

Global Initiative for Sentinel e-Health Network on Grid (GINSENG): Medical Data Integration and Semantic Developments for Epidemiology

S. Cipiere, G. Ereteo, A. Gaignard, N. Boujelben, S. Gaspard, Vincent Breton, F. Cervenansky, David R.C. Hill, T. Glatard, D. Manset, et al.

▶ **To cite this version:**

S. Cipiere, G. Ereteo, A. Gaignard, N. Boujelben, S. Gaspard, et al.. Global Initiative for Sentinel e-Health Network on Grid (GINSENG): Medical Data Integration and Semantic Developments for Epidemiology. Workshop on Clusters, Clouds and Grids for Health 2014, May 2014, Chicago, United States. pp.755-763, 10.1109/CCGrid.2014.45 . hal-01048737

HAL Id: hal-01048737
https://hal.science/hal-01048737

Submitted on 2 Mar 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

Global Initiative for Sentinel e-Health Network on Grid (GINSENG)

Medical data integration and semantic developments for epidemiology

Sébastien Cipière[a,b,c], Guillaume Ereteo[h], Alban Gaignard[i], Nouha Boujelben[e], Sébastien Gaspard[g], Vincent Breton[a,b], Frédéric Cervenansky[e], David R. C. Hill[a,b,d], Tristan Glatard[e,f], David Manset[g], Johan Montagnat[i], Jérôme Revillard[g], Lydia Maigne[a,b]

cipiere@clermont.in2p3.fr, guillaume.ereteo@mnemotix.com, alban.gaignard@cnrs.fr, nouha.boujelben@creatis.insa-lyon.fr, sgaspard@gnubila.fr, breton@clermont.in2p3.fr, frederic.cervenansky@creatis.insa-lyon.fr, david.hill@univ-bpclermont.fr, tristan.glatard@creatis.insa-lyon.fr, dmanset@gnubila.fr, johan.montagnat@cnrs.fr, jrevillard@gnubila.fr, maigne@clermont.in2p3.fr

(a) Clermont Université, Université Blaise Pascal, LPC, BP 10448, Clermont-Ferrand, France
(b) CNRS/IN2P3, UMR 6533, LPC, Aubière, France
(c) CNRS, UMR 6158, Université Blaise Pascal, LIMOS, Aubière, France
(d) ISIMA, Institut Supérieur d'Informatique, de Modélisation et de leurs Applications, BP 10125, Aubière, France
(e) Université de Lyon, CREATIS ; CNRS UMR5220 ; Inserm U1044 ; INSA-Lyon ; Université Lyon 1, Lyon, France
(f) McConnell Brain Imaging Centre, Montreal Neurological Institute, McGill University, Montréal, Canada
(g) GNUBILA, 174 Impasse Près En Bas, Argonay, France
(h) SCIC Mnemotix 318 Avenue de la Carraire, Mandelieu La Napoule, France
(i) Université de Nice Sophia Antipolis / CNRS UMR7271, I3S laboratory, Sophia Antipolis, Nice, France

Abstract— **The implementation of a grid network to support large-scale epidemiology analysis (based on distributed medical data sources) and medical data sharing require medical data integration and semantic alignment. In this paper, we present the GINSENG (Global Initiative for Sentinel e-Health Network on Grid) network that federates existing Electronic Health Records through a rich metamodel (FedEHR), a semantic data model (SemEHR) and distributed query toolkits. A query interface based on the VIP platform, and available through the e-ginseng.org web portal helps medical end-users in the design of epidemiological studies and the retrieval of relevant medical data sets.**

Keywords: Sentinel e-health network, medical metamodel, Electronic Health Record ontology, medical semantic web interface

I. INTRODUCTION

Epidemiology studies require valid and reliable data in order to determine appropriate strategies for the control of specific indicators [1]. In order to ensure that a control strategy is effective and appropriate, the data need to be of high quality. Consequently, epidemiology requires a rigorous and systematic approach to data management. The importance of data management seems often underestimated, with greater emphasis placed on the study design, data collection, and data analysis. This can result in an ad-hoc approach to data management that ultimately affects the reliability and validity of the data collected and increases the workload involved in data curation.

Since 2010, the GINSENG project (e-ginseng.org) has grouped IT researchers and physicians in order to create a proof-of-concept network for epidemiology in Auvergne dedicated to cancer surveillance and perinatal health. Presently, patient records are faxed or delivered using a simplified e-messaging between healthcare professionals to be duplicated in their databases. This process is not efficient, time consuming and can be responsible for errors caused by a dual registration of the medical information. Otherwise, interactions between healthcare professionals remain ineffective for the patient follow-up, mainly because of a lack of interoperability between information systems. Furthermore, the French Health Watch Institute (InVS), equivalent to CDC (Centers for Disease Control and Prevention) for the USA, is in charge of publishing indicators about population health and particularly about cancer incidence [2], [3], [4], [5]. To produce such indicators, InVS relies on regional data warehouses set up to collect relevant information to support statistical and epidemiological studies about cancer incidence, mortality, prevalence or screening [6]. Until now, cancer registries remain few and not easily interoperable with other medical databases.

