Building a Semantic Health Data Warehouse: Evaluation of a search tool in Clinical trials

Romain Lelong, Lina F. Soualmia, Julien Grosjean, Mehdi Taalba, Stéfan J. Darmoni

Submitted to: Journal of Medical Internet Research
on: March 05, 2019

Table of Contents

Building a Semantic Health Data Warehouse: Evaluation of a search tool in Clinical trials

Romain Lelong, MSc; Lina F. Soualmia, PhD; Julien Grosjean, PhD; Mehdi Taalba, MD; Stéfan J. Darmoni, MD, PhD

Corresponding Author:
Romain Lelong, MSc

Phone:
Fax:
Email: romain.lelong@gmail.com

Abstract

Background: The huge amount of clinical, administrative and demographic data recorded and maintained by hospitals can be consistently aggregated into Health Data Warehouses (HDWs) with a uniform data model. In 2017, Rouen University Hospital (RUH) initiated the design of a Semantic Health Data Warehouse (SHDW) enabling both semantic description and retrieval of health information.

Objective: Our objectives were: first, to present a proof of concept of this SHDW, based on the data of 250,000 patients from RUH and second, to assess its ability to assist health professionals to select patients in a clinical trials context.

Methods: The SHDW relies on three distinct semantic layers: (a) a Terminology and Ontology (T&O) portal, (b) a Semantic Annotator and (c) a Semantic Search Engine and a Not Only SQL (NoSQL) layer to enhance data access performances. The system adopts an entity-centered vision which contrasts with the usually patient-centered vision adopted by existing systems such as Informatics for Integrating Biology and the Bedside (i2b2). This vision notably provides generic search capabilities able to express data requirements in terms of the whole set of interconnected conceptual entities that compose health information. We assessed the ability of the system to assist the search for 95 inclusion and exclusion criteria originating from five randomly chosen Clinical Trials from RUH.

Results: The system succeeded in fully automating 39.19% of the criteria and was efficiently used as a pre-screening tool for 72.97% of them.

Conclusions: The semantic aspect of the system combined with its generic entity-centered vision enables the processing of a large range of clinical questions. However, an important part of health information remains in Clinical Narratives and we are currently investigating novel approaches (deep learning) to enhance the semantic annotation of those unstructured data.

(JMIR Preprints 05/03/2019:13917)
DOI: https://doi.org/10.2196/preprints.13917

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **(a) Please make my preprint PDF available to anyone at any time (recommended).**
(b) Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
(c) Only make the preprint title and abstract visible.
(d) No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **(a) Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
(b) Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will rema
(c) Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="

Original Manuscript

Building a Semantic Health Data Warehouse: Evaluation of a search tool in Clinical trials

Romain LELONG[1,2] , Lina F. SOUALMIA[1,2,3] , Julien GROSJEAN1,3 ,
Medhi TAALBA[1] , and Stéfan J. DARMONI[1,3]

1 Department of Biomedical Informatics, Rouen University Hospital, 76031 Rouen Cedex, France
3 TIBS - LITIS EA 4108, Normandy University, Rouen, France
4 INSERM, U1142, LIMICS, Sorbonne University, 75006 Paris, France

Corresponding author: Romain LELONG, Phone: +33 232 888 898, e-mail: romain.lelong@gmail.com, Address: Département d'Informatique et d'Information médicale, Porte 21, Cour Leschevin, Hôpital Charles-Nicolle, 1 rue de Germont, 76000 Rouen, France

Abstract:

Background: The huge amount of clinical, administrative and demographic data recorded and maintained by hospitals can be consistently aggregated into Health Data Warehouses (HDWs) with a uniform data model. In 2017, Rouen University Hospital (RUH) initiated the design of a Semantic Health Data Warehouse (SHDW) enabling both semantic description and retrieval of health information.
Objective: Our objectives were: first, to present a proof of concept of this SHDW, based on the data of 250,000 patients from RUH and second, to assess its ability to assist health professionals to select patients in a clinical trials context.
Methods: The SHDW relies on three distinct semantic layers: (a) a Terminology and Ontology (T&O) portal, (b) a Semantic Annotator and (c) a Semantic Search Engine and a Not Only SQL (NoSQL) layer to enhance data access performances. The system adopts an entity-centered vision which contrasts with the usually patient-centered vision adopted by existing systems such as Informatics for Integrating Biology and the Bedside (i2b2). This vision notably provides generic search capabilities able to express data requirements in terms of the whole set of interconnected conceptual entities that compose health information. We assessed the ability of the system to assist the search for 95 inclusion and exclusion criteria originating from five randomly chosen Clinical Trials from RUH.
Results: The system succeeded in fully automating 39.19% of the criteria and was efficiently used as a pre-screening tool for 72.97% of them.
Conclusion: The semantic aspect of the system combined with its generic entity-centered vision enables the processing of a large range of clinical questions. However, an important part of health information remains in Clinical Narratives and we are currently investigating novel approaches (deep learning) to enhance the semantic annotation of those unstructured data.

Keywords: Data Warehousing; search engine; semantics; clinical trial; patient selection

Introduction

Background and significance

Hospitals maintain important health data that can be used in various contexts: first and foremost clinical care and then as data re-usability, clinical decision support systems [1], clinical research and cohort selection [2], education [3], [4], indicators, etc. However, the exploitation of

these data remains difficult for several reasons. First, the data is produced and maintained by different systems and health professionals and is consequently spread over multiple sources and even across multiple establishments. Second, the amount of data generated results in problematic management of Big Data. For instance, according to [5], [6], in the United States of America (USA), the health-care system alone reached 150 exabytes (1.5 × 10 20 bytes) in 2011 and will reach the yottabyte scale (10 24 bytes) in the near future. Moreover, the health data produced is of different nature, some data are natively structured (e.g. Diagnosis Related Group (DRG) codings, laboratory tests results, etc.) but an important part of medical information remains in unstructured free-text Clinical Narratives (CNs) [7] (e.g. Admission notes, history and physical reports, discharge summaries, radiology reports, pathology reports, etc.). This unstructured information is particularly relevant in the context of cohort selection tasks. However, in [8], the authors found that not only unstructured data were essential to resolve between 59% and 77% of some clinical trials criteria, but also that combining the use of structured and unstructured data enabled leverage of patient recruitment. In order to process unstructured data, the main approaches rely on Natural Language Processing (NLP) methods [9], [10]. The background knowledge, as represented in terminologies and ontologies (which describe the domain), plays a crucial role in any clinical NLP task [11]. A common approach to Information Retrieval (IR) in clinical unstructured text outside the basic full-text search consists in partially restructuring the original texts using semantic annotators (e.g. MetaMap [12]) that map words or expressions to concepts from domain knowledge databases.

Consistently aggregating all these scattered, big, complex and diversely structured data, is in fact, the role of HDWs. A HDW is defined as a grouping of data from diverse sources accessible by a single data management system [13]. This kind of data repository centralizes clinical, demographic and administrative data within a uniform and consistent data model. Many HDWs have been proposed worldwide. From a holistic point of view, the majority of these solutions provide aggregated data mainly focusing on patient data as a result. Furthermore, they do not necessarily allow full and independent visualization and retrieval of the different atomic entities conceptually composing the whole scope of clinical information (e.g. STRIDE [14], DW4TR [15]). This is nevertheless particularly important in an IR context as potential clinical questions and inquiries from health professionals are formulated in terms of their vision of the conceptual organization of data which derives from the actual patient management process. The Enterprise Data Trust [16] relies heavily on industrial solutions in order to cope with the huge amount of data. Many solutions also implement generic frameworks such as I2b2. This, however, implies concessions to conciliate the original conceptual representation of data with the data model required by the framework (e.g. The European Hospital George Pompidou HDW [17]). Furthermore, many standardized controlled vocabularies used to semantically describe health information do not always provide access to concepts in French and access to the data through these T&Os is not always provided for the whole set of data notably as far as unstructured data are concerned (e.g. EMERSE [18], STRIDE [14]).

In this context, in 2017, the Biomedical Informatics and Information Department of RUH initiated the conception and the development of a SHDW. The SHDW functionally relies on three independent semantic layers: layer one, the Cross-Terminological Health Terminologies and Ontologies Portal (HeTOP) [19], which provides the background knowledge necessary to semantically describe the health data; layer two a semantic annotator, the ECMT [20], [21] (Extracting Concepts From Multiple Terminologies), which enables the annotation of unstructured data; and layer three, the Semantic Search Engine (SSE) [22], [23], [24] and a web application interface Semantic Access to Health Information (ASIS), which enable access and retrieval of health data through different conceptual entities composing health information. A generic Entity–Attribute–Value (EAV) data model and a NoSQL layer (layer zero) enables data structuring, while preserving the original conceptual data model.

The objectives of this study were: first, to present a Proof Of Concept (POC) of this SHDW, based on the data of 250,000 patients from RUH and second, to assess its ability to assist health

professionals to select patients in a clinical trial context. Since November 2018, this POC has integrated all the data of 1.8 million patients from RUH.

Related work

Clinical data warehousing manages health data from hospitals and is a well addressed research field. Few generic frameworks and components exist. I2b2 [25], [26] is a datamart used in more than 200 hospitals worldwide. Initiated within the Massachusetts General Hospital in 2004, I2b2 was developed by the Harvard Medical School and is funded by the National Institutes of Health. It enables the integration of clinical and genomic data into an EAV model known as the star schema. I2b2 enables retrieval of patients' data using graphically built queries and enables querying of free-texts and coded information. Another example of a distributed solution is the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) [27]. This EAV model tends to standardize data from HDW at structure level as well as at representation level (i.e. terminologies, vocabularies, etc.).

In France, a few open-source solutions exist such as Dr. Warehouse [28] the CN–oriented data warehouse of Necker Children's Hospital but two solutions really stand out from the others: the ConSoRe system [29] which is used in some French Oncology Hospitals and the query engine Biomedical data Warehouse of the HOsPital whose French acronym is (eHOP) [30], [31] that is being deployed in six University Hospitals in Western France.

