Web Semantics: Science, Services and Agents on the World Wide Web 76 (2023) 100774

Contents lists available at ScienceDirect

Web Semantics: Science, Services and Agents on the World Wide Web

journal homepage: www.elsevier.com/locate/websem

Decentralized semantic provision of personal health streams

Jean-Paul Calbimonte [a,b,*], Orfeas Aidonopoulos [a], Fabien Dubosson [a], Benjamin Pocklington [a,b], Ilia Kebets [c], Pierre-Mikael Legris [c], Michael Schumacher [a]

[a] *University of Applied Sciences and Arts Western Switzerland HES-SO, Sierre, Switzerland*
[b] *The Sense Innovation and Research Center, Lausanne and Sion, Switzerland*
[c] *Pryv SA, Morges, Switzerland*

ARTICLE INFO

Article history:
Received 3 July 2022
Received in revised form 14 November 2022
Accepted 18 January 2023
Available online 20 January 2023

Keywords:
Semantic streams
Personal data streams
Machine learning
Decentralized healthcare data
Open source

ABSTRACT

Personalized healthcare is nowadays driven by the increasing volumes of patient data, observed and produced continuously thanks to medical devices, mobile sensors, patient-reported outcomes, among other data sources. This data is made available as streams, due to their dynamic nature, which represents an important challenge for processing, querying and interpreting the incoming information. In addition, the sensitive nature of healthcare data poses significant restrictions regarding privacy, which has led to the emergence of decentralized personal data management systems. Data semantics play a key role in order to enable both decentralization and integration of personal health data, as they introduce the capability to represent knowledge and information using ontologies and semantic vocabularies. In this paper we describe the SemPryv system, which provides the means to manage personal health data streams enriched with semantic information. SemPryv is designed as a decentralized system, so that users have the possibility of hosting their personal data at different sites, while keeping control of access rights. The semantization of data in SemPryv is implemented through different strategies, ranging from rule-based annotation to machine learning-based suggestions, fed from third-party specialized healthcare metadata providers. The system has been made available as Open Source, and is integrated as part of the Pryv.io platform used and commercialized in the healthcare and personal data management industry.

© 2023 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

1. Introduction

Personal data streams are generated and made available for processing every second, in different domains including social networks, transportation, energy, or healthcare. In the latter case, one must add the additional challenge of handling sensitive information, which is connected to several technical, ethical, and logistic restrictions. The advances in patient monitoring, Internet of Things (IoT) healthcare devices, and big data processing have certainly had a positive impact on the quantity of information that can be collected for patients, either at home or in a healthcare facility. While the volume and velocity of healthcare data streams is already difficult to handle with traditional stream processing systems, the requirements of data integration across devices and institutions have only partially been satisfied by current solutions in the market. Interoperability is one of the key requirements in this context, which has traditionally led to the implementation of data integration mechanisms usually enforced through a top-down approach. However, this approach has been shown to be problematic, considering the increasing need to give data ownership and control to the patients.

Decentralized data management offers an alternative solution, allowing users and patients to flexibly decide where their information is stored and managed. Personal data clouds, or fully decentralized Web solutions as Solid [1] provide potential solutions, although they generally lack the capacity of managing streaming data. Stream reasoning and RDF stream processing has produced important results in this direction [2], combining the rich semantic capabilities of ontology-based data representation, with continuous processing of dynamic data.

In the specific case of healthcare data management, decentralization plays a key role in scenarios like neuro-rehabilitation [3]. After a hospitalization following a stroke or related conditions, patients often need to confront a long process of recovery in a complex environment. These patients have to navigate through a combination of public and private rehabilitation clinics, cabinets, and physiotherapists making regular visits. Sharing data crossing institutional barriers and electronic patient records is a considerable challenge, even more considering the need for protecting sensitive information. In other scenarios, e.g., rare diseases [4], patients are often forced to handle their own records, sometimes

[*] Corresponding author at: University of Applied Sciences and Arts Western Switzerland HES-SO, Sierre, Switzerland.
E-mail address: jean-paul.calbimonte@hevs.ch (J.-P. Calbimonte).

https://doi.org/10.1016/j.websem.2023.100774
1570-8268/© 2023 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

in paper form, given the need for multiple opinions and expertise from scattered healthcare professionals and providers. Nowadays these records contain not only classical patient records, but also rapidly changing information, i.e., streams generated by wearable sensors, specialized medical devices, and even patient-reported observations. Semantic data models have the potential of allowing the usage of healthcare data across devices and providers, with consistent annotations according to well known standards. Therefore, it has become necessary to provide both semantics and streaming features for personal health data management, following a decentralized approach.

In this paper we describe SemPryv, a system that enables the provision of semantically rich streams of personal healthcare data. The system uses Pryv.io as middleware for the management of the streams. It implements different strategies for the semantization of the incoming data events, associating high-level ontology concepts to them. The systems allows for manual and rule-based annotation of streams, as well as automatic prediction through Machine Learning models. Moreover, it allows the provision of semantic streams for healthcare data through the HL7 FHIR standard [5], making it possible to export and share the data streams with other systems and applications. The system can be linked to any ontology provider for the concept annotations, such as BioPortal [6], and it can be accessed through an API or a dedicated UI for healthcare experts. The system has been made available as Open-Source in Github,[1] and a running instance is available for demonstration purposes[2] This paper extends the initial idea and preliminary architecture presented on a poster [7], now completed with fully automated semantic suggestions, full implementation, FHIR import/export, a demonstrator and an Open-Source release. The system is also offered as part of the Pryv.io platform, which is commercialized for decentralized and privacy-compliant management of healthcare and other types of personal data.