The GINSENG project develops a dedicated grid infrastructure to support large-scale epidemiology analysis based on distributed, heterogeneous and confidential medical data sources (hospitals, cancer screening associations, cancer centers and pathology labs, etc.). It addresses the well-identified problems of increasing the number of interesting medical variables and observations, improving the representativeness of the data (typological and geographical), and maximizing the validity of information [7], [8]. This objective does not require nominative patient data even if unambiguity on patient identification and data ownership is

necessary to correlate data entries. Data interoperability can be greatly enhanced by using a semantic approach to define, describe information to exchange and facilitate the exploitation of data available in multiple data sources [9], [10]. Therefore, the integration of dedicated technologies to deal with a semantic description of medical data is needed to aggregate additional information about patients' follow-up and highlight variables of interest regarding epidemiology.

In a first instance, two medical fields are benefitting from this infrastructure: cancer and perinatal health. Preliminary studies have been undertaken in the "Sentinel Network Cancer Auvergne" with different laboratories and hospitals in the Auvergne Region (France) [11], [12].

In this paper, we present a network infrastructure based on lightweight dedicated services for a distributed approach of medical data integration and fusion (section 2). This data integration relies on a common data metamodel and on a fast patient identification technique presented in section 3. Section 4 describes a semantic data model allowing for data to be linked in a knowledge graph. The last section presents integration of an API built on the VIP platform [13] to query this knowledge graph.

II. Network Architecture for e-Health and Epidemiology

A. Global infrastructure

The GINSENG Network architecture is detailed in Fig. 1. It adopts a flexible Service-Oriented Architecture (SOA) to harness the complexity of such a distributed system. It federates different medical databases, from pathology lab cancer screening to PACS systems, etc. Each database is duplicated on a grid server (gateway) located at the home institution. An administration server is in charge of managing user authentication and security features. Users are authenticated using their French professional health cards identifiers. This way, physicians can log in the GINSENG network with their unique professional ID, and transparently access to distributed computing resources by means of credential delegation. Each gateway is organized in different layers of services:

- The Anonymization Service offers various data anonymization possibilities according to medical data types being manipulated within the system (e.g. DICOM headers blanking/encryption/removal, image blanking/scrambling etc).
- The Grid Interface Service bridges with all other grid middleware services such as computational services.
- The Database Interface Service bridges with all XML and relational database back-ends. Storage management and data access are managed by AMGA (ARDA Metadata Grid Application) [14]. AMGA is installed as an additional layer on top of a DataBase Management System (DBMS), usually MySQL, and provides a structured tree to manage folders with flexible access control mechanisms for individual data items based on Access Control Lists (ACLs).

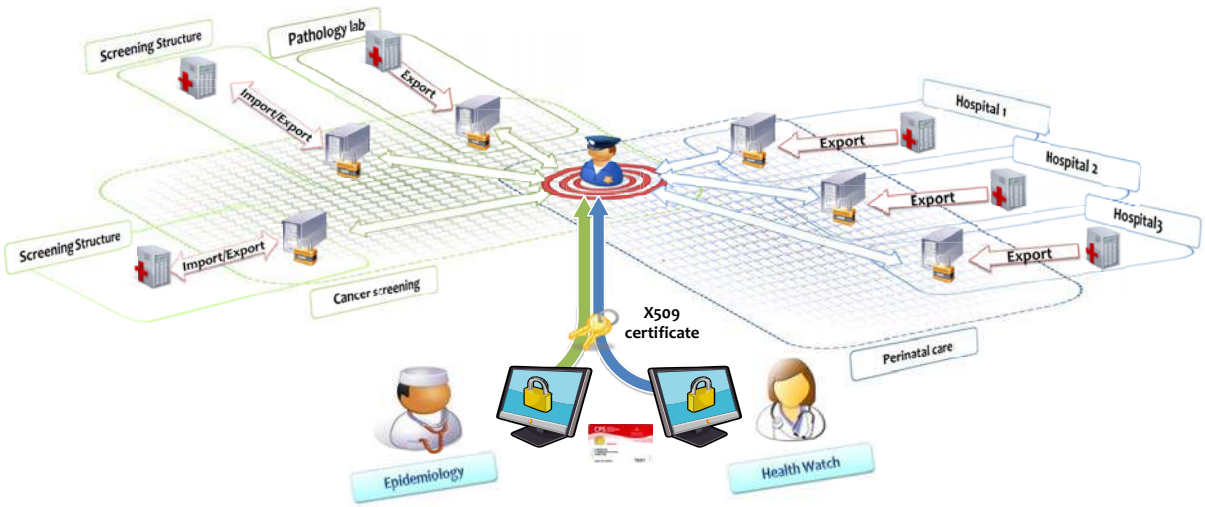


Figure 1. GINSENG network architecture

B. *Specificities for medical data management*

Each stakeholder of the GINSENG project owns a grid server hosting two kinds of services:

The *medical data upload and standardization service* automatically uploads the medical database. Each time the database is uploaded (usually out of working hours, once a month), new patients are identified and specific IDs are allocated for their registration in the GINSENG network [15]. In order to control patient duplicates, a patient identification technique, based on a probabilistic record linkage method taking into account field similarity in the calculation of field weights [16], [17] has been implemented.