Due to the specificity of the data and their private and sensitive aspect, HDWs are specific systems which are used locally in Hospital Information Systems (HISs) rather than distributed and ready-to-use solutions and many specific HDWs have been developed worldwide in addition to the previously cited generic solutions:

The Stanford Translational Research Integrated Database Environment (STRIDE) (USA) [14] project focuses on a Clinical Data Warehouse supporting clinical and translational research. It was initiated in 2003 at Stanford University when the functionalities of I2b2 and caBIG were not considered optimal. An Oracle database and an EAV data model derived from the HL7 Reference Information Model (RIM) standard are used for data storage and representation. Several (mainly English) standardized terminologies are used to represent important biomedical concepts and their relationships (e.g. Systematized Nomenclature Of MEDicine – Clinical Terms (SNOMED–CT), RxNorm, 9th revision of the International Classification of Diseases – Clinical Modification (ICD–9–CM), Current Procedural Terminology (CPT), etc.). STRIDE provides hierarchical concept-based retrieval as far as structured data are concerned and provides full-text search access to more than six million CNs. The system is based on an n-tiered architecture and the querying of the data is distributed along several client applications whose scope targets patient cohort selection, cohort chart review, clinical data extraction, research data management and specimen data management. The querying is done graphically using drag and drop interface based components and returns aggregated data as a result without exposing individual patient data.

EMERSE (Michigan, USA) [18] is an Electronic Health Records (EHR)–oriented system exclusively providing full-text search capabilities into free-text clinical notes.

The Windber Research Institute (USA) developed the Data Warehouse for Translational Research (DW4TR) [15] system to support multiple translational research projects through highly-structured medical information represented in three dimensions (viz. clinical data, molecular data and temporal information). Data are collected into an Oracle Relational DataBase Management System (RDBMS) with an EAV data model and are subsequently hosted in an extensible data model that organizes it into a structure of hierarchical modules inherited from especially developed ontologies. It provides two graphical querying interfaces designed to provide aggregated data dedicated to data analysis (e.g. mean, standard deviation, counts, categorical data, chronological view, etc.).

The Enterprise Data Trust [16] is an industrial HDW initiated in 2005 at the Mayo Clinic

(50,000 employees, Rochester, Minnesota, USA). It collects patient care, education, research, and administrative data to support IR, business intelligence and high-level decision making. The Enterprise Data Trust strongly relies on industrial technologies (e.g. IBM InfoSphere Information Server, Teleran iSight & iGuard, SAP BusinessObjects, Sybase PowerDesigner, etc.) and enables integration and exploitation of important volumes of data (e.g. more than 7 million unique patients, 64 million diagnoses, 268 million test results etc.). The Architecture and functionalities of the Enterprise Data Trust rely on legacy technical components and long-standing governance works on data/metadata management, data modeling and standardized vocabularies. Those initiatives provide the HDW with a reliable organization of information on patient, genomic, and research data as well as querying capabilities for cohort selection and aggregate retrieval.

In 2008, the European Hospital George Pompidou (Paris, France) initiated a HDW [17] based on the I2b2 framework. It is strongly integrated in the clinical information system of the hospital which relies on several industrial solutions (e.g. ONECALL from McKesson, Act management (CPOE) from MEDASYS, integration platform from THALES). The core HDW infrastructure relies on an Oracle database for storage (1.2 million patients, 1 million stays) and the I2b2 framework for data representation. Several client applications are connected to the system to provide technical access to the data but mainly use I2b2 client as far as researchers are concerned. The SMart EYE DATabase (SMEYEDAT) [32] is an ophthalmologic-specialized HDW developed at the University Eye Hospital in Munich in Germany. SMEYEDAT is based on a Microsoft SQL database updated daily from the HIS and uses a star-like patient-centered data model for data representation. The Qlikview [33] (Qliktech, Radnor, Pennsylvania, USA) tool was implemented as an analytic tool to visualize and explore patient data. This interface enables patient selection using criteria and views specific to the domain.

Material and methods

Overall, the first prerequisite pertaining to the design of a HDW-based system is the extraction of data from the HIS. This can be basically achieved in two ways: (a) by setting up a data stream from the production environment (or a replicated database) to the HDW data storage component, or (b) by using Extract–Transform–Load (ETL) scripts. As far as the SHDW is concerned, ETL scripts are used. The following section describes the targeted sources of data of the HIS of RUH.

Data source

Since 1992, RUH has collected and maintained patient identity data (e.g. name, date of birth, gender, etc.), clinical data (e.g. biological test results, medical procedures, visit records, letters, discharge summaries, etc.), administrative data, and less frequently omic data [34]. The data are produced by different sub-systems and applications of the Information System (IS) of RUH. A sub-system called CPage Dossier Patient (CDP) partially aggregates some of this important data such as laboratory results, DRGs, procedures and clinical documents. Other data remain in other sub-systems that have to be accessed separately. Overall, RUH maintains the data of 1.8 million patients which represents approximately 14.4 million visits, 11.9 million clinical documents (free-texts recorded since 2000), and 107 million unitary laboratory tests (e.g. Na, K, etc.) (Recorded since 2004). Since November 2018, the SHDW POC presented in this study includes the whole set of data. However, this study is based on a randomly chosen subset of data from 207,357 patients, 1.7 million visits, 671,442 clinical documents and 14.2 million unitary laboratory tests. ETL scripts are used to incorporate data from the production environment repositories into an Oracle database. Figure 1 summarizes included data according to their specific domain.

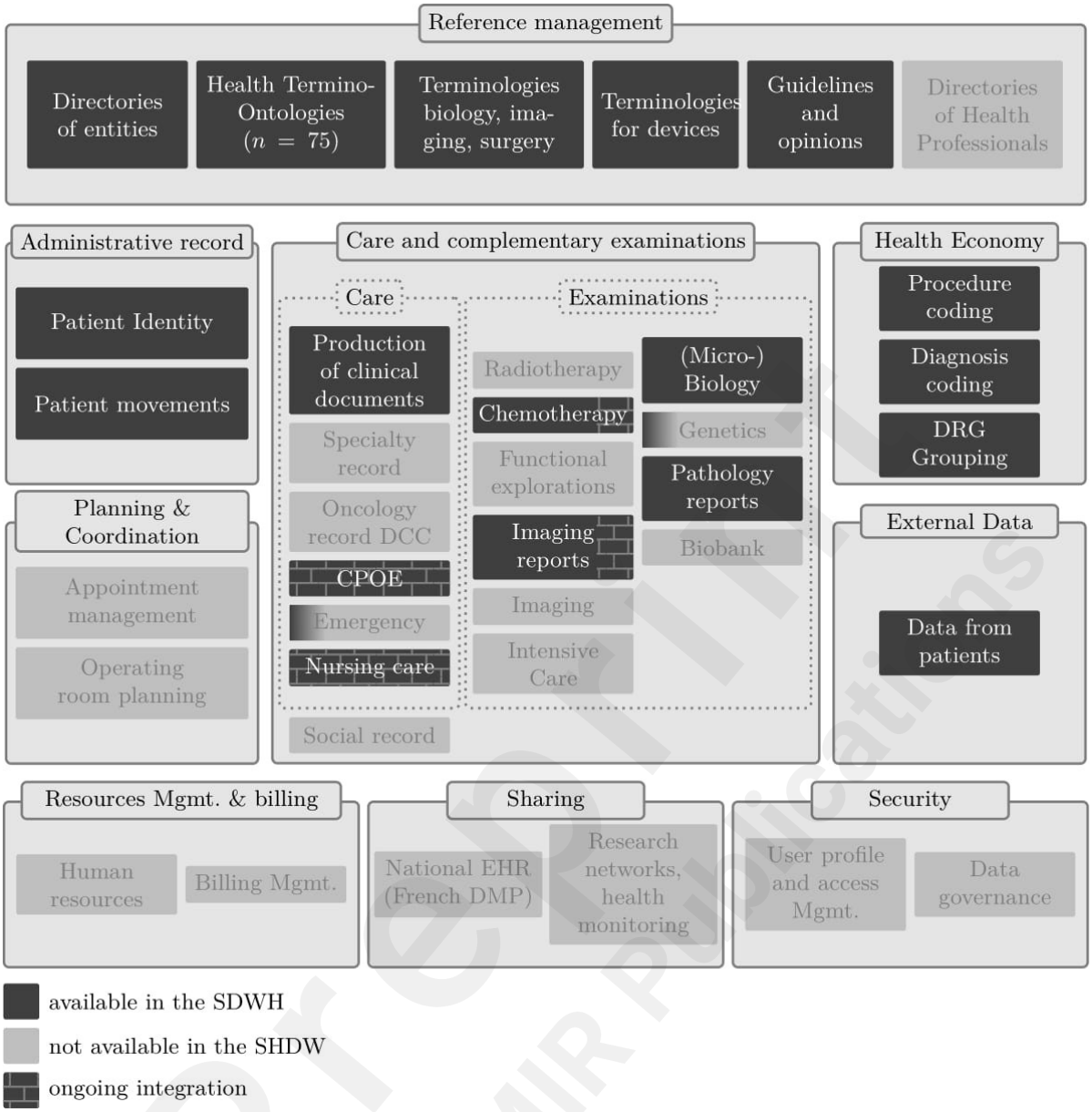


Fig. 1: Functional coverage of the SHDW in terms of data according to each domain (viz. Reference management, Administrative record, Care, Examinations, Health economy, Planning & Coordination, External Data, Resource management & Billing, Sharing and Security). Data already included in the SHDW are represented by a dark gray opaque background, whereas a light gray background indicates that data are not included and not planned to be in the short or medium term. Background partially or totally covered with bricks corresponds to data for which inclusion is in progress or is planned in the short term or medium term.

The SHDW currently focuses on clinical data and more broadly on health data according to a patient-centered strategy. In addition to structured patient data, the different data pertaining to multiple admissions and events at RUH are collected (e.g. diagnoses, biology, procedures, movements, etc.). The reference controlled vocabularies (i.e. reference management domain) necessary to the understanding of those data are notably widely collected and maintained. In contrast, pure management and administrative data such as appointment and planning data, billing data, data governance are not likely to be included in the short term. All those data are integrated into a modular architecture which is described in the following section.

Overall Architecture of the SHDW

Much health information remains in CNs [7]. The 11.9 million clinical documents in French of RUH consequently play a strategic role in the context of the SHDW. Since its creation in 1995, our research team has strongly investigated French IR research domains through T&Os (and more broadly Knowledge Organization Systems (KOSs)) which has led to the development of several search tools mostly dedicated to IR from documentary and bibliographical resources [35], [36]. However, the complexity of the clinical data and more broadly of SHDWs as a whole required the pooling of several of these acquired skills and tools. The SHDW enables the semantic retrieval of health data in French based on several T&Os and consequently relies on two data sets: a domain knowledge database and a health database maintaining clinical and patient data. The functionalities of the SHDW are ensured by the collaboration of three distinct layers, where each layer consumes data from the above(s) layer(s) (see Figure 2):

1. The Cross-Terminological Health Terminologies and Ontologies Portal (HeTOP) [19],
2. The semantic annotator Extracting Concepts From Multiple Terminologies (ECMT) [20], [21], [37],
3. The Semantic Search Engine (SSE) [22], [23], [24].