The remainder of the paper describes the state of the art in Section 2, the architecture of SemPryv (Section 3), and the overall interoperability approach (Section 4), followed by the description of the semantization approach (Section 5). Experimentation is presented in Section 6, while discussion and future work in Section 7, before the conclusions (Section 8).

2. Related work

Decentralized semantic data solutions have long been studied in the literature, and several approaches and proposals have been presented, for instance for distributed querying [8], linked data provision [9], or collaborative querying [10]. Regarding data privacy, solutions like Solid [11] have shown the potential of decentralized Web architectures, applied to social networks and other domains.

The inclusion of stream processing in decentralized architectures has only gathered attention recently, especially in the context of complex event processing [12], or IoT and fog computing [13]. However, these systems often lack the capabilities for semantic data exchange. Although some preliminary works have been presented to bridge the gap between decentralized agents and stream reasoners [14], there is still a need for technical solutions adapted to the needs of healthcare streaming data processing.

RDF stream processing (RSP) and stream reasoning have produced valuable academic results in the past decade [2], aiming at extending Semantic Web technologies for handling streaming information. Examples of RSP engines include C-SPARQL [15], CQELS [16], SPARQLStream [17], EP-SPARQL [18], RSP4j [19], among others. Even if these engines have demonstrated strong continuous querying capabilities, and even some degrees of reasoning, they still lack the possibility of decentralized deployment, as well as the possibility to deal with strict data privacy constraints.

Regarding healthcare data provision systems, semantic technologies have been used extensively for achieving interoperability [20,21], not only for clinical exchange, but also for secondary use in research [22]. In particular, IoT-based systems have been developed to cope with the large volume of observations collected by medical devices [23]. Given the sensitive nature of health-related data, many previous works have also addressed privacy concerns, for example in electronic health record sharing [24] and linkage [25]. The semantic mediation of different healthcare systems has also been addressed in approaches like [26], where health profiles are collected and curated for wider provision. Blockchain technologies have also been proposed to share healthcare records with certain guarantees on confidentiality, traceability and anonymity [27,28] Other works provided an overview of how semantic interoperability has been addressed in different subdomains, including infectious diseases [29], or oncology [30].

Nevertheless, it remains challenging to combine the power of semantic models with the streaming nature of health data, considering the increasing restrictions and privacy guarantees required in the health domain. The system presented in this paper intends to cover this gap, taking a practical approach, while relying on existing standards as much as possible.

3. SemPryv architecture & features

The SemPryv system addresses the challenges stated in the previous section, more precisely providing APIs to consume and produce inter-operable streams of healthcare data, following the HL7 FHIR standard [5] and incorporating semantic annotations. Fig. 1 depicts the main functionalities provided by SemPryv. First of all it provides an API for semantic ingestion and provision of personal health data streams, which are managed by the Pryv.io platform. Pryv.io is a privacy-centered middle-ware, used as a foundation to develop risk-controlled e-health applications with confidence respect to IT and regulatory requirements. The two key pillars of the Pryv.io platform are: decentralization and privacy. As opposed to traditional healthcare data management systems, Pryv.io stores each data account separately and independently, making it possible to deploy streams of different people in entirely different servers or cloud providers [31]. Data access in Pryv.io can be delegated through token-based authorization, so that patients could explicitly grant or deny access to clinicians or institutions if necessary. Following Pryv.io, data is organized as a hierarchy of streams, each of which is composed of different events. These events are timestamped and may have arbitrary data types, ranging from text, to numerical values, images, or video. The stream contents originate from different sources including wearable devices, IoT streams, and health records and observations. SemPryv addresses the heterogeneity of these data sources, by enriching the with semantic concepts of external ontologies. Given the diversity of potential personal data sources, the accurate semantization of the data is a primary concern in order to provide an added value over the collected information. Moreover, through the Pryv interfaces, the SemPryv system has access to encryption and consent/authentication management. Although in the current implementation of SemPryv is connected only to ontologies in BioPortal such as SNOMED-CT [32], it is open for plugin-based extensions to other ontology providers .

Considering data protection restrictions, SemPryv can be deployed in a fully decentralized configuration. In fact, multiple

[1] Sempryv in Github: https://github.com/pryv/sempryv.

[2] Available at: https://sempryv.ehealth.hevs.ch.

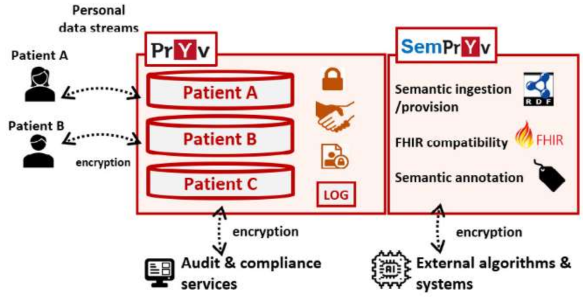


Fig. 1. Schematic view of the features of SemPryv, including its connections with the Pryv.io middleware.