The data is then structured through the FedEHR format (see section 3.A) in order to be queried via the web portal e-ginseng.org.

The *medical data query service* allows external users to query the grid server, according to the local security and authentication policy fixed by VOMS.

Two data query methodologies are considered in the GINSENG project. The SQL relational database query language can be used for direct access to the native database entries (see Section 3) and the SPARQL semantic graph query language can be used to access highly structured and semantically-rich knowledge graphs (see Section 4). In both cases, a data federation layer (named *Distributed Query Processing*) is needed to query multiple data stores distributed over data providing sites.

III. MEDICAL DATA INTEGRATION

A. *Electronic Health Record implementation*

FedEHR, the Federated Electronic Health Record data warehouse platform has been developed to store a large variety of medical concepts. FedEHR allows building vendor-neutral and cloud-enabled patient-centric EHR data warehouses. It has been primarily designed for biomedical research.

FedEHR makes it possible to add medical data according to their original format or standards such as the Systematized Nomenclature of Medicine (SNOMED-CT)[1] or the International Classification of Diseases (ICD-10)[2]. Based on grid/cloud technologies, it provides an extensible repository for patient data. When queried, the FedEHR service can behave as a centralized database giving access to all the databases connected on the network.

The EHR standard provides a metamodel made of general entities common to all medical fields: the *Patient*, the *Visit*, the *MedicalEvent* and the *ClinicalVariable*. Those four entities are organized hierarchically in order to represent the medical history of the patient: each patient undergoes visits, each visit is related to medical events and each medical event is a clinical dataset. It has to be noted that a clinical variable can be linked to a clinical variable dataset.

[1] http://www.ihtsdo.org/snomed-ct

[2] http://www.who.int/whosis/icd10

The mapping between medical datasets and FedEHR can be settled using an XML language in order to specify more information on data types. Those medical events, referenced in different medical databases are glued together using a FedEHR organization after patient identification.

B. *Medical database queries*

FedEHR implements an AMGA-based multi-tier repository for data stored in XDS (Cross-enterprise Document Sharing), XDS-I (Cross-Document Sharing for Imaging), DICOM (Digital Imaging and Communications in Medicine), and HL-7 (Health Level Seven) standards, and can handle any other data type. It combines the abstraction power of AMGA together with a proprietary big data technology to provide a powerful data warehouse to securely share medical information across the healthcare enterprise and transnationally, while preserving patient privacy and data copyrights. Using FedEHR, federated data sources are replicated locally, transformed and made available to the GINSENG network.



Figure 2. FedEHR Query generator GUI

Using AMGA, FedEHR behaves in each center as a standard SQL database. Inherited from this architecture, AMGA SQL queries can be performed through FedEHR web services returning an aggregation of results from each data center. This solution carries a serious drawback

when information about one patient is distributed in different sites. Using standard AMGA queries, distributed patient information fragments are not properly joined and the result of a query could be incomplete. To address this issue, FedEHR provides a former technology called SQL². Based on the structured medical database, involving clinical variables depending of medical events, FedEHR is performing SQL queries at two levels. A first set of queries is executed following the methodology of a standard AMGA query on the different metadata (Medical Events or Clinical Variables). Using result of these first level queries, temporary tables are created on a query-enabled computing element, which can be either a physical server, a virtual machine, instantiated in a cloud or a job in a grid enabled environment. Once created, these tables enable standard SQL query to be run on the new database.

This query system is a simple and valuable solution avoiding data replication around data centers and ensuring data property to providers. From a user standpoint, clinical variables are presented through a specific GUI proposing a standard SQL query interface (Fig. 2). From a technical standpoint, for each SQL² query, only the needed data is transferred through a secured encrypted protocol to the computing element sending the query. This reduces bandwidth consumption and ensures a high level of security.

Thanks to FedEHR, GINSENG carries and extends the Grid intrinsic capabilities to address the complex and demanding requirements of health surveillance. FedEHR makes it possible to extend GINSENG to other medical fields.

