Semi-automatic Semantic Enrichment of Personal Data Streams

Jean-Paul Calbimonte[1], Fabien Dubosson[1], Ilia Kebets[2], Pierre-Mikael Legris[2], and Michael Schumacher[1]

[1] Institute of Information Systems,
University of Applied Sciences and Arts Western Switzelrand (HES-SO),
Sierre, Switzerland
{firstname.lastname}@hevs.ch
[2] Pryv SA, Lausanne, Switzerland
info@pryv.com

Abstract. Current information technologies allow people to acquire personal data related to their health, lifestyle, behavior, and activities, often using wearable and mobile devices. Personal data management technologies have emerged recently, in order to cope with the requirements of this type of data, ranging from personal clouds to self-storage solutions. Pryv.io is a comprehensive solution for managing this particularly sensible type of data streams, focusing both on data privacy and decentralization. In this paper, we describe SemPryv, a system aiming at providing a semantization mechanism for enriching personal data streams with standardized specialized vocabularies from third-party providers. It relies on third providers of semantic concepts, and includes rule-based mechanisms for facilitating the semantization process. A full implementation of SemPryv has been produced, pluggable to the existing Pryv.io platform, showing the feasibility of the approach.

1 Context & Motivation

The increasing amount of generated personal data allows for the development of personalized applications in different domains, usually related to health, lifestyle, or everyday activities. These often rely on different sources and acquisition modalities, including wearable devices, sensors, domotic technologies, or self-reporting methods. In this context, it is essential to provide data privacy guarantees, in order to avoid unintended access or disclosure. This difficulty to address this challenge is further increased due to the streaming nature of many of these datasets, which require infrastructure designed to manage high-volume and high-velocity information flows.

Pryv.io is a privacy-centric middleware, used as a robust data management foundation to develop risk-controlled mHealth, eHealth, and InsurTech applications with confidence and in respect to IT and regulatory requirements. Pryv.io is built based on two key pillars: decentralization and privacy. Unlike traditional centralized solutions, Pryv.io stores each data account separately and independently, making it possible to be even deployed on its own server [3]. Furthermore, data access can be delegated in a modular way, providing token-based authorization, e.g. for tertiary use by a clinician. Data is organized as a hierarchy of streams, each containing a series of events of different nature and type. Given the large heterogeneity of data sources in these areas, and the velocity of the data, it becomes essential to provide the means for automatically categorizing them according to standard ontologies and vocabularies, especially in the health domain. Given the diversity of potential personal data sources (e.g. from time series of a smartwatch to health record annotations), the accurate semantization of the data is a primary concern in order to provide an added value over the collected information.

This paper describes the SemPryv subsystem for stream data enrichment[3]. The goal of SemPryv is to provide semantization capabilities for the Pryv.io middleware, such that it can automatically propose semantic concepts, associated to the heterogeneous data streams managed by the platform. The data semantization makes it possible to enhance the data model, currently conformed by typed events. Associating high-level ontology concepts to the stream events enables new types of search and discovery functionalities in the middleware, which were not possible up to now. Also, it provides the means to link the Pryv datasets with existing standards and models used widely for cataloguing data in the health sector. In particular, the use of standards, such as HL7 FHIR [1], make it possible to export and share the Pryv.io data with other systems and applications, as long as it is annotated with semantic vocabularies. The system described in this paper focuses on both the service-oriented architecture of SemPryv and its interaction with existing ontology providers as BioPortal [4], as well as a dedicated UI that allows experts to confirm or choose from the semantics suggestions offered by the module. The implementation of the system shows the feasibility of our fully decentralized solution for semantization of personal data streams, relying on the widely used HL7 FHIR standards for interoperability.

2 SemPryv Architecture

The Pryv.io middleware is used to manage large and diverse streams of data coming from external platforms, wearable devices, and health record systems. These streams are organized through identifiers and tags that are later used for searching and querying. While it is technically possible to export and make the Pryv datasets available to external applications in different formats, standards such as HL7 FHIR [1] impose the necessity of adding explicit semantic annotations to the Pryv.io data streams, for instance using the SNOMED-CT [2] vocabulary. While this semantization process could be carried out manually, it is unrealistic and too time consuming to be realizable. The SemPryv module enables the addition of semantics to the datasets, in an automatic, or semiautomatic manner. Given the decentralized nature of Pryv.io, multiple instances can be used in order to store and manage isolated data streams. SemPryv is designed