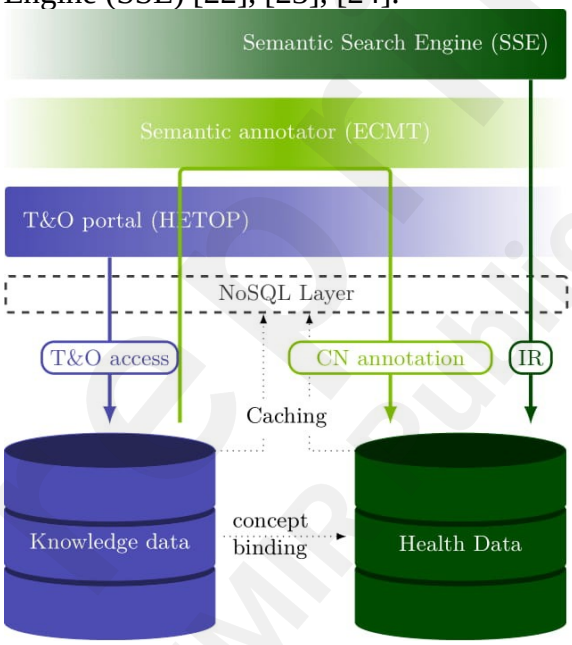


Fig. 2: Functional architecture of the SHDW which provides semantic IR functionalities form clinical data. The two data repositories “knowledge data” and “Health Data” respectively maintain the reference KOSs and the health data pertaining to the SHDW. These data are accessed through a NoSQL layer by the three distinct components HeTOP, ECMT and the SSE which, each, operate over a different range of data.

The HeTOP provides access to domain knowledge data. The ECMT matches words and expressions in natural language to domain knowledge concepts included in the HeTOP. In fact, ECMT enables the extraction of semantic information from unstructured data. Its functional scope consequently lies between domain knowledge data and clinical data. Together, the two components HeTOP and ECMT serve as a base for the semantic description of the clinical information in a computer–processable form. In contrast, the SSE, and the coupled web application, are dedicated to information retrieval tasks on health data by using this extracted semantic description.

Considering the amount of data, access to health data and domain knowledge data is made through a NoSQL layer [22] based on the Infinispan solution, an In Memory Data Grid (IMDG), on one server with 192 cores and 1 T B (i.e. 10 12 bytes) of Random–Access Memory (RAM) allowing vertical

scaling.

Each of these layers is functionally and technically detailed below.

Semantic representation

This section describes data and the methods for its storage and modeling; and presents ECMT which enables the link between knowledge data and actual clinical data.

Domain knowledge data

The HeTOP provides cross-lingual access to concepts originating from 75 T&Os. A set of 2,639,620 concepts and 10,735,905 terms are available mainly in English and French. However, 32 languages are available overall. Some of the T&Os have been partially or totally translated into French (e.g. SNOMED 3.5 (52.3%), MeSH descriptors (100%), National Cancer Institute Thesaurus (NCIt) (53.35%), Online Mendelian Inheritance in Man (OMIM) (79.67%), Human Phenotype Ontology (HPO) (72.19%), Radlex (22.1%), etc.). More broadly, 50% of the 2.64 million concepts accessible through the HeTOP are provided in French and 19.1% of the 10.74 million terms have a French translation. T&Os from the HeTOP come with their original sets of hierarchies and semantic relationships, but also with additional cross-terminological exact–, broader–to–narrower and, narrower–to–broader mappings performed manually or supervised by our health professionals at RUH.

As a primary use, the different concepts are bound to the different clinical entities (e.g. procedure and DRG codings, CN annotations, etc.), thus allowing a semantic description of the clinical information to be obtained. This allows both the refining and the broadening of IR tasks by exploiting the underlying semantic network formed by the concepts (i.e. by controlling the granularity and the depth with which this semantic network should be browsed in search processes).

Health Data Model

Health data are stored in a PostgreSQL [38] relational database. A generic and very adaptable physical EAV data model used in [34], [39] is used to integrate the data. This model structures the information in terms of objects, attributes, relationships and thus, defines an underlying Entity-Association modeling of the data. It enables the preservation of any original conceptual organization of the information without altering the physical data model and consequently maintains the desired vision of the data at conceptual level. A partial and simplified representation including the main entities and a limited number of relationships and attributes of the conceptual data model used for this study is shown in Figure 3. This model is used on a daily basis to satisfy the information needs of the different health professionals of RUH.

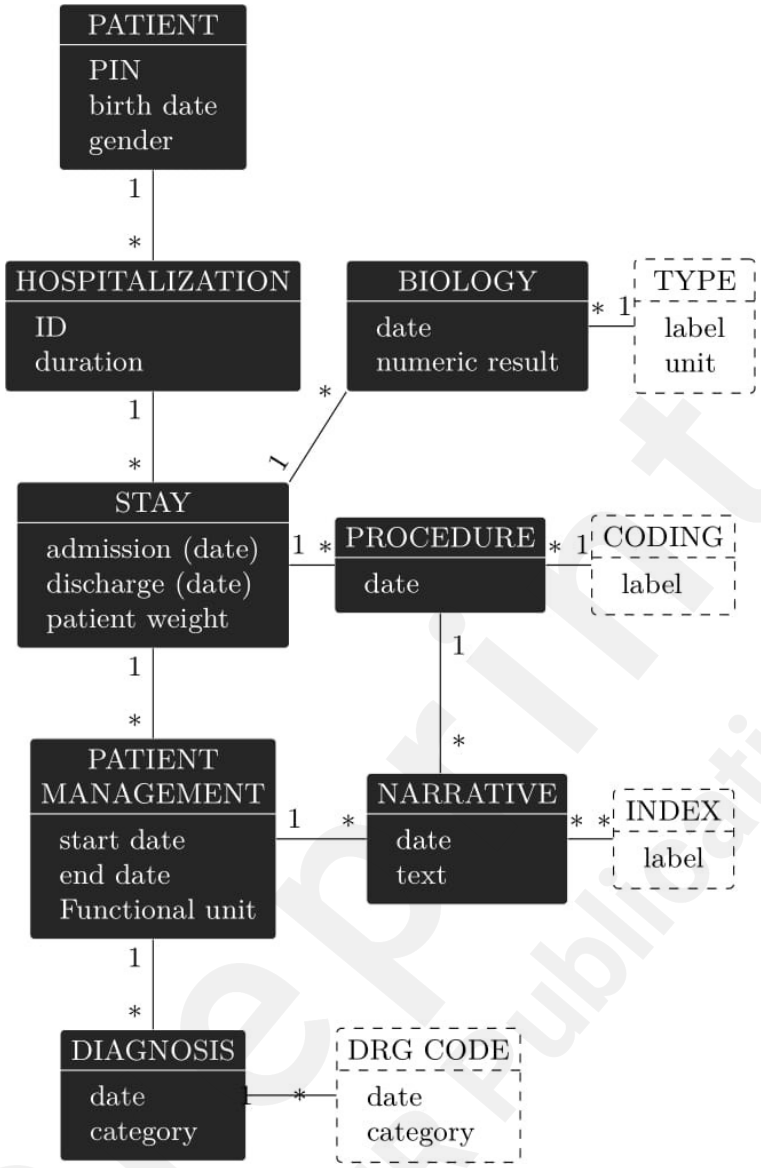


Fig. 3: Partial Conceptual Model of the SHDW represented as a directed and attributed graph. Entities corresponding to elements from T&Os are represented with dashed outlines.

Semantic annotator

The semantic annotator ECMT [20], [21], [37] matches the natural language words and expressions to the domain knowledge concepts included in the HeTOP. A bunch of semantic annotators have been proposed for English texts. Recently, Aurélie Névéol et al. [40] performed a literature review on NLP tools in health in languages other than English. In this study, French was the most studied language followed by German and Chinese. Nevertheless, most of the existing semantic annotators, usually extract concepts from the Unified Medical Language System (UMLS) Meta-thesaurus [41] (e.g. MetaMap [12]) or from mainly English T&Os such as the SNOMED–CT terminology (e.g. Text Analysis and Knowledge Extraction System (cTAKES) (SNOMED–CT, RxNorm) [42], NCBO Annotator [43]). French is little represented in the UMLS [44]. The HeTOP includes only 17 KOSs of the 2017 edition of the UMLS. However, the UMLS only manages 11 resources providing concepts in French and among the 978,233 Concept Unique Identifiers (CUIs) of the UMLS included in the HeTOP, only 143,762 (i.e. 14.7%) concepts in French originate from the UMLS. In contrast the HeTOP provides access in French to 428,854 (i.e. 43.8%) of them (almost three times more than the UMLS). Technically, the ECMT relies on the bag-of-word method for concept matching but also provides pattern-matching functionalities in particular to deal with

negation and contextual information such as numerical values in CNs. Functionally, the ECMT is used in query building processes to match user inputs to accurate sets of concepts but plays a major role in CN indexing.

In this study, 11,928,168 CNs of RUH have been indexed using ECMT over 55 terminologies available in French from the HeTOP server. This indexing process resulted in a total of 5,043,731,628 annotations. Some of the most redundant concepts were found clinically irrelevant and a manual filtering process was applied based on the top 5,000 most frequent medical concepts (e.g. the 27 million annotations with the concept university hospital were considered as irrelevant as the information was present elsewhere in the SHDW). A total of 2,087,784,055 annotations were retained after the filtering process. This set of semantic annotations served as a basis for the SSE in the semantic retrieval process.

Semantic Retrieval

Access to the data is allowed by a NoSQL layer before processing by the SSE.

NoSQL Layer

Due to the considerable amount of health data that needs to be retrieved and the well-known limitations of the RDBMSs in terms of scalability, a NoSQL layer was designed in order to interface access to all the data and improve data access performances. This layer is based on the IMDG Infinispan [45], [46]. It is a Java NoSQL solution that uses key–value hash tables as storage structure which allows efficient recovery of unitary data via the associated keys. Moreover, the hash tables are stored in memory and not on disk which leverages access times.

The NoSQL layer was conceived in a generic way to mirror the EAV data model used to structure health data (i.e. Java Object used as values in hash tables mimics the objects and relationships of the relational databases). This generic NoSQL layer consequently preserves the conceptual data model of health and knowledge data implicitly drawn by the EAV data model. A more detailed description of this layer and an overview of performance gain compared to relational RDBMSs systems are presented in [22].