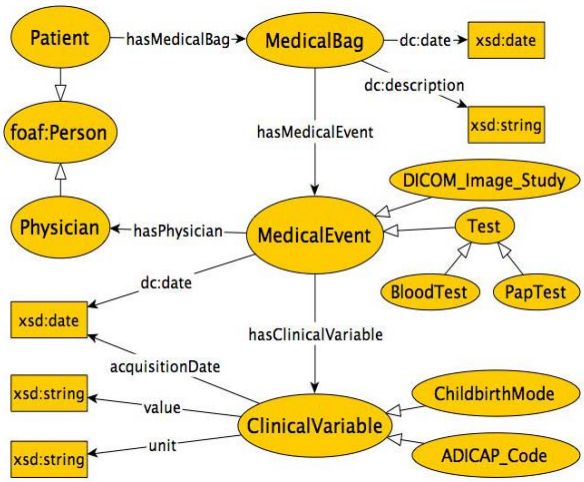


Figure 3. Core primitives of the semEHR ontology

IV. Data Semantic Description

A. Knowledge graphs based on electronic health records

To properly align heterogeneous data stores with widely accepted and well-understood concepts, the data semantics should be taken into account. The Semantic Web[3] addresses this problem with a rich set of widely accepted standards for attaching semantics to data and linking data sets (RDF), describing common concepts (RDFS and OWL), and enabling rich querying (SPARQL query language).

RDF (Resource Description Framework) represents distributed identities and concepts as a uniform directed and typed graph even if entities are located on different knowledge bases. It describes resources with triples (subject, predicate and object) that can be viewed as "the subject, verb and object of an elementary sentence", "a natural way to describe the vast majority of the data processed by machines" [18]. Unique Resource Identifiers (URIs) are used to uniformly identify entities and concepts of each description, linking descriptions and models across knowledge bases. Thereby, RDF allows publishing medical data from any autonomous infrastructure, while enabling to interconnect whole distributed datasets as a uniform Linked Data structure [19].

RDFS (Resource Description Framework Schema) and OWL (Ontology Web Language) are RDF-based standards to define ontologies. An ontology is defined by [20] as "a set of representational primitives with which to model a domain knowledge". The representational primitives are typically classes, attributes, and relationships among class members. The definitions of the representational primitives include information about their meaning and constraints on their logically consistent application. Ontologies offer a semantic layer to knowledge graphs that can be mined by Semantic Web compliant engines in order to consolidate inconsistent and incomplete data, and to discover new added information.

Finally SPARQL (Sparql Protocol And RDF Query Language) provides an expressive query language to retrieve triple patterns, as well as different protocols and formats to send queries and their results across networks. In the project, we publish each medical database with a SPARQL endpoint, enabling to send query and to retrieve results through HTTP.

This Semantic Web layer provides a reasoning capabilities layer enabling the data consolidation, and interoperability with external linked data sources, such as Dbpedia[4]. Fig. 3 represents the core of the Electronic Health Record ontology, semEHR, that we designed to describe the SQL medical data of FedEHR. "Patient" and "Physician" are typed with eponym subclasses of *foaf:Person*[5]. A "Patient" has a "Medical Bag" (*semehr:MedicalBag*) which contains multiple "Medical

[3] Semantic Web, W3C, http://www.w3.org/2001/sw/

[4] DBPedia Sparql Endpoint http://fr.dbpedia.org/sparql

[5] Friend Of A Friend ontology: http://xmlns.com/foaf/spec/

Events" (*semehr:MedicalEvent*) occurring during follow-up care. The measured clinical variables (*semehr:ClinicalVariable*) are attached to corresponding medical events with the property *semehr:hasClinicalVariable*. Properties are also defined to describe the different attributes of classes such as the date of a medical event or the value of a clinical variable.

This core is an extensible basis that can represent the variety of medical data encountered in the GINSENG project. In particular, we defined various subclasses of *semehr:MedicalEvent* in order to represent the taxonomy of the multiple types of medical events.

For instance, the class *semehr:Test* is a subclass of *semehr:MedicalEvent* which has multiple subclasses to represent different types of medical tests, including blood test and Pap Test.

Similarly, an extensible taxonomy of *semehr:ClinicalVariable* subclasses represents the different types of clinical variables that are measured during medical events, such as an ADICAP [6] code resulting from a Pap Test. Semantic Web compliant engines can automatically handle such polymorphic structures, which enables different granularity of querying.

B. Distributed linked health data querying with KGRAM-DQP

Regarding the context of the project, and the distribution of medical data, it seems more appropriate to implement a single data federator that provides a uniform view over a set of federated and autonomous healthcare data centers. The principle consists in decomposing and rewriting a unique SPARQL query into a set of sub-queries sent to the different medical databases. Results are retrieved, joined and merged into a single set of results finally presented in a unified way.

Corese/KGRAM [21] is a generic and versatile Semantic Web factory aiming at representing, querying and reasoning on Knowledge Graphs. It is fully compliant with the W3C Semantic Web standards (RDF/RDFS/SPARQL). It has been extended with federation capabilities (KGRAM-DQP) to address distribution scenarios, particularly relevant in the context of the GINSENG project and more generally in Linked Open Data [19] scenarios.

KGRAM-DQP [22] addresses multiple and possibly heterogeneous data sources, as illustrated in Fig. 4. It decomposes the initial SPARQL graph patterns into node/edge requests (triple pattern subqueries), which are sent to multiple data sources concurrently. A distribution component is responsible for federating the results of subqueries provided by multiple data sources, each interfaced through a "Producer" component.