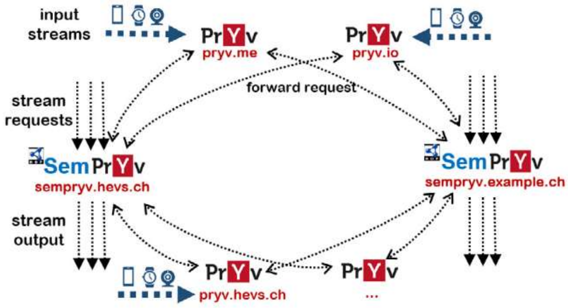


Fig. 2. Decentralized deployment of SemPryv instances, accessing different and multiple Pryv.io streaming data instances.

instances can be set up independently, thus introducing the possibility of multi-site personal data stream provision. As it is depicted in Fig. 2, both the stream generation/consumption and the semantization can be decentralized. As such, different instances of SemPryv can simultaneously connect with different Pryv.io streams, on different servers with dedicated endpoints. SemPryv is designed to act as a proxy for multiple instances of the Pryv.io platform, with the capacity to forward requests through its REST API. SemPryv is able to access any Pryv.io instance, provided that it has a valid authentication token.

The internal architecture of SemPryv is depicted in Fig. 3. SemPryv has two main constituent components: a back-end and a front-end. The back-end can be invoked by external applications to access data streams, or to input stream contents through the FHIR standard through a REST API. the back-end is also in charge of the semantization of streams with external medical vocabularies, as well as the training and prediction of stream metadata through Machine Learning models. The front-end is a Web user interface that domain experts can use to validate and semi-automatically annotate streaming data.

As explained in Fig. 2, the decentralized nature of SemPryv allows to deploy different back-ends and connect them to an arbitrary number of stream providers through Pryv.io. In practice, this entails that, for instance, a patient may have streams managed in different hospitals, yet semantically integrated through a single SemPryv instance. The back-end can also be connected to different ontology providers, i.e. services that allow retrieving a medical/healthcare ontology, or semantic vocabulary. The implementation allows for developing plugins in order to connect to different ontology providers, but for demonstration purposes, one was built connecting to the BioPortal [6] API, which grants access to vocabularies like LOINC or SNOMED-CT. As detailed in Section 5 the connection with these providers is used for automatic and semi automatic annotation validated or confirmed by an expert.

4. Data stream interoperability

A key requirement in healthcare stream data management is interoperability, linked to privacy protection. In SemPryv, we showcase how the system is able to provide endpoints for both requesting and submitting HL7 FHIR-compliant data streams. Given that SemPryv streams are composed of events, we encode these as Observations in the FHIR model, as exemplified in Listing 1. The FHIR standard can be serialized in different concrete formats, including RDF formats like Turtle. In the excerpt we can see a body temperature observation, including its actual value, units, and linked to its corresponding concepts extracted from both LOINC and SNOMED-CT. Multiple classifications and usage of multiple coding systems are allowed for every observation.

5. Stream enrichment: Annotations & prediction

As seen in the previous section, the FHIR representation of the streams require the linkage to semantic concepts form external ontologies. This process can be performed in SemPryv in three different ways: (i) Manually, by searching concepts in well-known ontology providers (such as BioPortal) that are connected to the platform through a unifying API; (ii) Semi-automated, i.e., where annotation suggestions are provided to the users. These suggestions are derived by predefined rules that experts can modify and save them in the system's knowledge graph; (iii) Fully automated, i.e., suggestions are provided by machine learning models that have been trained on existing data previously annotated, combined with annotations continuously provided by users.

As seen in Fig. 4 the user can navigate through the stream hierarchy and choose to add a new semantic code, searching from a dynamic list populated through the ontology providers described earlier (e.g. Bioportal). Once the expert identifies a semantic annotation form the list (Fig. 5) the observation is classified using this concept. Multiple annotations are possible at the stream level, and annotations can be inherited recursively by sub-streams and events inside of the hierarchy. In the example, the user can access her streams after authentication, and the top hierarchy displays two streams: Body Temperature (BT) and Heart (H). Heart is actually a parent to a child-stream named Heart Rate (HR). Through the search option, for instance for the body temperature, one can find a list of potentially associated SNOMED-CT codes.

In addition, SemPryv includes the possibility of using predefined rules expressed in its knowledge graph. These rules can be modified by administrators, and essentially allow the definition of close terms from different ontologies. For instance in the following example, the knowledge graph matches Pryv temperature streams to a SNOMET-CT code identified as: snomed-ct:386725007. Similarly the same is done for the concept of mass. SemPryv also allows matching these rules to different combinations of stream paths, according to the hierarchy established among streams.

