Knowledge guided integration of structured and unstructured data in health decision process

Marouane Radaoui
Univ. Manouba / ENSI / RIADI Laboratory
Manouba, Tunisia *raddaouimarouen@gmail.com*

Sana Ben Abdallah Ben Lamine
Univ. Manouba / ENSI / RIADI Laboratory
Manouba, Tunisia *sana.benabdallah@riadi.rnu.tn*

Hajer Baazaoui Zghal
Univ. Manouba / ENSI / RIADI Laboratory
Manouba, Tunisia *hajer.baazaouizghal@riadi.rnu.tn*

Chirine Ghedira Guegan
Univ. of Lyon / CNRS / IAE / Univ. of Lyon 3 / LIRIS / UMR 5205
Lyon, France *chirine.ghedira-guegan@univ-lyon3.fr*

Nadia Kabachi
Univ. of Lyon / Univ. of Lyon 1 / ERIC Laboratory / EA 3083
Villeurbanne, France *Nadia.Kabachi@univ-lyon1.fr*

Abstract

Data in the health domain is continuously increasing. It is collected from several sources, has several formats and is characterized by its sensibility (protection of personal health data). These characteristics make the management and the expert interaction with the collected data, in order to facilitate decision-making in Health Information Systems (HIS) a challenging field. In this paper, we propose a Knowledge guided integration of structured and unstructured data for health decision process. The knowledge is represented by domain ontology, which allows the integration of structured and unstructured data, stored in NoSQL format. Our motivation is to combine the confirmed advantages of ontologies and NoSQL databases both in data integration and decision aided processes. The proposed ontology has been implemented and evaluated using quality metrics. The approach was evaluated and results show response time optimization, compared with traditional approaches, and improvement of data relevance.

Keywords: knowledge-based Health data, ontology, data integration, NoSQL database.

1. Introduction

The huge amount of data collected in the health field raises several challenges related to big data in general and the health field in particular. We can mention the problems of storage, integration, analysis, exploitation, protection of personal health data and assistance to the Medical Decision. Considering these problems has become crucial for both the medical staff and the patient. Hence the crucial role of Health Information Systems (HIS) in the organization and efficiency of the health care process in general. Particularly, the monitoring and detection of the emergence of a disease such as influenza and the control of the epidemiological situation is the role of several health organizations. These organizations collect data from several health institutions such as public hospitals, laboratories, pharmacies, etc. However, these establishments are independent of each other. Moreover, social data is now an essential data source allowing knowledge generation, particularly in the medical field. Tweets are frequently considered as trust source, which can complete initial data or alert to unusual phenomenon. Consequently, collected medical data are multisource and heterogeneous. Heterogeneity is both at syntactic and semantic levels, and is considered among the major challenges for medical data integration systems.

In this paper, the heterogeneity of structured and unstructured medical data is addressed based on a domain ontology and on an integration process. The integration of multisource medical data plays an important role in tracking the dynamics of the health environment, which helps predict unexpected phenomena such as the spread of an epidemic.

Our contributions could be resumed around 1) building a medical domain ontology based on the structured data and the reuse of existing ontologies, 2) an approach allowing the integration of structured and unstructured heterogeneous and multisource data based on the proposed domain ontology, 3) historized integrated data storage in a NoSQL database, 4) an expert aided decision process to facilitate experts detecting the emergence of diseases or the spread of an epidemic.

The proposed ontology has been built and evaluated using quality metrics. The integration approach was illustrated by a case study and evaluated by response time measures.

The rest of this paper is organized as follows. In Section 2, we give an overview of related work on the use of ontologies and NoSQL databases in the medical data integration process and as a support of decision making in HIS. We also expose our motivations and the main goals of this paper. Then in Section 3, we detail our proposition and its main contributions. Section 4 is dedicated to present the implementation and evaluation of our approach. Finally, we conclude and present our future work in section 5.

2. Related Work and Motivation

An ontology is defined as a formal and explicit specification of a shared conceptualization [9]. Several languages have been proposed to represent ontologies like RDF (Resource Description Framework) and OWL (Web Ontology Language) [2]. In the medical field, several ontologies have been built. The National Library of Medicine (NLM) has taken the initiative of the Unified Medical Language System (UMLS) project to improve the ability of computer systems to understand the meanings of requests from specialists in the medical field [1]. Other initiatives were significant: Digital Anatomist project for the conceptualization of physical objects that make up the human body on a macroscopic scale [10], the MeSH [1] thesaurus used for indexing medical resources online and SNOMED[2], which is a controlled vocabulary, used in the clinical field.