[6] ADICAP thesaurus is a french codification for lesions of anatomic pathology http://www.adicap.asso.fr/

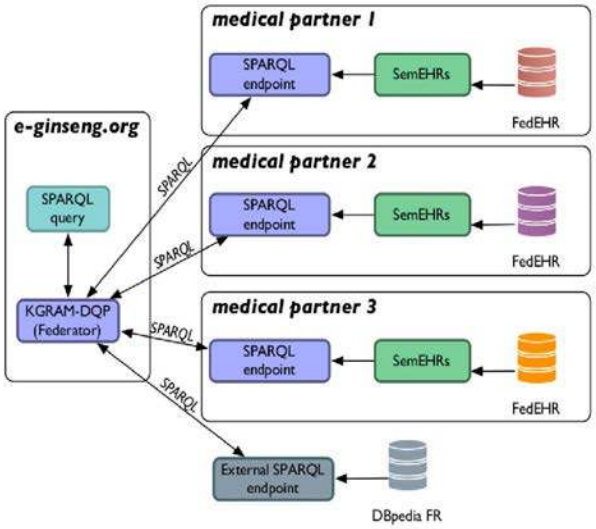


Figure 4. Distributed knowledge base architecture

Each "Producer" can either serve local or remote graph-based data sources. In addition, to address data heterogeneity issues, "Producers" also act as data mediators, rewriting incoming subqueries into the data source native query language, and finally transforming the native results into a graph-based representation.

In practice, KGRAM-DQP handles federated SPARQL queries through an iteration over all subqueries, which are sent concurrently to the data sources connected, wrapped into simple "CONSTRUCT" SPARQL clauses that generate triple results. Candidate results returned are finally joined into the result graph. In spite of subqueries concurrent execution, the federated approach can lead to inefficient querying scenarios, due to the amount of remote queries and transferred results finally joined into the federator. KGRAM-DQP implements a source selection mechanism and a set of static and dynamic optimizations, to reduce the inherent network overhead of federated querying.

C. Bridging local health data and open linked data

To illustrate the medical data federation capabilities, we implemented a scenario combining medical data (in this case the ADICAP code) and public demographic data for the needs of epidemiological studies. The GINSENG distributed knowledge base contains the necessary EHRs, while demographic data are provided by the French DBpedia instance (semantic data extracted from wikipedia). The SPARQL query illustrated in Fig. 5 first searches for patients whose disease is characterized by the "BHGSA3FO" ADICAP code (lines 4 to 6) and their postal code. These triple patterns are sent to the SPARQL endpoints of each medical partner. Then a SERVICE clause (line 8) indicates that the triple patterns enclosed (lines 10 to 12) are sent to the DBpedia data source. This subquery searches for the total population corresponding to postal codes of the Auvergne region in France. Once retrieved, results are joined, based on the postal code so as

to count the number of patients affected by the disease encoded "BHGSA3FO" per geographic area. The size of the corresponding population is then displayed.

This federated querying scenario has been simulated on a standard laptop computer. Three SPARQL endpoints have been implemented as illustrated in Fig. 4. Medical databases have been simulated for the moment with fictive anonymized patient information. Medical data size consists in around 5 Million triples without taking into account the DBpedia dataset. The query shown in Fig. 5 retrieves 209 results in about one minute. The query execution time can dramatically be reduced by manually grouping triple patterns into a UNION of SERVICE clauses leading to only few SPARQL queries sent over the network. Triple patterns grouping strategies open interesting optimization perspectives but need further investigation due to non-trivial data partitioning, mixing both horizontal (instances of the same data model split in several databases) and vertical (DBPedia data in a single store) data fragmentation.

```
1: PREFIX semehr: <http://www.mnemotix.com/ontology/semEHR#>
2: PREFIX dbpedia-owl: <http://dbpedia.org/ontology/>
3: SELECT (count(distinct ?patient) as ?nbPatients) (sum(?pop) as ?totalPop) ?postalCode WHERE {
4:     ?cv semehr:value "BHGSA3F0"^^xsd:string .
5:     ?patient semehr:hasMedicalBag/semehr:hasMedicalEvent/semehr:hasClinicalVariable ?cv .
6:     ?patient semehr:address/semehr:postalCode ?postalCode .
7:
8:     SERVICE <http://fr.dbpedia.org/sparql> {
9:         SELECT DISTINCT (str(?cp) as ?postalCode)  ?pop WHERE {
10:            ?s dbpedia-owl:region <http://fr.dbpedia.org/resource/Auvergne> .
11:            ?s dbpedia-owl:postalCode ?cp .
12:            ?s dbpedia-owl:populationTotal ?pop .
13:        }
14:    }
15: }
16: GROUP BY ?postalCode
17: ORDER BY ?totalPop
```

Figure 5. Example of GINSENG SPARQL Query

V. WEB INTERFACE TO SEMANTIC DATABASES

A. The Virtual Imaging Platform

The Virtual Imaging Platform (VIP) [13] is a web portal dedicated to the simulation and processing of medical data on distributed computing resources. VIP exposes scientific applications as services, mainly for medical simulation and neuro-image analysis. To handle the computing and storage needs of these applications, it is supported by the European Grid Infrastructure (http://www.egi.eu), where it consumes about 50 CPU years every month. To date, more than 500 users from 50 countries can access VIP.