Apart from the possibilities of performing manual search on existing annotations, and the use of custom rules, SemPryv can use existing classified streams of data to create pre-trained models using Machine Learning algorithms. This approach can use previously trained models that are fed into the system, and also integrate streams classified using the rules and/or manual annotation. In this manner, the suggestions are expected to improve in terms of quality as more data is annotated. As a demonstrator for this feature, the Open-Source SemPryv code-base includes the implementation of a Naive Bayes classifier for multinomial models. This model has been implemented using the Python *scikit-learn* library, and has been trained using two datasets, *Thryve* and *Pulso*. These datasets are focused mainly on exercises and daily

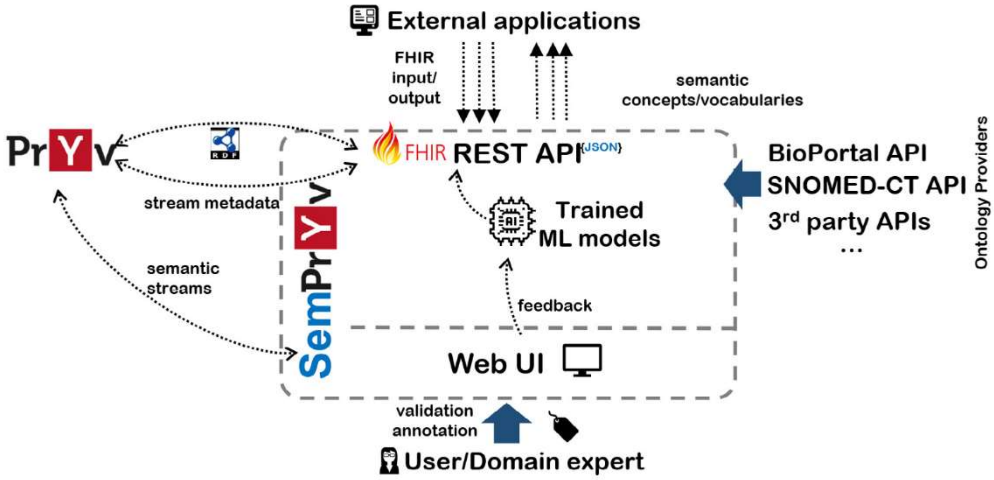


Fig. 3. SemPryv architecture: Main components including the user interface and the back-end API, as well as the interactions with the ontology providers.

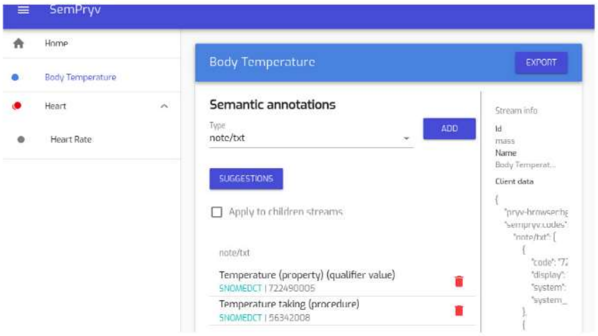
```
@prefix fhir: <http://hl7.org/fhir/> .
@prefix loinc: <http://loinc.org/rdf#> .
@prefix sct: <http://snomed.info/id/> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .

<http://hl7.org/fhir/Observation/bodyTemp01> a fhir:Observation;
  fhir:Resource.id [ fhir:value "bodyTemp01"];
  fhir:Observation.category [
     fhir:CodeableConcept.coding [
       fhir:Coding.code [ fhir:value "vital-signs" ];
       fhir:Coding.display [ fhir:value "Vital Signs" ] ] ];
  fhir:Observation.code [
    [ a loinc:8310-5;
      fhir:Coding.system [ fhir:value "http://loinc.org" ];
      fhir:Coding.code [ fhir:value "8310-5" ];
      fhir:Coding.display [ fhir:value "Body temperature" ] ],
    [ a sct:56342008;
      fhir:Coding.system [ fhir:value "http://snomed.info/sct" ];
      fhir:Coding.code [ fhir:value "56342008" ];
      fhir:Coding.display [ fhir:value "Temperature taking" ] ];
     fhir:CodeableConcept.text [ fhir:value "Temperature" ] ];
  fhir:Observation.subject [
     fhir:link <http://hl7.org/fhir/Patient/p100>;
     fhir:Reference.reference [ fhir:value "Patient/p100" ]; ];
  fhir:Observation.issued [ fhir:value "2021-02-04T14:27:00"^^xsd:dateTime];
  fhir:Observation.performer [
     fhir:link <http://hl7.org/fhir/Practitioner/m021>;
     fhir:Reference.reference [ fhir:value "Practitioner/m021" ] ];
  fhir:Observation.valueQuantity [
     fhir:Quantity.value [ fhir:value "39"^^xsd:decimal ];
     fhir:Quantity.unit [ fhir:value "degrees C" ];
     fhir:Quantity.system [ fhir:value "http://unitsofmeasure.org" ];
     fhir:Quantity.code [ fhir:value "Cel" ] ];
.
<http://hl7.org/fhir/Patient/p100> a fhir:Patient .
<http://hl7.org/fhir/Practitioner/m021> a fhir:Practitioner .
```

Listing 1: FHIR representation of an observation of a body temperature reading encoded in RDF Turtle format.

```
  "@graph": [{
      "@id": "pryv:temperature", "@type": "skos:Concept",
      "skos:notation": "note/txt",
      "skos:closeMatch": "snomed-ct:386725007", },
    {
      "@id": "pryv:mass", "@type": "skos:Concept",
      "skos:notation": "mass",
      "skos:closeMatch": "snomed-ct:118538004" },
    {
      "@id": "someRuleSet",
      "pryv:pathExpression": ".*/group",
      "pryv:mapping": ["pryv:mass", "pryv:temperature"]  },
```

Listing 2: Predefined rules mapping Pryv concepts to ontology concepts like SNOMED-CT.