One of the major drawbacks of Infinispan, and more generally of many in-memory key-value stores, is that no comprehensive query language is provided as opposed to the Structured Query Language (SQL) for RDBMS. Complex querying capabilities must be fully implemented from the basic Application Programming Interface (API) (i.e. obtaining and removing a value of a specific key) proposed by this kind of solution. In particular, in this study, neither join nor reverted index functionality is natively fully provided by Infinispan and requires respectively the maintenance of custom maps and the use of Lucene [47] tools to enable the search from concrete values (i.e. text, numerical and data values).

Semantic Search Engine (SSE)

The main purpose of the SSE is to deal with the multiplicity and the diversity of conceptual entities inherent to clinical and patient data (e.g. patients, stays, CNs, diagnosis, biological tests, etc.). Overall, the entire set of data originating from the SHDW can be seen as a comprehensive oriented attributed graph that can be queried by the SSE. This complex data structure is very different from documentary and bibliographic IR context, which has been studied these two last decades [36], [35], and which involves a limited number of entities and relationships and where data is classically more "flatly" structured with a limited depth (basically one resource entity possibly surrounded with several other entities such as an author or an editor entity). The SSE was designed to concentrate on semantic retrieval by allowing navigation through the semantic networks, not only included in the T&Os, but also those representing clinical data conceptual entities. From a more clinically coherent point of view, the data of the SHDW can be organized in four levels: (a) patient level corresponding to patient identity information, (b) hospital level defining the sources of

information (this level is currently not implemented as all the data originate from RUH), (c) stay level which defines much organizational and administrative information about the health care process and (d) health level enabling group medical procedures, biological tests, etc. [24]. As a HDW can be used in various contexts (e.g. health care, health research, secondary use of health data), access and search capabilities of the full scope of those types of information must be provided. Technically, the SSE is a Boolean and Entity-Oriented Search Engine. It enables the retrieval and display of data at any of the previous clinical levels. As mentioned in section above, the NoSQL key–value store used to interface data does not provide proper querying solutions. The SSE consequently relies on a specific query language based on formal grammar. It enables the expression of queries targeting any of the different conceptual entities selected through constraints focusing on attribute values and other linked entities [24]. The SSE is used through a Web application that enables the querying of clinical data using forms and string-based queries. This application is described below.

The Semantic Access to Health Information Web application: ASIS

The SSE provides a powerful means to select data using textual logical queries. To bypass the complexity of the query language syntax, we designed a user-friendly Web application known as ASIS. It enables the retrieval of clinical data by means of a form which generates a SSE–processable logical-based query. The clinical data selection process is divided into four numbered steps clearly identifiable on the graphical interface. Step 1 consists in building a set of constraints related to any desired entity of interest as patient, diagnosis (DRG), biological tests, stays, procedures, records (CNs), drugs, medical devices (see Figure 4). Constraints are built via: (1) the choice of the entity of interest, (2) the choice of the targeted metadata of this entity as date of birth (patient), gender (patient), type of biological test (biological test), date (procedures, biological tests, stays, etc.), coding (diagnosis, records and procedures, etc.), and finally (3) the entry of the inputs corresponding to the chosen entity and metadata as male/female for the gender metadata of a patient constraint, the desired numeric value for the biological test constraint, etc. To facilitate the reading of the interface, each type of entity is represented using a specific color (e.g. green for patient, red for diagnoses, green-cyan for biology, blue for stays, etc.). As the SHDW was conceived in order to focus on semantics, many metadata inputs concentrate on selecting T&Os and concepts by the user from fields auto-completed to facilitate the selection. For instance, constraints 2 and 3 enable retrieval of CNs indexed with the different concepts referring to type 1 and 2 diabetes (Figure 4). Step 2 consists in aggregating the constraints defined in step 1 into a Boolean query. In this form, constraints are represented as colored buttons showing their IDs, a short description of them and the numbers of results of the sub-queries corresponding to them. A click on a constraint button enables the visualization of the partial results corresponding to the constraints. The step 2 sub-form editable area enables the composition of the query using parentheses, Boolean operators (AND, OR, NOT), and the defined constraints that can be selected using an auto-completion feature. Nevertheless, the step 1 sub–form enables the predefinition of a basic Boolean query skull that is on–the–fly reported in step 2 and that can be later manually modified or left untouched in step 2. Step 3 consists in choosing the desired output entity type classified according to the three clinical information levels: patient, stay and health level. The choice of an entity type generates a button similar to constraint buttons in Step 4.

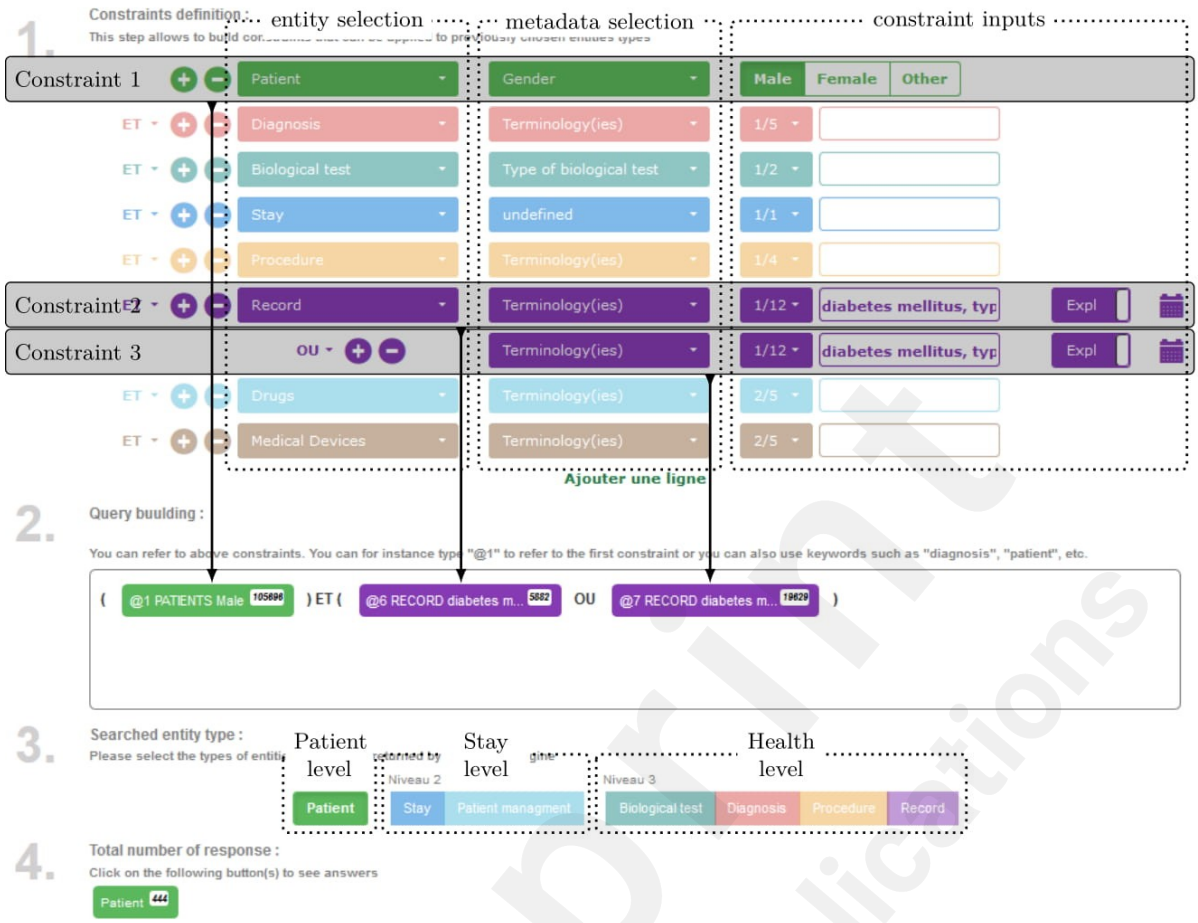


Fig. 4: The interface of the Semantic Access to Health Information (ASIS) web application and its four steps (1.) Definition of constraints, (2.) Composition of a Boolean query from atomic constraint defined in step 1, (3.) Selection of the desired output entity according to its clinical coherent level and (4.) Visualization of the results.

Evaluation methodology

Five clinical trials of RUH, consisting in a total of 95 criteria (36 inclusions/59 exclusions), were randomly selected. The ability of the system to automate patient recruitment was then assessed on each of those criteria, taken independently from both the originating clinical trial and the overall context of the clinical trials. For each criterion, a search strategy was designed. Each of them required the collaboration of a medical doctor (to clinically interpret the criteria and identify the different sources of information to target) and a computer engineer to master the ASIS tool querying process. The search for a single criterion can be done through multiple search directives (i.e. ASIS constraints) targeting different sources of information (i.e. entities). Those search directives are then aggregated into a single search strategy (i.e. a global query) by combining the different search directives using Boolean operations and relational links between the entities corresponding to each search directive. The different constraints that could reduce the accuracy of each search directive were also investigated. In this study, three characteristics are finally considered and linked to each other in order to more precisely identify the different capabilities and limitations of the system: (1) the global support level of the criteria by the system, (2) the targeted source of information, and (3) the obstacle and barriers that tend to lower the effectiveness of the search.

Each of the criteria was therefore classified into six levels of global support by the system:

Fully-supported level represents criteria that can be fully automated by the system with a search strategy that retrieves all and only the resources that fulfill the exact requirements of the

criteria, e.g. “18-year-old patients” and “patients with a neutrophil level below 1700/mm 3 ”, etc.

Accurately-supported level represents criteria that are based on consistently-recorded data in the IS and on reliable search. The result may, however, possibly include some irrelevant resources depending on the choices made in the elaboration process of the search strategy mainly about the choice of concepts to search and the exploitation of their semantic networks, e.g. “patients with hepatitis B or active hepatitis C” and “patient with acute kidney failure”.

Broadly-supported level represents criteria for which the search results in a lack of precision (i.e. inclusion of irrelevant resources or absence of relevant ones). These criteria can only be reliably answered partially. This implies a broadened search of the core requirement of the criteria and a manual post-filtering of the result and/or supervision by a health professional to decide whether, or not, the retrieved resources effectively fit the criteria, e.g. “patient with an evolving organic digestive and/or inflammatory pathology” and “patient with a badly regulated cardiac rhythm disturbance”.

Inaccurately-supported level represents criteria that cannot be searched precisely enough (both technically and in terms of data) in order to fulfill the core requirement of the criteria or to systematically provide consistent results, e.g. “pregnant woman or breastfeeding mother” and “patient admitted for a stomach hemorrhage resulting in a favorable evolution without surgery during the hospitalization”.