On the one hand, biomedical data is collected from multiple sources such as hospitals, laboratories, public health organizations and even social media. Several solutions were proposed to address the integration issue and both syntactic and semantic heterogeneity.

Mediated data integration systems involve installing a uniform mediation layer between users and data sources. With these systems, users query a fictional representation of data stored in their own stand-alone sources, geographically distributed and interconnected via the Internet. These systems provide direct access to data with source-specific languages and functions. Despite their poor performance due to query translation and internet connection, data security is almost guaranteed since they are stored and managed by their own sources.

Industry and researchers are attracted by NoSQL databases which are schema-less databases allowing easy and automatic scaling, better performance and high availability. Many researchers dealt with the data integration issue using NoSQL Databases [11] and [13] but not especially in the medical field.

Involving ontologies in data integration systems is becoming more and more a trend for many research projects to solve the problem of semantic heterogeneity. For mediation integration systems in the medical field, ontology facilitates interrogation and

[1] https://meshb.nlm.nih.gov

[2] http://www.snomed.org

homogenization of data from various sources [6] and [12]. Ontology is also evoked together with NoSQL databases for data integration. In the literature, several works dealt with ontologies and NoSQL in HIS. Recently, several researchers are interested in HIS use, implementations and especially interoperability [7]. Ontologies are interesting researchers not only by their capability to represent knowledge [5], but also as a decision support, like in [14], where the authors used ontologies in Health Data Management context to support collaboration in ambient assisted living and work reintegration to enable cooperation among different clinical and non-clinical stakeholders. Many researches have shown the efficiency of NoSQL databases particularly in data management and analysis in health information systems [4], [8] and [3].

Our ultimate motivation is to generate and capitalize on knowledge, especially that derived and validated by the expert. In this work, the knowledge is represented by ontologies and are employed in two levels: first level, to achieve structural and unstructured data NoSQL integration. Real structured medical data collected by the partner of the project is used in a medical domain ontology building. A tweets case study illustrates the unstructured data NoSQL integration process via the proposed ontology. In the second level, the ontology is used to assist the decision maker in the process of choosing the relevant unstructured data to integrate in the NoSQL historized data base and therefore to be considered in the process of detecting emerging diseases or an epidemic spreading. Our motivation is to combine the confirmed advantages of ontologies and NoSQL databases both in data integration and decision aided process in health information systems.

3. Ontology based integration of structured and unstructured data for NoSQL base Health Information System

In this paper we address the integration phase which allows to store heterogeneous and multi source structured and unstructured data in a NoSQL database (NoSQL DB) HIS based on a medical domain ontology.

Structured data is collected from several sources (hospitals, pharmacies…) while structured data come from social networks (tweets…). Figure 1 presents the architecture of our proposition.

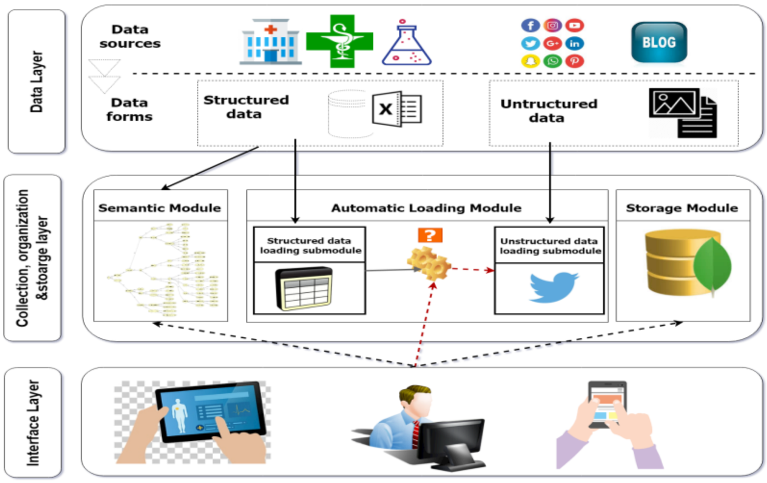


Fig 1. Proposed architecture

- **The data layer**: represents the sources of data and their types. Indeed, these sources can be hospitals, pharmaceutical organizations and even Internet sources (social networks, blogs, etc.) in case of an alert sent by the data analysis system. These sources provide data in several formats which can be structured (databases), semi-structured (XML files, csv. etc.) or unstructured (texts, images. etc.)
- **The Collection, organization & storage layer**: composed by three modules:
    - *Semantic module:* formed essentially by the domain ontology,
    - *Automatic loading module:* allows the loading of structured and unstructured

data in the NoSQL Database. Through a data anonymization module, it guarantees the protection of personal data. It also ensures data consistency using the semantic module,

- *Storage module:* constitutes the output of the integration phase, which is a document-oriented NoSQL database.