While scientific computing was the initial motivation for VIP, the current evolution of scientific practices toward so-called Big Data leads to the integration of various data repositories in the platform. The semantic databases exploited by the GINSENG project are an example of such integration. The VIP platform is accessible from the e-ginseng.org web portal through an iframe.

B. Interfaces to semantic databases

Interfaces to semantic databases available in VIP implement the following use cases:

1. Experts (e.g. epidemiologists and computer scientists) develop and publish queries;
2. End users (e.g. healthcare professionals or health watch authorities) parameterize and execute published queries;

Four tools were developed to implement these use cases: the Query Maker, Query Executor, Query History, and Query Explorer.

The Query Maker is a tool for SPARQL experts to develop and publish queries (Fig. 6). Developers can enter the body of a SPARQL query, and use placeholders to define parameters instantiated when the query is executed. The query can then be saved and possibly published to end users. The Query Maker allows to describe queries as complex as SPARQL permits, and to expose them to end-users. Future improvements could include features usually available in SPARQL development environments such as Flint (http://openuplabs.tso.co.uk) and the Datalift project (http://datalift.org), for instance syntax highlighting, syntax checking, auto-completion, etc.

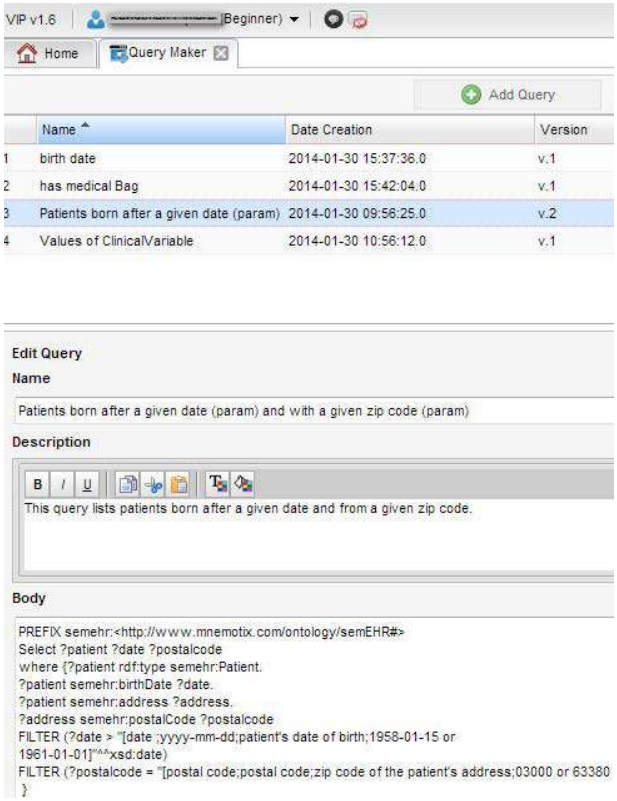


Figure 6. Query Maker in VIP. The highlighted text shows the syntax used to define parameters in the SPARQL query: [parameter name, type, description, example(s)]

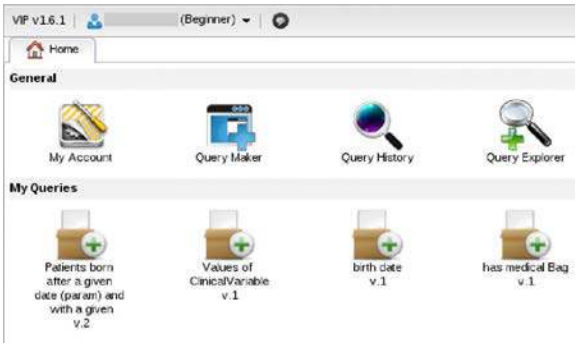


Figure 7. Query Executor in VIP (access to published queries)

With the Query Executor, end users can access, parameterize and launch public queries (Fig. 7 and Fig. 8).

When a query is executed, its parameter placeholders are replaced by user values to produce the final query body. The query body is then put in an execution queue and asynchronously processed by an agent that updates the status of the executions, from waiting to running and finally completed or failed. Query executions can be viewed and completed results can be retrieved from the Query History (Fig. 9).