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

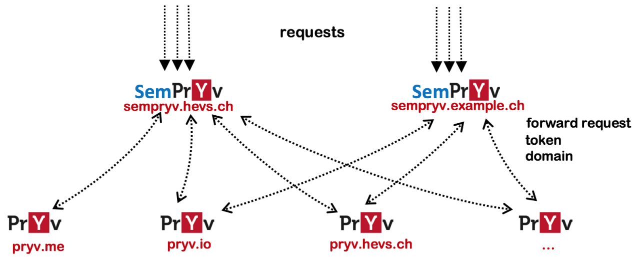


Fig. 1. Decentralized deployment in Pryv and proxy access through SemPryv: every Pryv.io instance can be enhanced with SemPryv, with an authorization token.

to act as a proxy for these instances, being able to forward requests to Pryv.io through its REST API (Figure 1). By passing an authentication token and the domain within the request, SemPryv is able to access any Pryv.io instance, and add the semantic annotation/suggestion features, as well as the FHIR support.

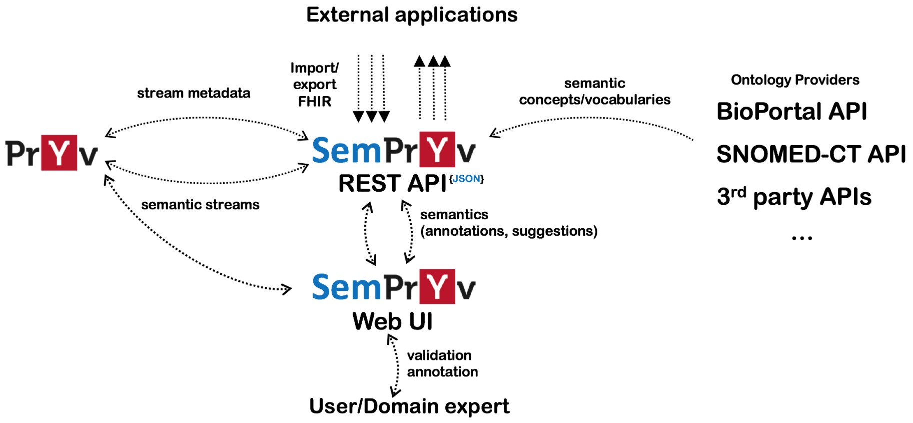
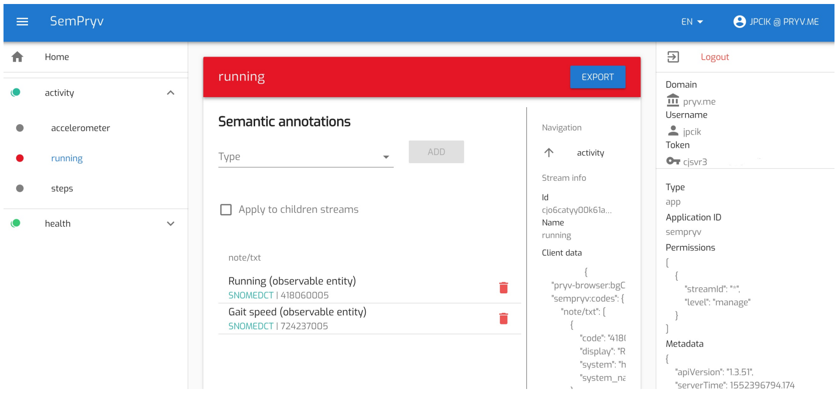


Fig. 2. SemPryv architecture: Prxv.io interactions with the SemPryv back-end and UI, as well as the ontology providers.

The architecture of SemPryv is depicted in Figure 2. SemPryv has two main components: a back-end that exposes the core services as a REST API, and a UI for end-users and experts. Besides the proxy capabilities mentioned before, the SemPryv back-end can connect to a series of providers for semantic vocabularies and ontologies. These may include existing APIs such as BioPortal [4], or other collections of relevant ontologies. SemPryv is able to query these providers in order to suggest relevant ontology terms for a given Pryv stream, or hierarchy of streams, which can then be validated, or confirmed by an expert through the SemPryv Web UI. The Pryv.io metadata can be then updated according to these suggestions and annotations. Additionally, the SemPryv back-end includes endpoints dedicated for the import/export of HL7 FHIR-compliant data streams, represented as bundle collections of observations.