- **Interface Layer**: permits the interaction of the expert with the HIS. In addition to the traditional interaction possibilities offered by an information system, the expert is involved in the process of building and validating the domain ontology. He is also responsible to decide of the necessity to force the loading of the unstructured data. In the next subsection, the semantic module will be detailed.

3.1. Semantic module

The central source of knowledge is the proposed domain ontology containing all the relevant domain concepts and their domain inter-relationships. (*cf.* Definition 1 and Figure 2 (a)).

Definition: The proposed ontology is defined as 3-tuple <*C*, *A*, *R*> where:

- *C*: is the set of concepts
- *A*: is the set of attributes
- *R*: is the set of relationships where $R = R_{tax} \cup R_{non\_tax}$ where
  R_{tax}: is the set of the taxonomic relationships
  R_{non_tax}: is the set of the non-taxonomic relationships.

A one-month lasting research and analysis of existing medical ontologies was entreprised in order to select the ones that will be reused and to extract the concepts that cover our domain of interest. Eight experts were involved in the understanding and selection process: four of them belong to the medical field and four to the computer science field. The relevant domain concepts in this ontology are identified and extracted in a first step from the official collected structured data, such as Person, diseases, location. Other relevant concepts are proposed by experts such as treatment and place_of_care.

We reused existing concepts from two ontologies (MeSH and EPO) and two vocabularies (ICD-11 and the international classification of medical and paramedical professions) presented respectively by WHO [3] and SFMG[4].

The concept Person is extracted from the ontology MeSH. Under this concept, we added four concepts; Person:Medical_Profession and Person:Paramedical_Profession are inherited from the international classification of medical and paramedical professions. We added also a concept Person:Patient and we divided it to 3 categories according to age groups: child, adolescent and adult. We subdivided each category also into sub-categories according to more minimized age groups, for example Person:Patient:Child is divided into baby, young child and adult child. Further to division by age group, the concept Person and its subclasses are subdivided according to gender. The concept Diseases is inherited from the ICD-11.

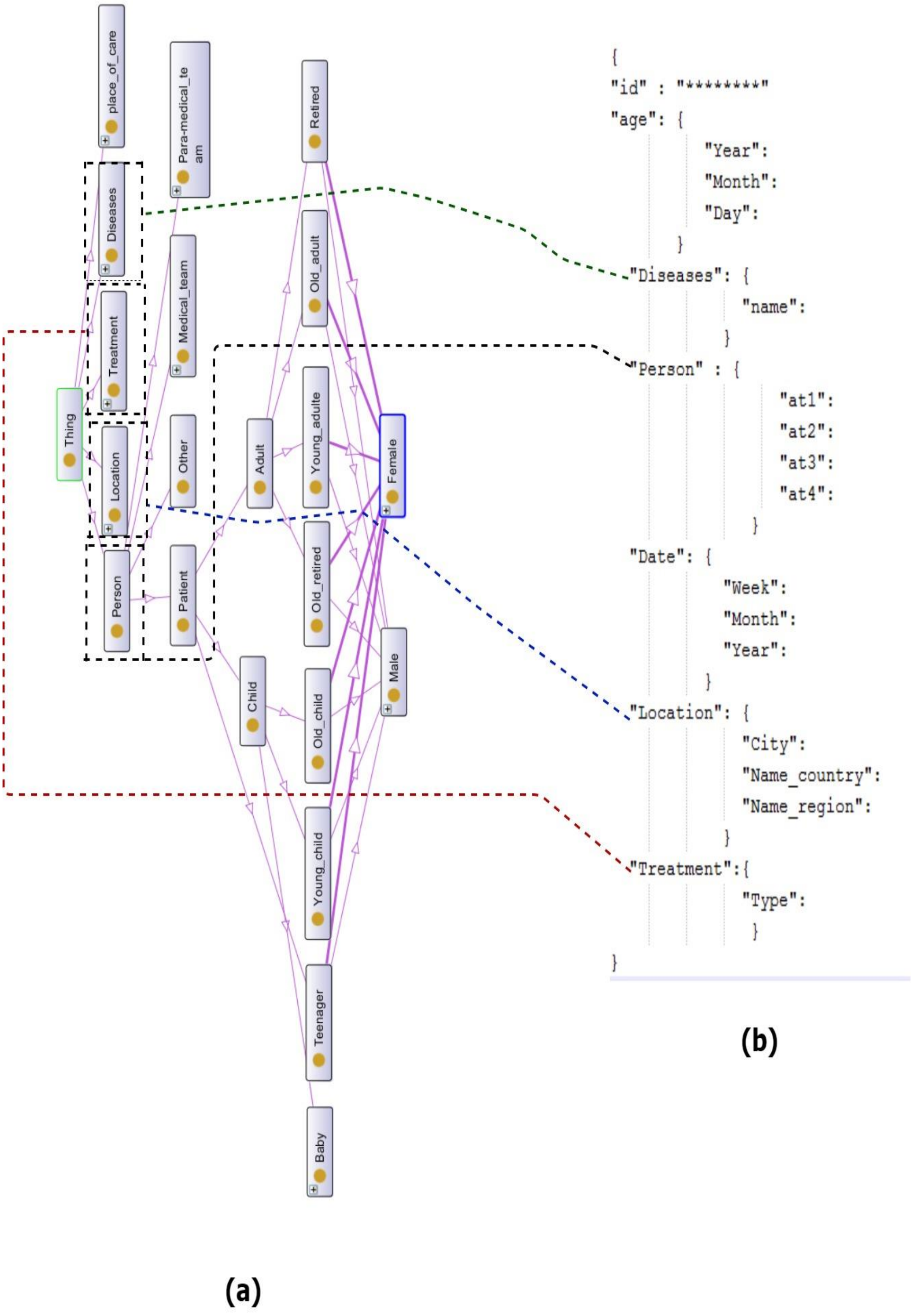
We point out that our proposed ontology consists in an import of concepts from the aforementioned standardized ontologies, that we refined, classified linked and enriched with properties according to the knowledge rules, for instance on low-level signals related to the epidemiology detection, suggested by the experts. The experts and the partners of the project suggested the following attributes added to the Person concept. Table 1 describes the Person's attributes.