Figure 8. Query Executor in VIP (query parameterization and launch)

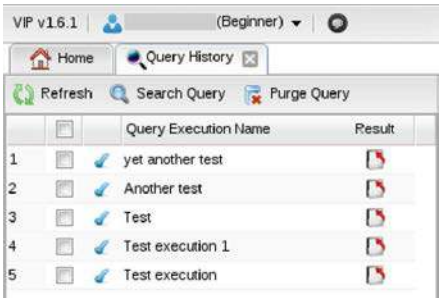


Figure 9. Query History in VIP

The Query Explorer is a tool for end users to specify and execute their own queries (Fig. 10). Its graphical user interface (GUI) completely hides the SPARQL backend. End-users can select variables in a graph, set the data sources from which these variables should be extracted, and possibly specify restrictions about their values. Underneath the GUI, a SPARQL query is built by the system: the SELECT clause of this query contains the variables selected by the user, the FROM clause contains the selected data sources, and the WHERE clause contains the variables and their restrictions. ORDER BY and GROUP BY clauses can also be specified using checkboxes available in advanced options. The resulting query can then be executed as previously described, and monitored in the Query History.

The Query Explorer allows end users to develop their own queries with no exposure to SPARQL. This, however, limits the range of queries that can be described. For instance, queries involving the OPTIONAL construct, aggregation functions such as SUM and AVG, or unions of multiple SELECT constructs currently cannot be described. Feedback from end users could help in determining whether this interface should be extended to enable a larger set of queries.

The tools described in this section are available and testable in the production instance of VIP at (https://vip.creatis.insa-lyon.fr).

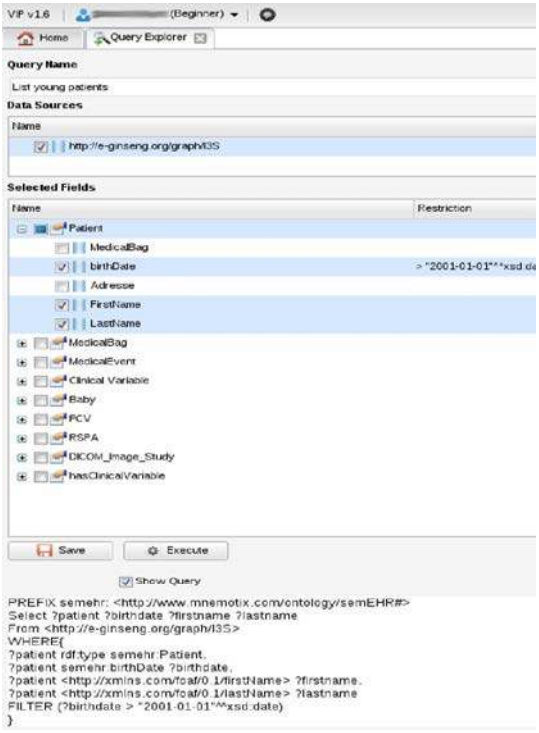


Figure 10. Query Explorer in VIP

VI. CONCLUSION

This article describes the work in progress applied to a surveillance network for cancer and perinatal health through a distributed architecture. The network will allow federating, in a fully secured way, medical databases located in hospitals in order to make it available to epidemiologists for health watch studies.

Different developments have been addressed to answer healthcare professionals' needs regarding patient's follow-up and pathology surveillance:

- Metamodels: the integration and the management of patient data sheets have been processed through the usage of metamodels adopting FedEHR metadata standards in the objective to extract pertinent information for epidemiologic studies.
- Data semantic: the elaboration of an RDF repository allowing for data to be linked and to be "grounded" in semantic descriptions using a set of triples organized into a RDF Graph. The semantic engine Corese/KGRAM is used to query the RDF database with SPARQL language.
- Web semantic: the integration of an API to query the RDF database using SPARQL endpoints. This API is integrated to the VIP platform and through an iframe into the e-ginseng.org web portal.

Future questions regarding the management of medical images will be triggered to complete the capabilities of the network by proposing close developments of tools and services validated by healthcare professionals; this is of key importance in order to offer the best achieved network.

ACKNOWLEDGMENT

This work was supported by the French National Research Agency under grant N° ANR-10-TECSAN-008-01.

REFERENCES

[1] A. Fink, "Epidemiological Field Work in Population-Based Studies," in *Handbook of Epidemiology SE - 11*, W. Ahrens and I. Pigeot, Eds. Springer Berlin Heidelberg, 2005, pp. 399–436.

[2] P. Guénel and S. Villeneuve, *Cancer du sein, professions et expositions professionnelles aux solvants organiques*. 2013, p. 55.

[3] F. Binder-Foucard, A. Belot, and P. Delafosse, "Estimation nationale de l'incidence et de la mortalité par cancer en France entre 1980 et 2012," 2013.

[4] P. Grosclaude, L. Remontet, A. Belot, A. Danzon, N. Rasamimanana Cerf, and N. Bossard, "Survie des personnes atteintes de cancer en France, 1989-2007. Etude à partir des registres des cancers du réseau Francim," 2013.

[5] D. Jezewski Serra and E. Salines, "Évaluation épidémiologique du programme de dépistage organisé du cancer colorectal en France," 2013.