Body Temperature

Semantic annotations

Type
note/txt

□ Apply to children streams

note/txt

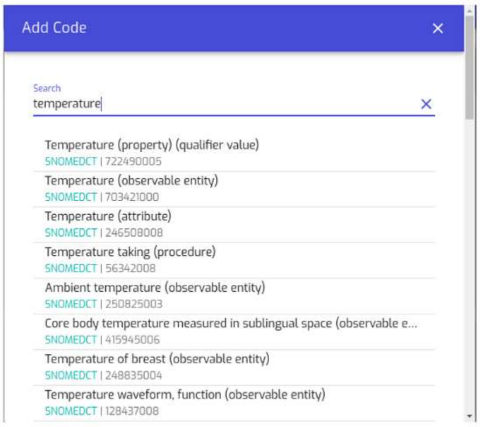
Temperature (property) (qualifier value)
SNOMEDCT | 722490005

Temperature taking (procedure)
SNOMEDCT | 56342008

Id
mass
Name
Body Temperat...
Client data

```
{
  "pryv-browser:bg
  "sempryv:codes"
    "note/txt": [
      {
        "code": "72
        "display":
        "system":
        "system_
      },
      {
```

Fig. 4. User interface of SemPryv allowing access to stream hierarchies, events, and their metadata. The UI also allows querying semantic metadata, adding annotations and providing suggestions fed by the machine learning models.



Search
temperature

Temperature (property) (qualifier value)
SNOMEDCT | 722490005

Temperature (observable entity)
SNOMEDCT | 703421000

Temperature (attribute)
SNOMEDCT | 246508008

Temperature taking (procedure)
SNOMEDCT | 56342008

Ambient temperature (observable entity)
SNOMEDCT | 250825003

Core body temperature measured in sublingual space (observable e...
SNOMEDCT | 415945006

Temperature of breast (observable entity)
SNOMEDCT | 248835004

Temperature waveform, function (observable entity)
SNOMEDCT | 128437008

Fig. 5. Suggested SNOMED-CT term annotations obtained from the BioPortal provider.

living activities. The data produced by the *Thryve*[3] wearable data API structures streams related to different measurements including: dailys steps, walked distance, daily burned calories, sleep duration, glucose level, daily weight, among many others. The second dataset originates from the *Pulso* fitness App, which also includes information about exercises and activities, as well as bio-demographic information. From these datasets, the already annotated streams have been used to train the classifier, and provide SemPryv with predicted semantic labels. The Naive Bayes classifiers shipped with SemPryv uses this trained model to provide potential suggestions automatically, but implementers can add additional annotated datasets or use other classification algorithms. The accuracy of the prediction models will of course be linked to the quality of the data used for the training, as well to the pertinence to the domain of application (e.g., activity monitoring, medical devices streams, etc.). For example, a *heart rate* stream in *Thryve* could be annotated as LP91312−6 in LOINC, and 428420003 in SNOMED-CT. However, if a stream about MRI measurements is input, and no similar data exists in the system, then the recommendations will be less accurate. The classifier provided in the project provides a extensible demonstrator so that domain-specific datasets can be incorporated as needed, for use cases beyond exercise and daily-living activities.

6. Experimentation

Different experiments have been conducted in order to analyze the potential use of SemPryv in different scenarios, especially for the provision of FHIR-compliant streams.

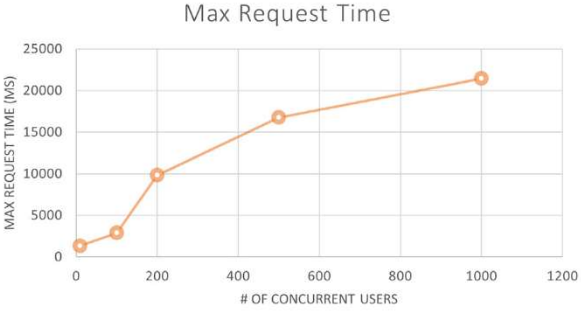


Fig. 6. Number of concurrent users against the maximum time taken to process a request.

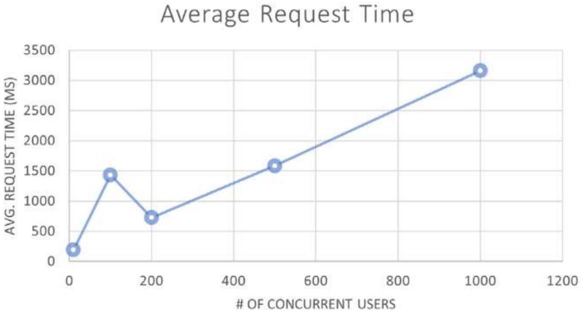


Fig. 7. Average request time for all amounts of concurrent users.