It is important to mention at this level, that the ontology is built based on the expert knowledge and validation. Indeed, the expert (medical staff) can, through an interface (interface layer), en-rich and complete the ontology by new concepts and relations to represent intrinsic knowledge or extracted knowledge from tweets sent by expert and non expert users. The final ontology allows the preparation of the prediction process of disease

[3] www.who.int

[4] www.sfmg.org

or unexpected phenomenon.



```
{
"id" : "********"
"age": {
        "Year":
        "Month":
        "Day":
      }
  "Diseases": {
            "name":
          }
  "Person" : {
              "at1":
              "at2":
              "at3":
              "at4":
            }
  "Date": {
          "Week":
          "Month":
          "Year":
        }
  "Location": {
            "City":
            "Name_country":
            "Name_region":
          }
  "Treatment":{
            "Type":
             }
}
```

Fig 2. Proposed domain ontology

Table 1. The Person's attributes

Concepts	Attribute	Subattribute	Type	Description
Person	Address	-	String	Describes person’s address
	Postal code	-	Number	Describes person’s city code

	Social Security	CNAM	Boolean	Indicates if a person has social security “CNAM”
		Complement	Boolean	Indicates if a person has social security’s complement or no
	Inhabitant	Collective	Boolean	indicates if the person lives with others
		Individual	Boolean	indicates if the person lives alone
	School	-	Boolean	Indicates if the person is still at school
	Job	-	Boolean	Indicates if the person has a job
	Homeless	-	Boolean	Indicates if person is homeless

3.2. NoSQL Storage Module

There are four main families of NoSQL databases: key-value model, document-oriented model, column-oriented model and graph-oriented model [14].

For its flexibility, standardization, large storage capacity and the possession of its own meta-data (json ou bson), the documents-oriented model is adopted to design the NoSQL DB. Figure (2 (b)) illustrates the design of the adopted document-oriented model based on the domain ontology structure. The integration of streaming data (from real time sources e.g. sensors, tweets, etc.) is also supported but this module, taking advantage of the scalability and extensibility of NoSQL.

3.3. Automatic loading module

Structured and unstructured data coming from several sources (hospitals, pharmacies, social networks ...) will be stored in the NoSQL database according to fixed strategy. Algorithm 1 details the loading steps. In fact, when there is not any trigging alert following an abnormal event (epidemic . . .), the data integration phase stores only the structured data (line 3). However, when triggering an alert or by the intervention of the expert (VarExpert==True), the data integration phase starts the storage of unstructured data as well (line 5).

