fill in their name.)*
- *Storage capacity/repository*
  - *during the research*
  - *after the research*

IDLab has a strong archiving environment, which is maintained by our IT support team, led by Brecht Vermeulen.
This enables preservation of all my research outcomes for at least five years thanks to nightly redundant backups.
At VITO, similar archiving with multiple backups of the cloud environment on which all files are stored is done.

To increase redundancy, I store all my research data in git repositories linked to my personal account.

445 / 700

What is the reason why you wish to deviate from the principle of preservation of data and of the minimum preservation term of 5 years?

NA

2 / 700

Are there issues concerning research data indicated in the ethics questionnaire of this application form? Which specific security measures do those data require? ***(optional)***

All human genoimic data used for this project will be obtained from publicly available repositories and therefore will not require additional security measures.

160 / 700

Which other issues related to the data management are relevant to mention?

NA

2 / 700

CONSENT

DECLARATION BY THE APPLICANT

General

In completing this application, the applicant confirms that to the best of their knowledge and belief, the information in this application is complete and correct.

The applicant will inform FWO immediately if the intended project cannot be carried out as foreseen or if a major change occurs that may hinder the planned implementation of the project.

The applicant declares that they have read and agree with the FWO regulations that form an integral part of the application documents published on the FWO website and that form the legal basis of the future contract. Furthermore, they take note that the FWO is committed to the principles of the European Charter for Researchers and the Code of Conduct for their Recruitment.

The applicant agrees that the data required for the application and follow-up are electronically stored and used by the FWO. The FWO will use the data provided by the applicant according to the legal requirements of data protection in Belgium, including the use of the anonymized data for statistical purposes and reports. As soon as the FWO has processed your application, you will receive a notification message. The FWO respects the 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) in regards to the processing of your personal data. For more information concerning the privacy policy of the FWO, we redirect you to our website:https://www.fwo.be/en/the-fwo/organisation/processing-personal-data-privacy/.

The applicant agrees that the FWO will forward the full application form including their personal data to, as far as applicable, the members of the FWO expert panels and to experts involved in the evaluation of their proposal in Flanders and abroad (EU and outside EU) and to a partner organization. Any of these receiving parties must declare in advance that they will treat data confidentially and that they will not forward the data or the knowledge gained to anyone nor use it for their own purpose. FWO will take the necessary safety measures to assure this data transfer to the aforementioned organizations or persons will take place in a secure and correct way. More information and details, if available, are published on the FWO website.

The applicant agrees that FWO will forward their private e-mail-address, as provided in the personal data section of the FWO E-portal to their host institution, among other non-personal data regarding their application. The receiving host institution must declare in advance that they will treat data confidentially and that they will not forward the data or the knowledge gained to anyone nor use it for their own purpose. FWO will take the necessary safety measures to assure this data transfer to the aforementioned organizations will take place in a secure and correct way. More information and details are published on the FWO website or can be requested via dpo@fwo.be.

Furthermore, the applicant agrees that the following information may be included in lists published by the FWO: title/abstract; full name of the beneficiaries/supervisors; host institution(s); scientific domains/disciplines/key words; start date and end date, allocated funding of the project.

The applicant declares that all information provided in the personal data section of the FWO E-portal is accurate and up-to-date according to the instructions of the respective programme (i.e. only the items in the E-portal that are applicable to the type of support you apply for should be filled out).

The applicant declares that it fully meets the definition of a research and knowledge-dissemination organization' as stated in Framework for State aid for research and development and innovation 2022/ C 414/01 [1].

Research Integrity

The FWO watches over the scientific integrity from the moment research funding is applied for until the execution of the research and the publication of the research results. Therefore, researchers benefiting from FWO support as well as their host institutions, (co-)supervisors and other collaborators involved in FWO research are required to adhere to the scientific integrity at all times.

To this end, elementary rules of behaviour have been laid down in the Ethical Code for scientific research in Belgium and the European Code of Conduct for Research Integrity. Both documents are included in the call for research proposals. The FWO assumes that each researcher has acknowledged these codes from the moment the application is submitted and undertakes to comply with their provisions in all stages of the proposed research. This also applies to their host institutions, (co-)supervisors and collaborators involved in FWO research, for whom the applicant bears partial responsibility.