A simulation of users accessing different streams through the SemPryv API was performed, first varying the number of requests (for 10k, 20k, 30k, and 40k simultaneous requests), and for each request count, different numbers of users (10, 100, 200, 500, and 1000). These tests were performed on the following hardware : Intel Xeon i7 E5-2670 2.50 GHz, 132 GB of RAM and a total disk size of around 4TB. The SemPryv service was run in three Docker containers - *sempryv-frontend*, *sempryv-backend* and *sempryv-proxy*. While SemPryv was not the only service to be running on the server, an inspection of incoming requests determined that any impact on its performance would have been negligible. In the evaluation the streams have been accessed using the SemPryv API, via REST calls. Other access mechanisms, including WebSockets and Webhooks have been incorporated recently in the platform, mainly for notification management. Regarding access control of the streams, for simplicity, the evaluation only considers stream events visible to the test user for annotation and exporting into FHIR resources.

In the evaluation, we first analyzed the number of concurrent users against the maximum time taken to process a request. It can be seen in Fig. 6 that after 100 users the maximum request time increases quickly, although slowing down after 500 users.

Regarding the average request time for all amounts of concurrent users (Fig. 7), we see an important decrease between 100 and 200 users. This is because after 100 users, requests start being rejected — so they are processed somewhat quickly, but then starting to take more and more time to complete.

When analyzing the percentage of requests that are successful plotted against the number of concurrent users (Fig. 8), it can be observed that after around 100 concurrent users, 80% of requests fail. This is due to the configuration of the server, which stops single IP addresses making too many requests — to protect from spam and denial of service attacks. In order to test the ability of the service with large amounts of unique users an enterprise-grade testing environment would be required.

Next, we analyzed the maximum time a request takes plotted against the total number of requests (Fig. 9). As it appears the

[3] https://thryve.health/.

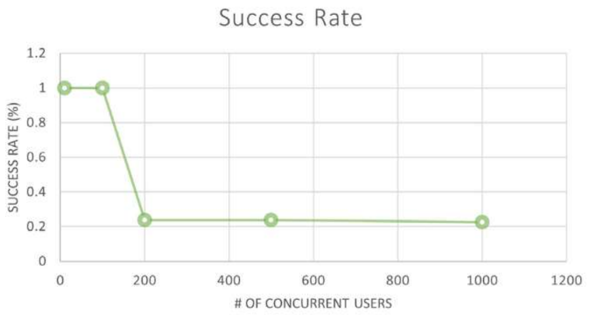


Fig. 8. Percentage of requests that are successful plotted against the number of concurrent users.

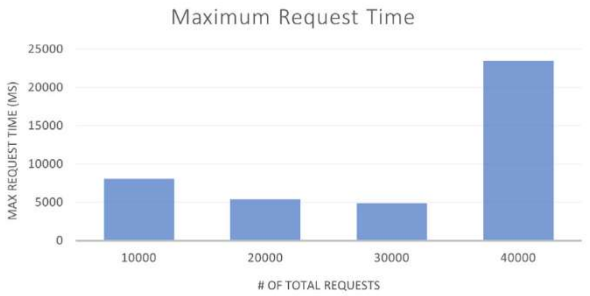


Fig. 9. Maximum time a request takes plotted against the total number of requests.

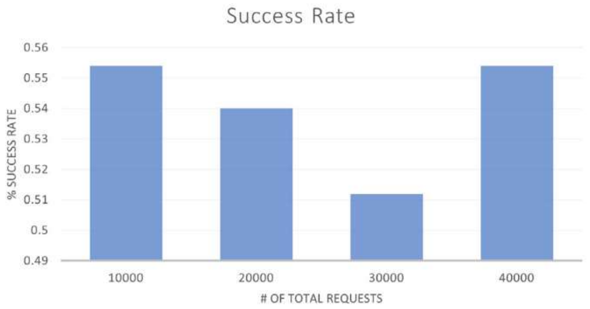


Fig. 10. Percentage of successful requests plotted against the total number of requests.

number of requests does not seem to have much of an effect on the slowest requests.

Regarding The % of successful requests plotted against the total number of requests in the set (Fig. 10), they all average out very closely, with only a difference of around 4%–5% overall

In summary the biggest factor regarding the FHIR interface is the number of concurrent users making requests. This is on a first level associated to the limitation of concurrent requests, and is required in order to protect the server infrastructure from external attacks and spam. The number of sequential requests has very little to no effect on the performance of the FHIR stream data provision. This indicates that this feature does not take an exorbitant amount of time to complete and that the bottleneck lays within the server infrastructure.

In order to scale the service to a basic commercial level it would need to be run on its own dedicated server, which would need to be configured and maintained by an administrator to handle the expected amount of traffic. The SemPryv service could also be scaled-up to accept a large amount of users and connections by using an orchestration system like Kubernetes (K8s). Kubernetes is capable of managing multiple servers, pooling them as a single resource and load-balancing requests, it achieves this by creating new instances of containers on-the-fly according to the load on the network.

7. Discussion and future work

SemPryv provides a practical framework for organizing personal data streams with semantic interoperability features, following a decentralized approach and considering privacy aspects. The proposed solution focuses as much as possible in already available solutions and technologies, as well as existing adopted standards. SemPryv is deployed as a middle-ware, designed to work as a bridge between client applications and multiple healthcare systems. This entails the necessity of using standards for healthcare record interchange, specifically through the HL7 FHIR endpoints. In this manner, different hospitals and clinics can expose selected portions of the data that they provision, each of them annotated with the suggested entities proposed by SemPryv. One example of this type of interactions is the SwissNeuroRehab[4] network, funded by Innosuisse. This initiative aims to connect patient clinical data records from multiple clinics, hospitals, and cabinets, so that a continuum of care can be established. Built on top of different systems, SwissNeuroRehab includes Pryv.io in its architecture, which is essential to allow the integration of incoming streams of data. A key advantage is the flexibility of the framework: SemPryv is not too prescriptive on the type of ontologies to be used for annotations, or in the type of devices that feed the stream. In the future we plan on integrating other ontologies as core modeling options for SemPryv, for instance the SSN ontology [33], which is relevant for personal and healthcare sensor data streams. Another addition would be to add continuous query languages as part of the SemPryv interface, following existing works such as CQELS [16].