Algorithm 1. LoadingAlgorithm(Date)

```
1:      Input: Data
2:      Begin
3:      LoadingStructuredData()
4:      If (alert==true or varExpert==true) Then
5:        loadUnstructuredData()
6:      EndIf
7:      End
```

Structured data loading

Since we deal with a medical context, sensitive data are present. A data is said sensitive if it identifies or has a potential to identify a person. To protect personal data, an anonymization technique should be applied on sensitive data. In this work, we choose to delete quasi-identifying or sensitive attributes from original data because personal data is used to identify the patient, while what is important for a stakeholder in the medical field is the state of health of a community of people. So the expert’s decision will not be influenced by the deletion of sensible attributes.

This technique shows its effectiveness because it is irreversible since it is impossible to generate the original data having the anonymized one. To apply this technique we went through two steps: (1) The definition of the list of sensitive attributes in the original data in provenance from the data layer; (2) Deleting the specified attributes and their associated values.

Unstructured data loading

Structured and official data sometimes are not sufficient for the decision makers especially in critical cases (the presence of an abnormal event). Therefore, in order to improve the decisions of experts, the storage of unstructured data can be a solution. Unstructured data storage begins if there is a medical alert or a decision of the expert interacting with the HIS.

Structured and official data sometimes are not sufficient for the decision makers especially in critical cases (the presence of an abnormal event). Therefore, in order to improve the decisions of experts, the storage of unstructured data can be a solution. Unstructured data storage begins if there is a medical alert or a decision of the expert interacting with the HIS.

To address the problem of unstructured data integration we handled a set of tweets as an example. We are then faced to the necessity of mapping the unstructured data (tweets in our case) to the proposed ontology concepts. For that purpose, a mapping algorithm is proposed cf. Algorithm 2). The principle role of Algorithm 2 is to decide whether a tweet matches with the domain of the proposed ontology (*cf.* Figure 3) so that we decide whether to store it in the NoSQL database or not. Algorithm 2 consists of 4 main steps. It transforms the tweet into RDF schema (RDFS), uses its components and compares their similarities with the ontological concepts calculated by the formula with a threshold similarity measure based on BabelNet. The algorithm returns "True" if the similarity of two components (subject, object) is greater than and "False" otherwise.

Notation in Algorithm 2:

T_N: Cleaned tweet / RDF_T:RDF schema extracted from cleaned tweet / sim_1, sim_2: similarities measures / α: Threshold of the similarity measure / $|E_c|$: Number of ontological concepts / it: iterator.

```
Algorithm 2. TweetMapping(tweet T, threshold α)
1:Input: tweet T, threshold α
2:Output: Boolean Decision
 3:  Begin
 4:  TN =Cleaning(T)
 5:  RDFT =TweetToRDFS(T)
 6:  sim1=sim2=it=0
 7:  Decision=False
 8:  while it<|$E_c$| and Decision == false do
 9:  begin
 10:        it2=0
 11:        sim1=Similarity($c_{it}$, Subject(RDFT ))
 12:        if (sim1 >= α) then
 13:        begin
 14:          Decision = True
 15:          while $it_2$ < $E_c$ and $sim_2$< α do
 16:          begin
 17:            sim2=Similarity($c_{it_2}$, Object(RDFT ))
 18:            if(sim2 <= α) then
 19:               begin
 20:                Decision = False
 21:               EndIf
 22:            it2=it2+1
 23:          EndWhile
 24:        EndIf
 25:        it=it+1
 26: EndWhile
 27: Return (Decision)
 28:End
```

The main steps of algorithm 2:

- *Cleaning (line 4):* It consists of eliminating all special characters such as , -, ...
- *RDF triplets building*: to construct RDF triplet we propose the TweetToRDFSfunction (line 5 Algorithm 2). This function is essentially based on a grammar dictionary or parser. In our work, we used the standard Sanford dictionary[5]. This tool allows to determine the grammatical function of the terms of a tweet (verb, subject, name, adverb, complement ...). Once the cleaned tweet analyzed and the grammatical function is determined for each term, the RDF triplet is built.

[5] https://nlp.stanford.edu/software/lex-parser.html/Download

- *Calculation of a similarity measure (line 8 – line 28):* having the RDF triple which corresponds to the tweet in input, Algorithm 2 measures the degree of similarity between the components of the constructed triple and the concepts of the proposed domain ontology. The measure of similarity between two concepts is defined as:

$$WUP(c_1, c_2) = \frac{2 \times depth(LCS(c_1, c_2))}{depth(c_1) + depth(c_2)} \qquad (1)$$

Where:
$depth(c_i)$: The depth of the concept c_i
$LCS(c_i, c_j)$: The most specific shared concept between c_i and c_j
We use this measurement in two cases. The first one computes the similarity measurement between the RDF triplet subject and ontological concepts (Line 11). If this measurement is greater than , the algorithm computes the second similarity measurement, which is used to compare between the RDF triple object and the ontological concepts (line 17 of the algorithm 2) and compares it with to return True or False. The "Decision" variable takes the value true if the similarities measures (sim1 and sim2) are greater than the threshold "α", then the tweet is stored and otherwise it is rejected. The output of this phase is the NoSQL DB with integrated structured and unstructured data.