If there is any doubt about the applicability or implementation of a provision, the host institution and/ or the researcher responsible for the project at hand will contact the FWO administration in order to clarify or make concrete arrangements about the relevant provision.

[1] an entity (such as universities or research institutes, technology transfer agencies, innovation intermediaries, research-oriented physical or virtual collaborative entities), irrespective of its legal status (organised under public or private law) or way of financing, whose primary goal is to independently conduct fundamental research, industrial research or experimental development or to widely disseminate the results of such activities by way of teaching, publication or knowledge transfer. Where such entity also pursues economic activities the financing, the costs and the revenues of those economic activities must be accounted for separately. Undertakings that can exert a decisive influence upon such an entity, in the quality of, for example, shareholders or members, may not enjoy preferential access to the results generated by it. (Definition of a 'research and knowledge-dissemination organisation').

I agree

Yes

PHD FELLOWSHIP APPLICATION

study results and percentiles

DECLARATION

More info here: NL / EN

First and last name applicant	Elias Crum
First and last name (main) promotor	Ruben Verborgh
(Main) Host institute PhD fellowship	Ghent University
Official name relevant diploma	*Master*
Institute where diploma was obtained	Loyola University Chicago
Academic year diploma was obtained	2022
Study result (XX.xx %)	100%
Number of students in study group	13
Percentile (PXX) **OR** Rank in study group	1st position

Additional remarks:

The applicant and the promotor hereby confirm that the information is correct and acknowledge that the application may be declared inadmissible if the information proves to be incorrect.

Signature applicant:

Signature promotor

Upload this signed declaration as pdf in the application form (e-portal, application form, tab 'personal data')

If a diploma from a non-Flemish university is concerned: add the obtained info from that university to this document.

Heather E. Wheeler, PhD
Associate Professor, Department of Biology
Bioinformatics Graduate Program Director
1032 W. Sheridan Road
Chicago, IL 60660
e-mail: hwheeler1@luc.edu
Phone: 773-508-3629
URL: http://hewlab.org

February 16, 2024

Dear Fellowship Committee,

I confirm Eli Crum received his MS Bioinformatics degree from Loyola University Chicago in August 2022 with a 4.0 GPA, see official transcript for details. In 2022-23, we had 13 MS Bioinformatics graduates, so Eli's ranking based on GPA is 1/13.

Sincerely,

Heather E Wheeler

Heather E. Wheeler, PhD

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*

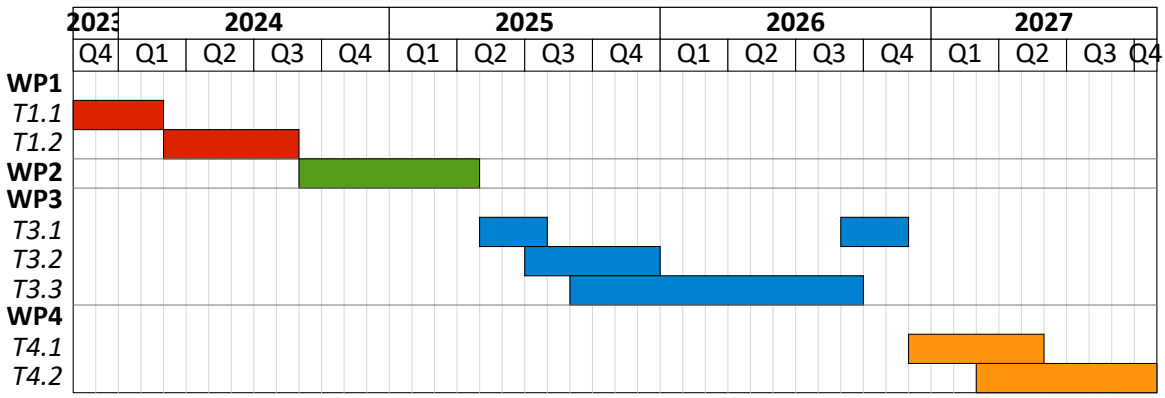
A Gantt chart details when tasks will be undertaken. The primary work packages of my Ph.D. are WP1 and WP3. WP2 is a supporting work package and WP4 will be completed by combining all other work packages. WP1 and WP3 are dependent, meaning that WP1 must be partially completed before WP3 can begin. Thus, 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.