[6] N. Bossard, M. Velten, L. Remontet, A. Belot, and S. Bara, "Survie des patients atteints de cancer en France: principaux resultats de la premiere etude du reseau des registres francais des cancers (Francim)," *Oncologie*, 2007.

[7] J. A. Magnuson and J. Paul C. Fu, *Public Health Informatics and Information Systems*. Springer London, 2014.

[8] V. Breton, K. Dean, T. Solomonides, I. Blanquer, V. Hernandez, E. Medico, N. Maglaveras, S. Benkner, G. Lonsdale, S. Lloyd, K. Hassan, R. McClatchey, S. Miguet, J. Montagnat, X. Pennec, W. De Neve, C. De Wagter, G. Heeren, L. Maigne, K. Nozaki, M. Taillet, H. Bilofsky, R. Ziegler, M. Hoffman, C. Jones, M. Cannataro, P. Veltri, G. Aloisio, S. Fiore, M. Mirto, I. Chouvarda, V. Koutkias, A. Malousi, V. Lopez, I. Oliveira, J. P. Sanchez, F. Martin-Sanchez, G. De Moor, B. Claerhout, and J. A. M. Herveg, "The Healthgrid White Paper.," *Stud. Health Technol. Inform.*, vol. 112, pp. 249–321, 2005.

[9] H. Peixoto, J. Machado, J. Neves, and A. Abelha, "Semantic Interoperability and Health Records," in *E-Health SE - 30*, vol. 335, H. Takeda, Ed. Springer Berlin Heidelberg, 2010, pp. 236–237.

[10] J. Vejvalka, P. Lesny, T. Holecek, K. Slaby, H. Krasnicanova, A. Jarolimkova, and H. Bouzkova, "Semantic description of health record data for procedural interoperability," in *2nd International ICST Conference on Electronic Healthcare, eHealth 2009, September 23, 2009 - September 25, 2009*, 2010, vol. 27 LNICST, pp. 124–130.

[11] P. De Vlieger, J.-Y. Boire, V. Breton, Y. Legre, D. Manset, J. Revillard, D. Sarramia, and L. Maigne, "Sentinel e-health network on grid: developments and challenges.," *Stud. Health Technol. Inform.*, vol. 159, pp. 134–145, 2010.

[12] P. De Vlieger, J.-Y. Boire, V. Breton, Y. Legré, D. Manset, J. Revillard, D. Sarramia, and L. Maigne, "Grid-enabled sentinel network for cancer surveillance.," *Stud. Health Technol. Inform.*, vol. 147, pp. 289–294, 2009.

[13] T. Glatard, C. Lartizien, B. Gibaud, R. da Silva, G. Forestier, F. Cervenansky, M. Alessandrini, H. Benoit-Cattin, O. Bernard, S. Camarasu-Pop, N. Cerezo, P. Clarysse, A. Gaignard, P. Hugonnard, H. Liebgott, S. Marache, A. Marion, J. Montagnat, J. Tabary, and D. Friboulet, "A Virtual Imaging Platform for multi-modality medical image simulation," *IEEE Trans. Med. Imaging*, vol. 32, no. 1, pp. 110–118, 2013.

[14] B. Koblitz, N. Santos, and V. Pose, "The AMGA Metadata Service," *J. Grid Comput.*, vol. 6, no. 1, pp. 61–76, 2007.

[15] S. Cipière, P. De Vlieger, D. Sarramia, D. R. C. Hill, and L. Maigne, "Development of a Metamodel for Medical Database Management on a Grid Network: Application to Health Watch and Epidemiology for Cancer and Perinatal Health," in *Cluster, Cloud and Grid Computing (CCGrid), 2012 12th IEEE/ACM International Symposium on*, 2012, pp. 892–897.

[16] M. A. Jaro, “Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa , Florida,” *Methodology*, vol. 84, no. 406, pp. 414–420, 1989.
[17] W. Winkler, “String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage.,” *Proc. ASA Sect. Surv. Res. Methods*, pp. 1184–1187, 1990.
[18] T. Berners-Lee, J. Hendler, and L. Ora, *The Semantic Web*. Scientific American, 2001.
[19] C. Bizer, T. Heath, and T. Berners-Lee, “Linked Data - The Story So Far,” *Int. J. Semant. Web Inf. Syst.*, vol. 5, no. 3, pp. 1–22, 2009.
[20] T. Gruber, “Ontology,” in *Encyclopedia of Database Systems*, L. Liu and M. T. Özsu, Eds. Springer US, 2009, pp. 1963–1965.
[21] O. Corby and C. Faron-Zucker, “The KGRAM Abstract Machine for Knowledge Graph Querying,” *Web Intell. Intell. Agent Technol. IEEE/WIC/ACM Int. Conf.*, vol. 1, pp. 338–341, 2010.
[22] O. Corby, A. Gaignard, C. Faron-Zucker, and J. Montagnat, “KGRAM Versatile Inference and Query Engine for the Web of Linked Data,” in *IEEE/WIC/ACM International Conference on Web Intelligence(WI’12)*, 2012.