Concerning the semantic enrichment of the data streams we plan to further enhance the rule-deducing approach for establishing suggestions of semantic metadata, in cases where a bootstrapping process is required. As for the machine learning-based approach, we also plan on evaluating large scale enrichment of data streams, and addressing some of the limitation of potential lack of ground truth.

In regard to privacy protection, this is one of the key points of the underlying Pryv.io storage system. It is GDPR compliant, allowing users to have complete control of their data, even if it is used by different healthcare facilities. Moreover, it provides several features for data privacy protection, including fine grained access control settings (who has access and for which specific data element), built-in consent management, encryption for data during transmission, segmentation of data and aliasing, selective sharing of (pseudo) anonymized data, a dedicated audit module, isolation of per-user data in back-ups, etc. SemPryv benefits from these features, enabling semantic provision on top. In the future we also plan to extend customization of external ontologies, which are currently configured by an administrator.

The prototype is publicly available, as referenced earlier, and the code has been made available as Open Source. We intend to further use the system in several health monitoring scenarios such as neurological rehabilitation, and physiotherapy, in order to provide further evidence of its use.

8. Conclusions

In this paper we have described SemPryv, a system that allows the semantic enrichment of personal data streams, set up in a fully distributed environment. The proposed approach has been fully implemented, comprising not only the semantization but

[4] https://www.swissneurorehab.ch/.

also (i) its integration with external providers such as BioPortal, (ii) the implementation of an interoperability bridge through HL7 FHIR, and (iii) a rule-based automated suggestion feature and machine-learning prediction of stream semantics. The system is currently deployed, showcasing the use of semantics in real-life scenarios and on integrated with a commercial solution. The SemPryv approach relies on two main principles. First, on the reuse of consolidated vocabularies, ontologies, and taxonomies that are standardized and widely used in the domains of application. This is the case for well-known standards (e.g. SNOMED-CT, LOINC; UCUM, RxNorm) which have been curated to enable interoperability among applications. Second, SemPryv uses different, but complementary approaches for proposing semantics for a given dataset, depending on: the available data, metadata, and previous inferences.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Links to code are provided in the paper.

Acknowledgment

Supported by the InnoSuisse, Switzerland grant 26547.1 SemPryv project.

References

[1] E. Mansour, A.V. Sambra, S. Hawke, M. Zereba, S. Capadisli, A. Ghanem, A. Aboulnaga, 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, 2016, pp. 223–226.

[2] D. Dell'Aglio, E. Della Valle, F. van Harmelen, A. Bernstein, Stream reasoning: A survey and outlook, Data Sci. 1 (1–2) (2017) 59–83.

[3] W. Deng, I. Papavasileiou, Z. Qiao, W. Zhang, K.-Y. Lam, S. Han, Advances in automation technologies for lower extremity neurorehabilitation: A review and future challenges, IEEE Rev. Biomed. Eng. 11 (2018) 289–305.

[4] S. Courbier, R. Dimond, V. Bros-Facer, Share and protect our health data: an evidence based approach to rare disease patients perspectives on data sharing and data protection-quantitative survey and recommendations, Orphanet J. Rare Dis. 14 (1) (2019) 1–15.

[5] D. Bender, K. Sartipi, HL7 FHIR: An agile and restful approach to healthcare information exchange, in: Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems, IEEE, 2013, pp. 326–331.

[6] M. Salvadores, P.R. Alexander, M.A. Musen, N.F. Noy, BioPortal as a dataset of linked biomedical ontologies and terminologies in RDF, Semant. Web 4 (3) (2013) 277–284.

[7] J.-P. Calbimonte, F. Dubosson, I. Kebets, P.-M. Legris, M. Schumacher, Semi-automatic semantic enrichment of personal data streams, in: Poster Proceedings of the 15th International Conference on Semantic Systems (SEMANTiCS 2019), 9-12 September 2019, 2019.

[8] C. Aebeloe, G. Montoya, K. Hose, A decentralized architecture for sharing and querying semantic data, in: European Semantic Web Conference, Springer, 2019, pp. 3–18.

[9] A. Polleres, M.R. Kamdar, J.D. Fernández, T. Tudorache, M.A. Musen, A more decentralized vision for linked data, Semant. Web 11 (1) (2020) 101–113.

[10] A. Grall, H. Skaf-Molli, P. Molli, M. Perrin, Collaborative sparql query processing for decentralized semantic data, in: International Conference on Database and Expert Systems Applications, Springer, 2020, pp. 320–335.

[11] A.V. Sambra, E. Mansour, S. Hawke, M. Zereba, N. Greco, A. Ghanem, D. Zagidulin, A. Aboulnaga, T. Berners-Lee, Solid: a Platform for Decentralized Social Applications Based on Linked Data, Tech. Rep., MIT CSAIL & Qatar Computing Research Institute, 2016.