4. A case study

In this section, we present a case study that illustrates how our proposal integrates structured and unstructured data into a NoSQL database to assist the expert’s decision process. The integration process is based on a domain ontology which conceptualizes the expert’s and the domain’s knowledge. Two scenarios are proposed. The first scenario illustrates the loading of structured data, which is done periodically. This task guarantees the confidentiality of personal data by applying an anonymization technique (cf. Section 3.3). In the usual case of interaction with the HIS, the expert can for example query the NoSQL database on « the rate of influenza infection in the population of students aged between 16 and 27 years ». If the returned value is less than a fixed threshold (for e.g. the national average rate), the loading process continues on the treatment. Otherwise, an alert is triggered and the second scenario is launched. In this second scenario, the expert intervenes at two levels: first, to modify the structured data’s loading period from 7 to 3 days. Second, to start the loading of unstructured data from social networks (tweets), in order to better track this anomaly or unusual phenomenon. The pro-posed approach, allows the evaluation of the confidence of unstructured data. For example, the tweet "#an_epidemic_is_spreading_in_the_university_environment" is processed according to the steps of algorithm 2 (build an RDF triplet based on the tweet, then mapping the RDF triplet with the ontology concepts. The similarity measure (cf. Equation (1)) indicates that the current tweet matches the ontology concepts and the automatic loading module stores it in the NoSQL database. While, the tweet "France_has_won_the_world_Cup_2018" is rejected because of its weak confidence value based on the domain ontology. Storing structured and unstructured data provides decision support for experts; to set strategies to limit the risks of a disease (for example, launching a vaccination campaign against influenza) and is a tool for predicting and preventing unexpected events by assessing the confidence of unstructured data.

5. Implementation and experimental evaluation

5.1. Data set and implementation environment

The data are provided by the public organization operating in the health sector. They present influenza data covering the period from 2009 to 2017. These data relate to patients, regions, diseases (influenza) and age groups of patients. With the observatory experts, we built a set of tweets, sent by experts and citizens that presents the unstructured data. The creation of the test tweets is based on the collection of tweets published by users of this network, including doctors, official medical organizations and even any type of user. Modification of real tweets are done by changing their context or rephrasing them in order to obtain a set of tweets that can contain noise. The aim here is to vary the data source and the users. From more technical point of view the frameworks that have been employed in the implementation of the prototype are: JAVA an object-oriented programming language; MongoDB: document oriented NoSQL database; Jena API: that allows the management of domain ontologies; OWL (Ontology Web Language): that represents and describes ontologies and BabelNet API: that represents the lexical resource.

5.2. Results and discussion

In this section, we evaluate the proposed domain ontology and the proposed integration approach.

Ontology implementation and evaluation

Different approaches have been proposed to evaluate an ontology, targeting a number of different criteria and metrics. In this work, the following criteria are chosen to evaluate the proposed ontology:

- *Base metrics***:** Basic metrics include simple metrics such as the number of classes, axioms, objects, and so on. These metrics indicate the amount of ontology elements.
- *Schema metrics:* concern the design of the ontology. The metrics in this category indicate the richness, width, depth, and inheritance of an ontological schema design. The most significant indicators in this category are:
  - Attribute richness (**AR**): is defined as the average number of attributes (slots) per class. It is computed as the number attributes for all classes (att) divided by the number of classes (C),
  - Inheritance Richness (**IR**): describes the distribution of information across different levels of the ontology. It is defined as the average number of subclasses per class,
  - The relationship richness (**RR**) of a schema: is defined as the ratio of the number of (non-inheritance) relationships (P), divided by the total number of relationships defined in the schema (the sum of the number of inheritance relationships (H) and non-inheritance relationships (P)),
  - Axiom-Class ratio: This metric describes the ratio between axioms and classes. It is calculated as the average amount of axioms per class
  - Class Relation Ratio: This metric describes the ratio between the classes and the relations in the ontology.