Non-supported level represents criteria for which the system fails to properly select the relevant resources or for which a search strategy is hardly feasible, e.g. “patient with a regular consumption of liquorice or derived substances” and “abdominal pain presenting once a week during the last 3 months associated with two of the following criteria [...]”.

Not-Applicable (N/A) and Instruction level represents criteria that either does not connect to the medical domain or that corresponds more to instructions than real requirements, e.g. “patient participating in another clinical trial” and “contraception will be required during the treatment”, etc..

Six types of source of information were identified: (P) Patient structured data as age, gender, etc., (D) DRG data corresponding to structured diagnosis coded with the 10th revision of the International statistical Classification of Diseases and related health problems (ICD–10), (S) stay data and other organizational structured data as medical units, (B) biological structured data, (N) CN unstructured data as full-text and/or Automatic Indexing including drug data, and (I) for information that is not within the scope of RUH IS.

Finally, the different obstacles or barriers that lower the effectiveness of the search were recorded for each atomic search directive and were distributed among six categories: $(\varnothing)$ for search directives that are free of any obstacles, (d) for data obstacles corresponding to inconsistently provided or insufficiently accurate data from the IS, (s) for difficulties to perform an accurate search in CN or DRG data as complex information search, (t) for technical limitations of the system as chronological querying handling or search for quantitative values in CNs (partially implemented), (c) for subjective and/or generic criteria implying the interpretation or value judgment of a Health professional, and (e) when it is necessary to meet the patient to complete the criteria.

The global support levels of criteria observed in this study are first detailed in section “Global support of criteria”. A two-sided Wilcoxon signed-rank statistical test is used to examine the different levels of support of inclusion vs. exclusion criteria. The three sets of scores detailed in this methodology section are then matched with each other in “Observed sources of information and limitations” in order to objectify and identify the concrete abilities and limitation of the system.

Results

Global support of criteria

As a primary and holistic result, the support levels of the 36 inclusion criteria and the 59 exclusion criteria from the five randomly selected clinical trials of RUH are shown in Table I. The percentage of criteria for each of these levels was recorded.

According to the methodology used to classify criteria, Three of the six levels of support, “Full”, “Accurate” and “Broad” could be considered as contributing to cohort selection. Taken together, the system was consequently able, at least partially, to automate the search for 41.67% of inclusion criteria vs. 66.09% of exclusion criteria. This lower support of inclusion requirements tends to affect more the ability of the systems to assist cohort selection tasks. This is mainly due to the fact that clinical trials usually rely on fewer inclusion criteria than exclusion criteria, which make inclusion requirements more critical prerequisites. In fact, among the five clinical trials used in this study, the number of exclusion criteria exceeded the number of inclusion criteria by 20.14% on average.

A fairer and more reliable measure was also investigated. N/A criteria represented 22.1% of the criteria of this study. This type of criteria is not in the scope of an HDW-based system and should consequently be set aside. Moreover, 66% of these criteria were attributed to inclusion criteria. Excluding “N/A” criteria, the percentages of criteria for which the system was able to contribute increased to 68.18% for inclusions and 75% for exclusions. When only considering support levels that did not imply post-filtering (i.e. only “Full” and “Accurate”), 40.90% of inclusion criteria could be answered compared to 38.46% of exclusion criteria.

A two-sided Wilcoxon signed-rank statistical test was used to compare the levels of support of inclusion vs. exclusion criteria. In order to perform that test, a mean support score was calculated for each subset of inclusion or exclusion criteria of each clinical trial. The calculation of these means was made by assigning to each support level a score from 0 to 100. The test were not significant with an homogeneous distribution of the scores but a trend was observed towards better support of inclusion criteria compared to exclusion criteria for distributions that weighted "Full" criteria twice as much as the others. The mean support score of inclusion criteria was constantly greater than the mean support score of exclusion criteria for each clinical trial. The tests resulted in observed statistics T = 15 with a p–value equal to P = .06 which, even if slightly greater than the 5% significance level, suggested a better support of inclusion criteria.

Support	**Inclusion criteria**			**Exclusion criteria**			**Total**		
Level	n	p (%)	I (%)	n	p (%)	I (%)	n	P (%)	I (%)
Full	6	16.67	[4.5, 28.8]	5	8.47	[1.4, 15.6]	11	11.58	[5.1, 18.0]
Accurate	3	8.33	[0.0, 17.4]	15	25.42	[14.3, 36.5]	18	18.95	[11.1, 26.8]
Broad	6	16.67	[4.5, 28.8]	19	32.20	[20.3, 44.1]	25	26.32	[17.5, 35.2]
Inaccurate	4	11.11	[0.8, 21.4]	6	10.17	[2.5, 17.9]	10	10.53	[4.4, 16.7]
None	3	8.33	[0.0, 17.4]	7	11.86	[3.6, 20.1]	10	10.53	[4.4, 16.7]
N/A	14	38.89	[23.0, 54.8]	7	11.86	[3.6, 20.1]	21	22.10	[13.8, 30.4]
Total	36	100.00		59	100.00		95	100.00	

Table I: Number (n), percentage (p) and 95% confidence interval of the percentage p (I) of criteria for each support level and type (inclusion or exclusion).

Observed sources of information and limitations

The results obtained in Table I should nevertheless be regarded more qualitatively than quantitatively as regards the 95% confidence intervals which show widths of 20.37% on average (15.08% when inclusion and exclusion criteria are taken together). In order to achieve that goal, both the targeted sources of information and the observed limitations for each support level of each criterion were investigated.

The support level of the criteria according to the combination of information sources required to search them are displayed in Table II.

Support Level	*P*	*S*	*B*	*D*	*N*	*I*	*SN*	*PN*	*DB*	*BN*	*DN*	*NI*	*PSD*	*SDN*	*DNI*
Full	4	0	7	0	0	0	0	0	0	0	0	0	0	0	0
Accurate	1	0	0	8	0	0	0	1	3	0	5	0	0	0	0
Broad	0	2	2	6	7	0	1	0	0	1	4	0	0	2	0
Inaccurate	0	0	0	0	5	0	0	0	0	1	1	0	1	1	1
None	0	0	0	0	2	3	0	0	0	0	3	1	0	1	0
N/A	0	0	0	0	0	21	0	0	0	0	0	0	0	0	0
Total	5	2	9	14	14	24	1	1	3	2	13	1	1	4	1

Table II: Number of criteria of each support level according to the combination of sources necessary to search them.

Setting aside N/A criteria, 63.51% of criteria could be answered using a single search directive (i.e. by exploiting a single source of information) against 36.49% that required combined search directives. The calculation of the mean scores of level of support of these two groups of criteria resulted in scores between “Accurate” and “Broad”. 23.40% of single search directives versus 0% of combined search directives concerned fully-supported criteria.

The different sources of information were not uniformly distributed. Patients (P) and stays (S) structured data were involved in the search for only 7.37% and 8.42% of the criteria, respectively. In contrast, the top two sources of information, CNs (N) and diagnoses (D), were involved in the search for 38.95% and 37.89% of criteria, respectively.

The percentages of involvement of sources of information and the percentages of involvement of observed limitations for each support level are presented in Figure 5.

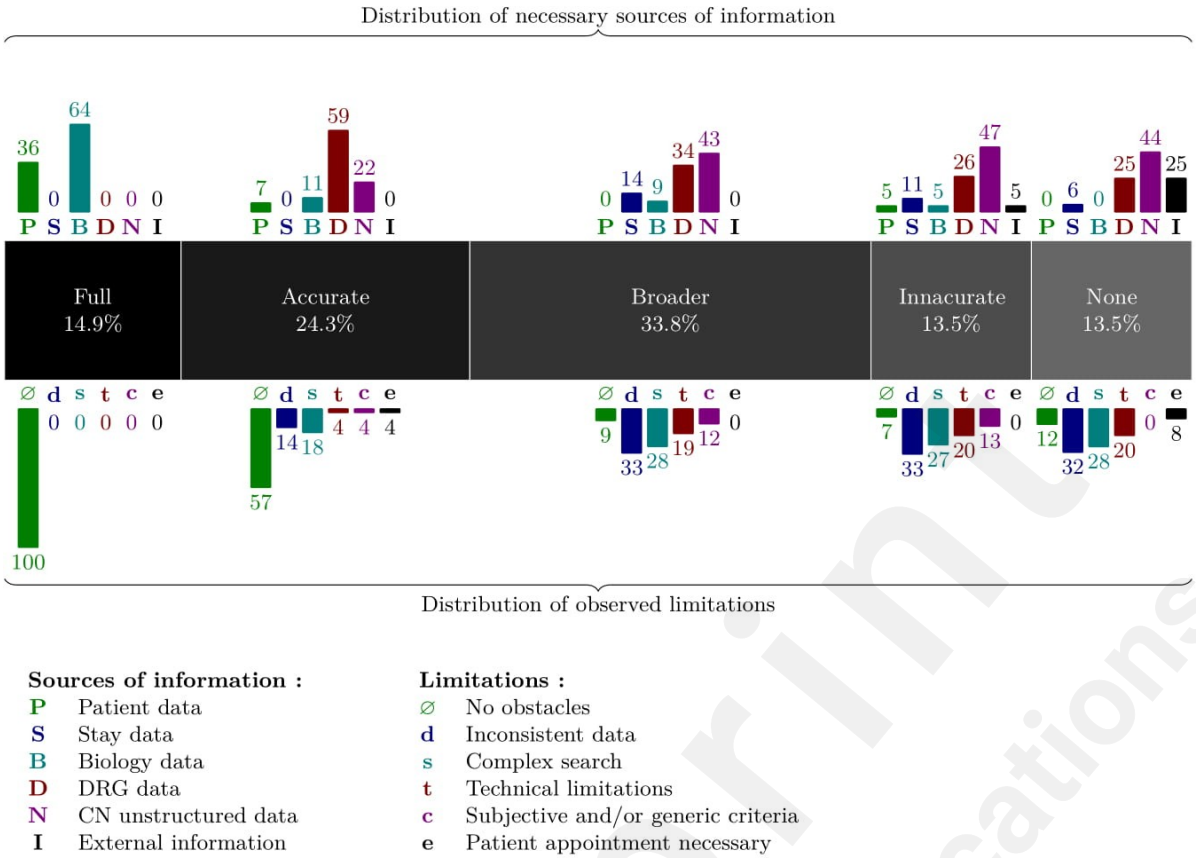


Fig. 5: The central gray band gives the percentage of criteria of each support level excluding N/A criteria. The upper bars, show, for each support level, the percentages of involvement of each source of information in the search of criteria. The lower bars, show the distribution (in percentage) of the different obstacle categories identified as lowering the effectiveness of the search of criteria.