[12] V.P. De Almeida, S. Bhowmik, G. Lima, M. Endler, K. Rothermel, DSCEP: an infrastructure for decentralized semantic complex event processing, in: 2020 IEEE International Conference on Big Data (Big Data), IEEE, 2020, pp. 391–398.

[13] P. Dazzi, M. Mordacchini, Scalable decentralized indexing and querying of multi-streams in the fog, J. Grid Comput. 18 (3) (2020) 395–418.

[14] J.-P. Calbimonte, D. Calvaresi, M. Schumacher, Decentralized management of patient profiles and trajectories through semantic web agents, in: Seamntic Web for Healthcare SWH Co-Located with ISWC, 2020, pp. 19–29.

[15] D.F. Barbieri, D. Braga, S. Ceri, E. Della Valle, M. Grossniklaus, C-SPARQL: a continuous query language for RDF data streams, Int. J. Semant. Comput. 4 (01) (2010) 3–25.

[16] D. Le-Phuoc, M. Dao-Tran, J. Xavier Parreira, M. Hauswirth, A native and adaptive approach for unified processing of linked streams and linked data, in: International Semantic Web Conference, Springer, 2011, pp. 370–388.

[17] J.-P. Calbimonte, O. Corcho, A.J. Gray, Enabling ontology-based access to streaming data sources, in: International Semantic Web Conference, Springer, 2010, pp. 96–111.

[18] D. Anicic, P. Fodor, S. Rudolph, N. Stojanovic, EP-SPARQL: a unified language for event processing and stream reasoning, in: Proceedings of the 20th International Conference on World Wide Web, 2011, pp. 635–644.

[19] R. Tommasini, P. Bonte, F. Ongenae, E. Della Valle, Rsp4j: An api for rdf stream processing, in: European Semantic Web Conference, Springer, 2021, pp. 565–581.

[20] X. Zenuni, B. Raufi, F. Ismaili, J. Ajdari, State of the art of semantic web for healthcare, Procedia-Soc. Behav. Sci. 195 (2015) 1990–1998.

[21] H. Liyanage, P. Krause, S. De Lusignan, Using ontologies to improve semantic interoperability in health data, BMJ Health Care Inform. 22 (2) (2015).

[22] C. Gaudet-Blavignac, J.L. Raisaro, V. Touré, S. Österle, K. Crameri, C. Lovis, et al., A national, semantic-driven, three-pillar strategy to enable health data secondary usage interoperability for research within the swiss personalized health network: Methodological study, JMIR Med. Inform. 9 (6) (2021) e27591.

[23] S. Balakrishna, M. Thirumaran, Semantic interoperability in IoT and big data for health care: a collaborative approach, in: Handbook of Data Science Approaches for Biomedical Engineering, Elsevier, 2020, pp. 185–220.

[24] S. Martínez, D. Sánchez, A. Valls, A semantic framework to protect the privacy of electronic health records with non-numerical attributes, J. Biomed. Inform. 46 (2) (2013) 294–303.

[25] Y. Lu, R.O. Sinnott, Semantic privacy-preserving framework for electronic health record linkage, Telemat. Inform. 35 (4) (2018) 737–752.

[26] F.A. Satti, T. Ali, J. Hussain, W.A. Khan, A.M. Khattak, S. Lee, Ubiquitous Health Profile (UHPr): a big data curation platform for supporting health data interoperability, Computing (ISSN: 1436-5057) 102 (11) (2020) 2409–2444, http://dx.doi.org/10.1007/s00607-020-00837-2.

[27] M. Kumar, S. Chand, MedHypChain: A patient-centered interoperability hyperledger-based medical healthcare system: Regulation in COVID-19 pandemic, J. Netw. Comput. Appl. 179 (2021) 102975.

[28] R. Jabbar, N. Fetais, M. Krichen, K. Barkaoui, Blockchain technology for healthcare: Enhancing shared electronic health record interoperability and integrity, in: 2020 IEEE International Conference on Informatics, IoT, and Enabling Technologies (ICIoT), IEEE, 2020, pp. 310–317.

[29] X. Gansel, M. Mary, A. van Belkum, Semantic data interoperability, digital medicine, and e-health in infectious disease management: a review, 38 (6) 1023–1034 http://dx.doi.org/10.1007/s10096-019-03501-6.

[30] T.J. Osterman, M. Terry, R.S. Miller, Improving cancer data interoperability: the promise of the minimal common oncology data elements (mCODE) initiative, JCO Clin. Cancer Inform. 4 (2020) 993–1001.

[31] S. Goumaz, White paper: data in Pryv, 2018, URL https://pryv.com/data_in_pryv/.

[32] K. Donnelly, SNOMED-CT: The advanced terminology and coding system for ehealth, Stud. Health Technol. Inform. 121 (2006) 279.

[33] A. Haller, K. Janowicz, S.J. Cox, M. Lefrançois, K. Taylor, D. Le Phuoc, J. Lieberman, R. García-Castro, R. Atkinson, C. Stadler, The modular SSN ontology: A joint W3C and OGC standard specifying the semantics of sensors, observations, sampling, and actuation, Semant. Web 10 (1) (2019) 9–32.