In our work, we refer to OntoMetrics[6], a web-based tool for ontology evaluation that calculates, validates and displays statics about a given ontology. We applicate this evaluation to our ontology and the results are summarized in table 2.

Table 2. Quality metrics of the proposed ontology

Basic metrics				**Schema metric**				
Axioms	**Classes**	**Proprieties**		**AR**	**IR**	**RR**	**Axiom/relation ratio**	**Class/relation ratio**
		Data prop	**Object prop**					
798	332	13	16	0.048	1.033	0.04	2.40	0.93

[6] https://ontometrics.informatik.uni-rostock.de

Automatic loading process evaluation

First Evaluation: Structured data loading. To evaluate the structured data loading, we used two metrics: (1) the data loading time and (2) the response time to a query. Experimental results are summarized in respectively in Tables 4 and 5. Request examples are given in Table 3.

Loading time remains constant regardless of the stored records is the perfect scenario for loading data. Nevertheless, it is impossible in reality because there is no ideal data integration system.

Analyzing Table 4, we can notice that the loading time has increased from 5 to 13 seconds when the number of records has been increased from 2600 to 10400 for the traditional approach. While it went from 4.2 to 7 seconds only for the same variation in the number of recordings, which highlights the effectiveness of our integration approach compared to classical approaches.

Analyzing Table 5, we notice that increasing the data volume does not cause a remarkable degradation of the system's capacity to answer the user's queries. For the request R1, the response time is 7ms for 2600 records and 19ms for 10400 records. While, for the same query, the traditional system responds respectively in 10ms and 35ms for the same number of records. The no schema nature of the NoSQL system eliminates the need for schema transformation between different data sources. This transformation causes a waste of time for the traditional databases.

Table 3. Queries Description

Request	SQL (traditional approach)	MongoDB (our approach)
R1	Select Count(id_person) From Person P, Region R, Diseases D Where (R.Name_country='Tunisia') AND (D.Name='Influenza') AND (R.id_region=P.id_region) AND (D.id_Disease=P.id_Disease),	>PersoDiagMedi.DB.find(f \Region.Name_country" :" Tunisia" , \Diseases.Name" :" Influenza" g).count();
R2	Select Person From Person P, Date D Where D.Year BETWEEN '1993' AND '2003' AND (P.id_Date=D.id_Date)	>PersoDiagMedi.DB.find(f \Year" :f \$ gte" :1993, \$ lte" : 2003g g)

Table 4. Loading execution time

	2600	5200	7800	10400
Our approach	4.2 ms	5.3 ms	6 ms	7.3 ms
Traditional approach	5 ms	8.3 ms	11 ms	13 ms

Table 5. Obtained results

Query	Records Number	Time request (mongo R)	Time request (SQL R)
R1	2600	7 ms	10 ms
	7200	16 ms	24 ms
	10400	19 ms	35 ms
R2	2600	5 ms	8 ms
	7200	10 ms	17 ms
	10400	14 ms	26 ms

Second Evaluation: Unstructured data loading. The second experiment aims to show the importance of the semantic module (the domain ontology) to evaluate the confidence of the tweets and then to improve the relevance of the stored unstructured data in the NoSQL DB. For that, we realized two tests: the first (Test 1) is to store the tweets without involving the domain on-ology in the data loading process and the second test (Test 2) consists of using the semantic module. For both tests, we calculated the precision and the recall. The result of this evaluation is presented in Figure 3.

Analyzing Figure 3, we can notice that the precision-recall curve of the first test (without ontology) is located below the precision-recall curve of the second test (using the domain ontology). This justifies that our approach improves the values of precision and recall compared to the traditional approaches. Therefore, our approach maximizes the relevance of the tweets stored in the DB and by the way optimizes the storage space since it stores only the confident tweets.