Only continuously provided and fully-structured data were used to answer fully-supported criteria. The only sources of information used were patient structured data (P) and biological data (B). Fully-supported criteria were consequently based on very precise characteristics not subject to errors or ambiguity and relying on numeric or symbolic data such as “female or male patient of 18-75 years old” and “patient with glycated hemoglobin ≤ 6.5% or ≥ 8%”.

Accurately-supported criteria were mostly searched in DRG data (D). In practice, these criteria either rely on a single source of information (e.g. “HIV-positive patient”, “type 2 diabetic subject”, etc.) or on the combination of data consistently provided or properly coded, as “known active hepatopathy, [...], transaminase and/or alkaline phosphatase levels twice the normal level of the laboratory.” (DRGs and Biological data), or “men aged 18-70 years or women aged 18-70 years in menopause” (patient data and CNs).

From a holistic point of view, we observed that DRGs (D) and CNs (N) were the two major sources of information used. Both were involved in the search strategy of approximately 38% of criteria (57.9% if taken together). The support of the criteria by the system decreased as the exploitation of CNs (N) took precedence over DRG data (D). The exploitation of unstructured data (N) was consequently considered as the major challenge for the SHDW in this study.

The search accuracy obstacles (s) category represented 20.2% of all obstacles and 84.4% of these obstacles were attributed to CN search (N). However, the exploitation of the semantics (i.e. synonyms, hierarchical and semantic relationships) through the automatic indexing of CNs (N) by the ECMT and the ability of the SSE to combine multiple search directives (using Boolean operators) enabled a broad search support of 26.3% of criteria. Even when post-filtering was required, the system could be used effectively as a pre-screening tool. For instance, the search for the

criterion “Patients with severe heart failure (including NYHA Class III and IV)” was done through the search for “heart failure” in DRG data (D) and the search for “NYHA Class III” and “NYHA Class IV” in CN data (N). Separately, it resulted in 11,880 diagnoses and 3,311 CNs. The combination of both search directives into a single search enabled the recruitment of only 36 patients.

Data inconsistency (d) was also a major challenge (22.1% of all obstacles) and was found across different sources of information including DRGs (23%) and Stays (S) (17%). Many data are sparsely recorded in the IS even outside CNs (e.g. weight of the patient as structured data for each stay (S), diet plan in CNs, hypersensitivity to substances in DRG data (D), etc.).

This lack of consistency of information tends to explain the focus on CNs (N) of inaccurately and none supported criteria. In practice, these criteria suffer from the association of concurrent obstacles often including a data consistency obstacle (d). For instance, both data inconsistency (d) and technical limitations (t) were found for the non-supported criteria “regular consumption of alcohol exceeding 60g per day”. Information on alcohol consumption was in fact not provided consistently in CNs (N) and technically it would have required: (a) the extraction of a quantitative value from CNs and (b) the processing of this value as data (partially implemented). As another example, the criterion “Patient with a Creatinine Clearance ≤ 50 ml/mn according to Cockroft formula” was inaccurately managed by searching instead for the biological tests of creatinine higher than 100 µmol/L. The criterion strongly relies on specific calculation functionalities not provided by the system and based on sparsely provided data (e.g. weight of the patient).

As regards efficiency, the NoSQL layer used to access the data gave access performances that were considered extremely satisfactory. Based on the data of 250,000 patients, each of the search directives used for this study took less than 2 seconds. As far as the POC integrating the entire patient dataset (1.8 million patients) is concerned, similar performances were observed except for some specific queries targeting and returning huge amounts of biological tests which exceeded one minute.

Discussion

To our knowledge, no formal evaluation of the criteria for clinical trial inclusion and exclusion has been performed using a SHDW. The system based on a SHDW presented in this study could be successfully used to fully automate 39.19% of the criteria. Moreover, with a limited post filtering process, it could be efficiently used as a prescreening tool for 72.97% of the criteria.

A slightly better level of support of inclusion vs. exclusion criteria was observed for each clinical trial. When considering all the criteria of a single clinical trial, there was an increase in the overall recruitment of patients.

However, there are still many criteria (i.e. 27%) that cannot be searched or that can only be partially searched by the system. Several mishandled sources of information along with specific limitations of the system are apt to explain these results. DRG and CN data remain an important source of information for none-supported or inaccurately-supported criteria. Consistent and systematic recording of necessary information in the IS is not always performed. Furthermore, this information often resides within unstructured CNs. Consequently, more advanced methods of information extraction from those unstructured data such as the extraction and exploitation of quantitative values from CNs (which is only partially implemented in our system) or the on-the-fly computation of relevant measures (e.g. Body Mass Index) could drastically improve the capabilities of the system.

Furthermore, despite the growing interest in statistical machine learning methods, rule-based NLP methods remain predominant as far as clinical information extraction is concerned mainly due

to their potential of interoperability and interpretability [9], [10]. Nevertheless, since 2018, our team has engaged new research on the semantic annotator ECMT in order to investigate the development of a hybrid approach between Bag-Of-Words Algorithm and Word embeddings.

The general philosophy of the system relies on a generic representation of clinical information. It enables the independent search and visualization of each conceptual entity (e.g. patient, biology, diagnoses, CNs etc.) that composes the entire health information of the SHDW. We believe that this entity-oriented vision gives added-value to the IR systems dedicated to HDW compared to existing solutions, such as I2b2, which usually adopt a patient-centered vision and provide the user with aggregated data and lists of patients as a result. Notably, the system allows the search to be conducted in an iterative manner by visualizing the search of each entity before aggregating all of them into a comprehensive and coherent search.

In addition, the underlying powerful query language used by the system makes the querying of entity-based co-occurring events more generic and more intuitive (i.e. searching several events occurring in the same stay, hospitalization, medical units, etc.). In contrast, that kind of functionality is usually proposed through user-friendly but predefined and specific forms (e.g. STRIDE, I2b2). Temporal and chronological aspects are a topic of interest of many IR systems (e.g. DW4TR) and are particularly relevant to IR in clinical data. Temporal querying (i.e. querying data occurring at a definite moment in time) can be achieved by the underlying search engine and its associated specific query language but the ASIS web interface still needs to be enhanced to provide specific forms able to generate the entire set of proper string-based queries. In contrast, the querying of chronologically co-occurring events (i.e. searching events occurring before, after, at the same time or within a definite time frame compared to another) is not well supported. Our department is currently discussing generic technical upgrades of the SSE that will enable us overcome those limitations but also offer powerful functionalities beyond the scope of time handling.

Technically, our system relies only on free solutions. It accesses the data through an IMDG NoSQL layer that offers very satisfactory performances with the data of 250,000 patients but which requires the design and the development of IR functionalities usually provided by the SQL when a RDBMS is used. Since November 2018, all the data of the 1.8 patients from RUH have been integrated into the POC with relative constant performance (i.e. most of the queries tested in this paper are still under the five second threshold considered acceptable by health professionals). An optimized version is nevertheless scheduled for January 2019.

Conclusion

A Health Data Warehouse is defined as a grouping of data from diverse sources accessible by a single data management system [13] which centralizes clinical, demographic and administrative data within a uniform and consistent data model. In this study, a Proof Of Concept of a Semantic Health Data Warehouse, based on the data of 250,000 patients from Rouen University Hospital is presented along with a graphical interface Semantic Access to Health Information. The system provides semantic Information Retrieval capabilities and relies on three distinct semantic layers. The system was evaluated for its ability to assist patient recruitment in five randomly selected clinical trials from Rouen University Hospital. The system showed encouraging results in accurately automating the search of the criteria and good results when used as a pre-screening tool. However, this study underlines some limitations of the system especially in relation to information extraction from unstructured Clinical Narratives which is still an essential source of information. Since November 2018, all the data of 1.8 million patients from Rouen University Hospital have been included in the Proof Of Concept and an optimized version is scheduled to be used as early as January 2019.

Acknowledgments

The authors are grateful to Nikki Sabourin-Gibbs, Rouen University Hospital, for help in editing the manuscript.

Bibliography

1. P. J. O'Connor, J. M. Sperl-Hillen, W. A. Rush, P. E. Johnson, G. H. Amundson, S. E. Asche, H. L. Ekstrom, and T. P. Gilmer. Impact of electronic health record clinical decision support on diabetes care: A randomized trial. The Annals of Family Medicine, 9(1):12–21, jan 2011. DOI: 10.1370/afm.1196.
2. Chaitanya Shivade, Preethi Raghavan, Eric Fosler-Lussier, Peter J. Embi, Noemie Elhadad, Stephen B. Johnson, and Albert M. Lai. A review of approaches to identifying patient phenotype cohorts using electronic health records. Journal of the American Medical Informatics Association, 21(2):221–230, mar 2014. DOI: 10.1136/amiajnl-2013-001935.
3. Matthew D. Krasowski, Andy Schriever, Gagan Mathur, John L. Blau, Stephanie L. Stauffer, and Bradley A. Ford. Use of a data warehouse at an academic medical center for clinical pathology quality improvement, education, and research. Journal of Pathology Informatics, 6(1):45, 2015. DOI: 10.4103/2153-3539.161615.
4. Kali VanLangen and Greg Wellman. Trends in electronic health record usage among US colleges of pharmacy. Currents in Pharmacy Teaching and Learning, 10(5):566–570, may 2018. DOI: 10.1016/j.cptl.2018.01.010.
5. Mike Cottle, Waco Hoover, Shadaab Kanwal, Marty Kohn, Trevor Strome, and N. Treister. Transforming health care through big data strategies for leveraging big data in the health care industry. Institute for Health Technology Transformation, http://ihealthtran. com/big-data-in-healthcare, 2013.
6. Wullianallur Raghupathi and Viju Raghupathi. Big data analytics in healthcare: promise and potential. Health Information Science and Systems, 2(1), feb 2014. DOI: 10.1186/2047-2501-2-3.
7. J. Petro. Natural language processing in electronic health records. https://www.kevinmd.com/blog/2011/09/natural-language-processing-electronic-health-records.html, 2015.
8. Preethi Raghavan, James L. Chen, Eric Fosler-Lussier, and Albert M. Lai. How essential are unstructured clinical narratives and information fusion to clinical trial recruitment? AMIA Summits on Translational Science Proceedings, 2014:218, 2014. PMID: 25717416, PMCID: PMC4333685.
9. Yanshan Wang, Liwei Wang, Majid Rastegar-Mojarad, Sungrim Moon, Feichen Shen, Naveed Afzal, Sijia Liu, Yuqun Zeng, Saeed Mehrabi, Sunghwan Sohn, and Hongfang Liu. Clinical information extraction applications: A literature review. Journal of Biomedical Informatics, 77:34–49, jan 2018. DOI: 10.1016/j.jbi.2017.11.011.
10. Kory Kreimeyer, Matthew Foster, Abhishek Pandey, Nina Arya, Gwendolyn Halford, Sandra F. Jones, Richard Forshee, Mark Walderhaug, and Taxiarchis Botsis. Natural language processing systems for capturing and standardizing unstructured clinical information: A systematic review. Journal of Biomedical Informatics, 73:14–29, sep 2017. DOI: 10.1016/j.jbi.2017.07.012.
11. Son Doan, Mike Conway, Tu Minh Phuong, and Lucila Ohno-Machado. Natural language processing in biomedicine: A unified system architecture overview. In Methods in Molecular Biology, pages 275–294. Springer New York, 2014. DOI: 10.1007/978-1-4939-0847-9_16.
12. Alan R. Aronson and François-Michel Lang. An overview of MetaMap: historical perspective and recent advances. Journal of the American Medical Informatics Association, 17(3):229–236, may 2010. DOI: 10.1136/jamia.2009.002733.
13. International Organization for Standardization. Health informatics – Deployment of a clinical