3 SemPryv Suggestions & Annotations



running

Semantic annotations

Type

□ Apply to children streams

note/txt

Running (observable entity)
SNOMEDCT | 418060005

Gait speed (observable entity)
SNOMEDCT | 724237005

Stream info

Id
cjo6catyy00k61a...
Name
running
Client data

```
{
 "pryv-browser:bgC
 "sempryv:codes": {
  "note/txt": [
   {
    "code": "418(
    "display": "R
    "system": "h
    "system_n&
```

Domain
pryv.me
Username
jpcik
Token
cjsvr3

Type
app
Application ID
sempryv
Permissions

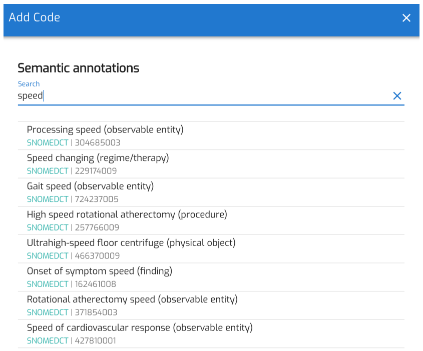
```
[
  {
    "streamId": "*",
    "level": "manage"
  }
]
```

Metadata

```
{
  "apiVersion": "1.3.51",
  "serverTime": 1552396794.174
```

Fig. 3. SemPryv: User interface, including access to streams hierarchies, events, and their metadata.

The architecture presented previously describes how the different components of the system interact with each other. Concerning the semantization process itself, the SemPryv module is flexible enough to adapt to different types of situations. For users and data integrators, the SemPryv UI (Figure 4) exposes the proposed semantics, queried from the 3rd party providers (e.g. BioPortal). Then, these suggestions can optionally be confirmed by an administrator before being consolidated into its corresponding Pryv instance. This semi-automatic semantization makes it possible to have full control over the type of semantics to be assigned. As an example, a body-weight stream in the Pryv.io middleware can be modeled as the weight of an individual according to SNOMED-CT, codified as: SNOMED-CT:27113001. Notice that multiple annotations can be attached to a given stream, and that these annotations can be inherited recursively by sub-streams and events inside of its hierarchy. Furthermore, other custom 3rd party ontology/vocabulary providers can be configured in order to feed the system. 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 mass. Then, the system also allows to match these rules to certain stream paths, defined using regular expressions.



Semantic annotations

speed

Processing speed (observable entity)
SNOMEDCT | 304685003

Speed changing (regime/therapy)
SNOMEDCT | 229174009

Gait speed (observable entity)
SNOMEDCT | 724237005

High speed rotational atherectomy (procedure)
SNOMEDCT | 257766009

Ultrahigh-speed floor centrifuge (physical object)
SNOMEDCT | 466370009

Onset of symptom speed (finding)
SNOMEDCT | 162461008

Rotational atherectomy speed (observable entity)
SNOMEDCT | 371854003

Speed of cardiovascular response (observable entity)
SNOMEDCT | 427810001

Fig. 4. SemPryv UI: suggested annotations obtained from the BioPortal provider: SNOMED-CT terms.

```
"@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 1.1. Predefined rules mapping pryv concepts to SNOMED-CT.

4 Discussion & Future Work

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 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. The system is currently deployed, showcasing the use of semantics in real-life scenarios and on integrated with a commerical 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 (eg. 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. For the next iteration of the SemPryv module, we will further enhance the rule-inferencing approach for establishing suggestions of semantic metadata, in cases where a bootstrapping process is required. Once a critical mass of data is acquired, SemPryv will rely on incremental machine learning techniques to correlate previously annotated datasets with new incoming streams. The prototype is publicly available, as referenced earlier, and in the future an Open Source approach will be considered, given the potential reuse opportunities.

Acknowledgements Supported by the InnoSuisse grant 26547.1 SemPryv project.

References

1. Bender, D., Sartipi, K.: Hl7 fhir: An agile and restful approach to healthcare information exchange. In: Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems. pp. 326–331. IEEE (2013)
2. Donnelly, K.: Snomed-ct: The advanced terminology and coding system for ehealth. Studies in health technology and informatics **121**, 279 (2006)
3. Goumaz, S.: White paper: data in pryv (2018), https://pryv.com/data_in_pryv/
4. Salvadores, M., Alexander, P.R., Musen, M.A., Noy, N.F.: Bioportal as a dataset of linked biomedical ontologies and terminologies in rdf. Semantic web **4**(3), 277–284 (2013)