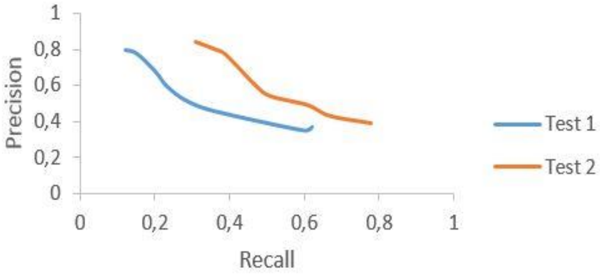


Fig 3. Precision-Recall

6. Conclusion and future works

In this paper, we propose a multi-source and heterogeneous health data integration. The idea is to store data in the same structure which is guided by knowledge. The consideration of formal and structured data only in the integration phase is limited in case of decision-making process, especially when it comes to critical situations (expansion of an epidemic, etc.). Thus considering unstructured data improve the decision making process, particularly collected from private or public social networks at real time while ensuring an efficient confidence evaluation technique, helps experts in their decision. To this end, our contributions treat with (1) the building of an ontology which conceptualizes the domain and the expert's knowledge; (2) a two-level use of the proposed ontology in the integration process: to centralize and to generate knowledge since it allows assessment of the confidence level of unstructured data in order to improve the relevance of data and then to optimize the storage space; (3) integrated data storage in a NoSQL database, and 4) an expert aided decision process in the health field. The evaluation of our proposal and its comparison with classical approaches have shown its efficiency and a valuable response time. The use of the ontology in the integration process have also shown an improvement of the precision-recall values of test requests. In our work in progress, we are testing the efficacy and response time in a large-scale case by real-time treatment of tweets, increasing the number of official organisms acting as sources of structured data and reducing the loading period of structured data. We intend to perform the implementation of a predictive analysis algorithm dedicated to the experts.

Acknowledgements

This work was funded by the SCUSI (Coopérations scientifiques et académiques internationales) program of the region Auvergne Rhône-Alpes in France for the project “PersoDiagMedi”, Number 1700938003.

References

1. Bodenreider O.: The unified medical language system (umls): integrating biomedical terminology, 32 (Database-Issue), 267-270 (2004)
2. Bruijn J., Lara R., Polleres A., Fensel D.: OWL DL vs. OWL Flight: conceptual modeling

and reasoning for the semantic Web. In: Proceedings of the 14th international conference on World Wide Web, pp. 623-632. WWW 2005, Chiba-Japan (2005)

3. Divya .S , Shivaprasad. N.: Bigdata: A Survey on RDBMS and Various NOSQL Databases On Storing Medical Ima ges. International Journal of Advance Research and Development. 57-63 (2017)
4. Ercan, M., Lane, M.: An Evaluation of NoSQL Databases for Electronic Health Record Systems. In: In: Proceedings of the 25th Australasian Conference on Information Systems, Auckland (2014)
5. Fareedi, A., Ghazawneh, A.: An Ontology Approach for Knowledge Acquisition and Development of Health Information System (HIS). In: Proceedings of the 27th International Conference on Information Systems Development, ISD 2018, Lund (2018)
6. Hua M., Frank J.M., Elizabeth G.,Yu-Ning W., Eric R., Robert Beck J.: Integration of prostate cancer clinical data using an ontology. Journal of Biomedical Informatics, 42 (6), 1035-1045 (2009)
7. Kobusinge, G., Pessi, K., Koutsikouri, D., Mugwanya, R.: An Implementation Process of Interoperability: A Case-Study of Health Information Systems (HIS). In: Proceedings of the 27th International Conference on Information Systems Development, ISD 2018, Lund (2018)
8. Madhava, V.: Big Data Electronic Health Records Data Management and Analysis on Cloud with MongoDB: A NoSQL Database. International Journal of Advanced Engineering and Global Technology, 3 (7), 96-949 (2015)
9. Mariano F., Asuncion G., Natalia J.: Methontology: from ontological art towards ontological engineering. In: In: Proceedings of the Engineering Workshop on Ontological Engineering, AAAI 1997, California (1997)
10. Neal PJ., Shapiro LG., Rosse C.: The digital anatomist structural abstraction: a scheme for the spatial description of anatomical entities. In: In: Proceedings of the AMIA annual Symposium. AMIA 1997. (1998)
11. Omar, H., Rachid D., Mohamed, T., Zouhair, I., Abdellah, B.: An Approach for Big Data Interoperability. Journal of Engineering and Applied Sciences, 13 (17), 7323-7328 (2018)
12. Raulok A., Sergio P., David P., Enrique A., Ruud van S., Sheng Y., Marian T., Francesca B., Carlos F., Alejandro P., Victor M.: A semantic interoperability approach to support integration of gene expression and clinical data in breast cancer. Computers in Biology and Medicine. 87 (1), 179-186 (2017)
13. Sanjoy S., Ningthoujam M., Dutta C., Kumar S., Potsangbam P., Chetia L., Nahar S., Sarker N., Basar A.,Das T.: NoSQL Data Model for Semi-automatic Integration of Ethnomedicinal Plant Data from Multiple Sources. Phytochemical Analysis 25 (6) 495-507 (2014)
14. Spoladore D.(eds) IFIP International Federation for Information Processing: Ontology-Based Decision Support Systems for Health Data Management to Support Collaboration in Ambient Assisted Living and Work Reintegration, pp. 341-352 (2017)