data warehouse. Standard, International Organization for Standardization, 2010.
14. Henry J. Lowe, Todd A. Ferris, Penni M. Hernandez, and Susan C. Weber. Stride–an integrated standards-based translational research informatics platform. In AMIA Annual Symposium Proceedings, volume 2009, page 391. American Medical Informatics Association, 2009. PMID: 20351886, PMCID: PMC2815452.
15. Hai Hu, Mick Correll, Leonid Kvecher, Michelle Osmond, Jim Clark, Anthony Bekhash, Gwendolyn Schwab, De Gao, Jun Gao, Vladimir Kubatin, Craig D. Shriver, Jeffrey A. Hooke, Larry G. Maxwell, Albert J. Kovatich, Jonathan G. Sheldon, Michael N. Liebman, and Richard J. Mural. DW4TR: A data warehouse for translational research. Journal of Biomedical Informatics, 44(6):1004–1019, dec 2011. DOI: 10.1016/j.jbi.2011.08.003.
16. C. G. Chute, S. A. Beck, T. B. Fisk, and D. N. Mohr. The enterprise data trust at mayo clinic: a semantically integrated warehouse of biomedical data. Journal of the American Medical Informatics Association, 17(2):131–135, feb 2010. DOI: 10.1136/jamia.2009.002691.
17. Eric Zapletal, Nicolas Rodon, Natalia Grabar, and Patrice Degoulet. Methodology of integration of a clinical data warehouse with a clinical information system: the hegp case. Studies in Health Tech nology and Informatics, 160(MEDINFO 2010):193–197, 2010. DOI: 10.3233/978-1-60750-588-4-193.
18. David A. Hanauer, Qiaozhu Mei, James Law, Ritu Khanna, and Kai Zheng. Supporting information retrieval from electronic health records: A report of university of michigan's nine-year experience in developing and using the electronic medical record search engine (EMERSE). Journal of Biomedical Informatics, 55:290–300, jun 2015. DOI: 10.1016/j.jbi.2015.05.003.
19. Julien Grosjean, Tayeb Merabti, Badisse Dahamna, Ivan Kergourlay, Benoit Thirion, Lina F. Soualmia, and Stéfan J. Darmoni. Health multi-terminology portal: A semantic added-value for patient safety. Studies in Health Technology and Informatics, 166(Patient Safety Informatics):129–138, 2011. DOI: 10.3233/978-1-60750-740-6-129.
20. Lina F. Soualmia, Chloé Cabot, Badisse Dahamna, and Stéfan J. Darmoni. Sibm at clef e-health evaluation lab 2015. In proceedings of CLEF 2015 - Conference and Labs of the Evaluation Forum, volume 1391 of CEUR Workshop Proceedings, 2015.
21. Chloé Cabot, Lina F. Soualmia, Badisse Dahamna, and Stéfan J. Darmoni. Sibm at clef ehealth evaluation lab 2016: Extracting concepts in french medical texts with ecmt and cimind. In 2016 Conference and Labs of the Evaluation Forum, CLEF, pages 47–60, 2016.
22. Romain Lelong, Lina F. Soualmia, Saoussen Sakji, Badisse Dahamna, and Stéfan J. Darmoni. Nosql technology in order to support semantic health search engine. In MIE 2018, 2018. DOI: 10.2196/jmir.3836.
23. Romain Lelong, Lina F. Soualmia, Badisse Dahamna, Nicolas Griffon, and Stéfan J. Darmoni. Querying ehrs with a semantic and entity-oriented query language. Studies in Health Technology and Informatics, 235(Informatics for Health: Connected Citizen-Led Wellness and Population Health):121–125, 2017. DOI: 10.3233/978-1-61499-753-5-121.
24. Romain Lelong, Chloé Cabot, Lina F. Soualmia, and Stéfan J. Darmoni. Semantic search engine to query into electronic health records with a multiple-layer query language. In MEDIR workshop, 2016.
25. Shawn N. Murphy, Michael E. Mendis, David A. Berkowitz, Isaac Kohane, and Henry C. Chueh. Integration of clinical and genetic data in the i2b2 architecture. In AMIA Annual Symposium Proceedings, volume 2006, page 1040. American Medical Informatics Association, 2006. PMID: 17238659, PMCID: PMC1839291.
26. S. N. Murphy, G. Weber, M. Mendis, V. Gainer, H. C. Chueh, S. Churchill, and I. Kohane. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). Journal of the American Medical Informatics Association, 17(2):124–130, feb 2010. DOI: 10.1136/jamia.2009.000893.

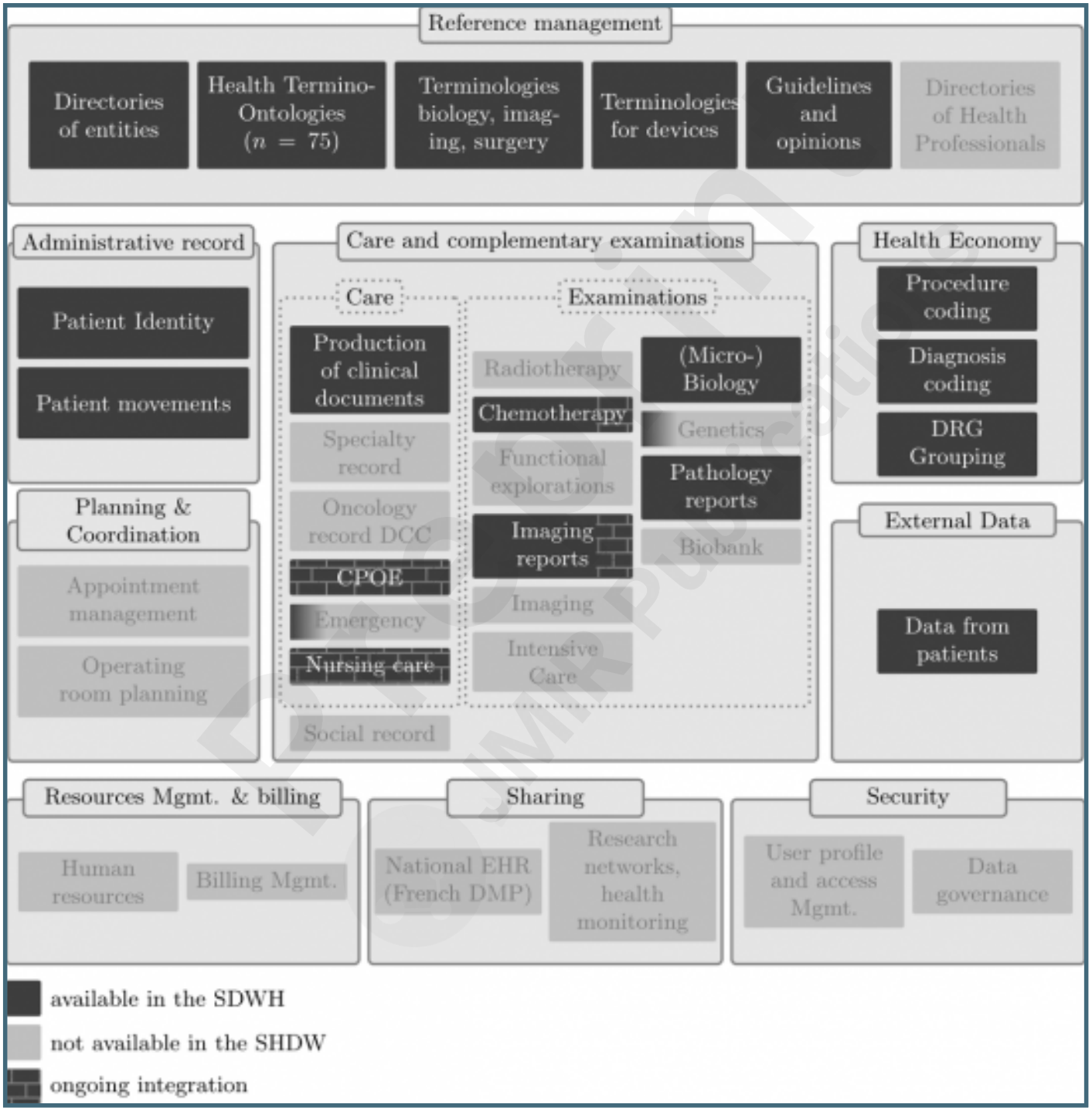
27. George Hripcsak, Jon D. Duke, Nigam H. Shah, Christian G. Reich, Vojtech Huser, Martijn J. Schuemie, Marc A. Suchard, Rae Woong Park, Ian Chi Kei Wong, Peter R. Rijnbeek, and et al. Observational health data sciences and informatics (ohdsi): Opportunities for observational researchers. Studies in Health Technology and Informatics, 216(MEDINFO 2015: eHealth-enabled Health):574–578, 2015. DOI: 10.3233/978-1-61499-564-7-574.
28. Nicolas Garcelon, Antoine Neuraz, Rémi Salomon, Hassan Faour, Vincent Benoit, Arthur Delapalme, Arnold Munnich, Anita Burgun, and Bastien Rance. A clinician friendly data warehouse oriented toward narrative reports: Dr. warehouse. Journal of Biomedical Informatics, 80:52–63, apr 2018. DOI: 10.1016/j.jbi.2018.02.019.
29. Pierre Heudel, Alain Livartowski, Patrick Arveux, Eddy Willm, and Christophe Jamain. ConSoRe : un outil permettant de rentrer dans le monde du big data en santé. Bulletin du Cancer, 103(11):949–950, nov 2016. DOI: 10.1016/j.bulcan.2016.10.001.
30. Marc Cuggia, Nicolas Garcelon, Boris Campillo-Gimenez, Thomas Bernicot, Jean-François Laurent, Etienne Garin, André Happe, and Régis Duvauferrier. Roogle: An information retrieval engine for clinical data warehouse. Studies in Health Technology and Informatics, 169(User Centred Networked Health Care):584–588, 2011. DOI: 10.3233/978-1-60750-806-9-584.
31. Denis Delamarre, Guillaume Bouzille, Kevin Dalleau, Denis Courtel, and Marc Cuggia. Semantic integration of medication data into the EHOP clinical data warehouse. Studies in Health Technology and Informatics, 210(Digital Healthcare Empowering Europeans):702–706, 2015. DOI: 10.3233/978-1-61499-512-8-702.
32. Karsten U. Kortüm, Michael Müller, Christoph Kern, Alexander Babenko, Wolfgang J. Mayer, Anselm Kampik, Thomas C. Kreutzer, Siegfried Priglinger, and Christoph Hirneiss. Using electronic health records to build an ophthalmologic data warehouse and visualize patients' data. American Journal of Ophthalmology, 178:84–93, jun 2017. DOI: 10.1016/j.ajo.2017.03.026.
33. Qlikview. URL:https://www.qlik.com/us/products/qlikview, WebCite®: http://www.webcitation.org/76Cws45GA. Accessed: 2019-02-15.
34. Chloé Cabot, Lina F. Soualmia, Julien Grosjean, Romain Lelong, and Stéfan J. Darmoni. Integrating and retrieving clinical and omic data in electronic health records. In 7th International Workshop on Knowledge Representation for Health Care (KRH4C) and 8th International Workshop on Process-oriented Information Systems in Healthcare (ProHealth), pages 154–159, 2015.
35. Stéfan J. Darmoni, Benoit Thirion, Jean-Phillipe Leroy, Magaly Douyere, Benoit Lacoste, Christophe Godard, Isabelle Rigolle, Martial Brisou, Stéphane Videau, Eric Goupy, Josette Piot, Myriam Quere, Saida Ouazir, and Habib Abdulrab. A search tool based on 'encapsulated' MeSH thesaurus to retrieve quality health resources on the internet. Medical Informatics and the Internet in Medicine, 26(3):165–178, jan 2001. DOI: 10.1080/14639230110064488.
36. Nicolas Griffon, Matthieu Schuers, Lina F. Soualmia, Julien Grosjean, Gaétan Kerdelhué, Ivan Kergourlay, Badisse Dahamna, and Stéfan J. Darmoni. A search engine to access PubMed monolingual subsets: Proof of concept and evaluation in french. Journal of Medical Internet Research, 16(12):e271, dec 2014. DOI: 10.2196/jmir.3836.
37. Chloé Cabot, Lina F. Soualmia, Julien Grosjean, Nicolas Griffon, and Stéfan J. Darmoni. Evaluation of the terminology coverage in the french corpus LiSSa. Studies in Health Technology and Informatics, 235(Informatics for Health: Connected Citizen-Led Wellness and Population Health):126–130, 2017. DOI: 10.3233/978-1-61499-753-5-126.
38. Postgresql. URL:https://www.postgresql.org/, WebCite®: http://www.webcitation.org/76CxVKdFk. Accessed: 2019-02-15.
39. Julien Grosjean. Modélisation, réalisation et évaluation d’un portail Multi-terminologique, Multi-discipline, Multi-lingue (3M) dans le cadre de la Plateforme d’Indexation Régionale

(PlaIR). PhD thesis, Université de Rouen, École doctorale Sciences Physiques, Mathématiques et de l'Information pour l'Ingénieur, 2014.
40. Aurélie Névéol, Hercules Dalianis, Sumithra Velupillai, Guergana Savova, and Pierre Zweigenbaum. Clinical natural language processing in languages other than english: opportunities and chal- lenges. Journal of Biomedical Semantics, 9(1), mar 2018. DOI: 10.1186/s13326-018-0179-8.
41. B. L. Humphreys, A. T. McCray, and D. A. B. Lindberg. The unified medical language system. Yearbook of Medical Informatics, 02(01):41–51, aug 1993. DOI: 10.1055/s-0038-1637976.
42. Guergana K. Savova, James J. Masanz, Philip V. Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C. Kipper-Schuler, and Christopher G. Chute. Mayo clinical text analysis and knowledge extraction system (cTAKES): architecture, component evaluation and applications. Journal of the American Medical Informatics Association, 17(5):507–513, sep 2010. DOI: 10.1136/jamia.2009.001560.
43. N. F. Noy, N. H. Shah, P. L. Whetzel, B. Dai, M. Dorf, N. Griffith, Jonquet C., D. L. Rubin, M.-A. Storey, C. G. Chute, and M. A. Musen. BioPortal: ontologies and integrated data resources at the click of a mouse. Nucleic Acids Research, 37(Web Server):W170–W173, may 2009. DOI: 10.1093/nar/gkp440.
44. Aurélie Névéol, Julien Grosjean, Stéfan Jacques Darmoni, and Pierre Zweigenbaum. Language resources for french in the biomedical domain. In LREC, pages 2146–2151, 2014.
45. Infinispan data grid platform. URL:http://infinispan.org/, WebCite®: http://www.webcitation.org/76Cw0We8b. Accessed: 2019-02-15.
46. Francesco Marchioni and Manik Surtani. Infinispan data grid platform. Packt Publishing Ltd, 2012. ISBN: 978-1-84951-822-2.
47. Apache lucene. URL:https://lucene.apache.org/, WebCite®: http: //www.webcitation.org/76CxHXK5k. Accessed: 2019-02-15.

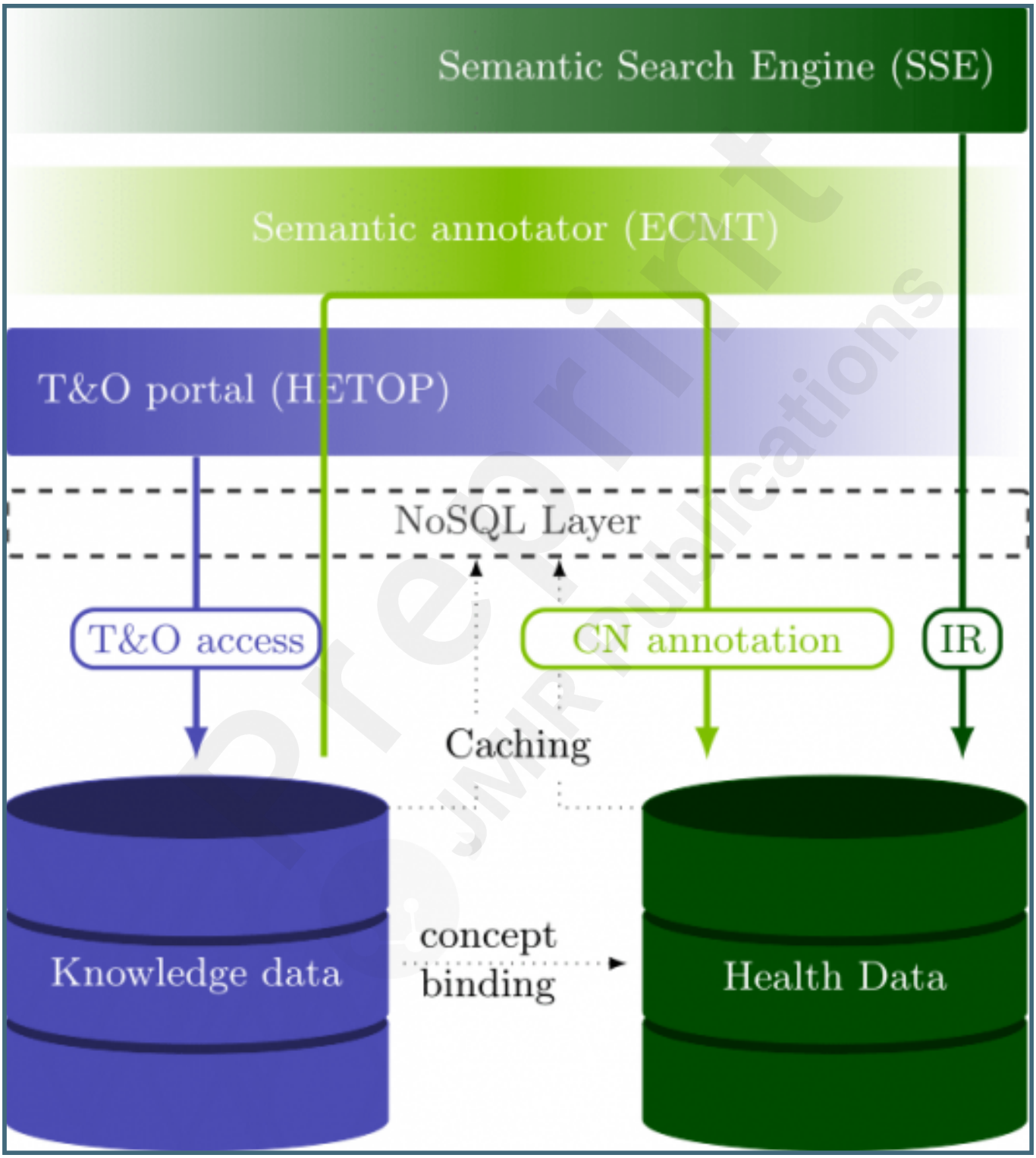
Supplementary Files

Figures

Functional coverage of the SHDW in terms of data according to each domain (viz. Reference management, Administrative record, Care, Examinations, Health economy, Planning & Coordination, External Data, Resource management & Billing, Sharing and Security). Data already included in the SHDW are represented by a dark gray opaque background, whereas a light gray background indicates that data are not included and not planned to be in the short or medium term. Background partially or totally covered with bricks corresponds to data for which inclusion is in progress or is planned in the short term or medium term.



Functional architecture of the SHDW which provides semantic IR functionalities form clinical data. The two data repositories “knowledge data” and “Health Data” respectively maintain the reference KOSs and the health data pertaining to the SHDW. These data are accessed through a NoSQL layer by the three distinct components HeTOP, ECMT and the SSE which, each, operate over a different range of data.



Partial Conceptual Model of the SHDW represented as a directed and attributed graph. Entities corresponding to elements from T&Os are represented with dashed outlines.