Journal of Information & Knowledge Management
Vol. 19, No. 3 (2020) 2050021 (19 pages)

DOI: 10.1142/S0219649220500215

World Scientific
www.worldscientific.com

Using Semantic Web Technologies and Multi-agent System for Multi-dimensional Analysis of Open Health Data

Salma El Hajjami[*,¶], Mohammed Berrada[*,‖], Mostafa Harti[†,**] and Gayo Diallo[‡,§,††]

[*]*IASSE Laboratory*
ENSA, Sidi Mohammed Ben Abdellah University
Fez, Morocco
[†]*LIMS Laboratory*
Sidi Mohammed Ben Abdellah University
Fez, Morocco
[‡]*Team ERIAS, Bordeaux Population*
Health Research Centre
INSERM, UMR 1219 Université de Bordeaux
Bordeaux, France
[§]*LaBRI, CNRS, UMR 5800*
Université de Bordeaux
Talence, France
[¶]*salma.elhajjami@usmba.ac.ma*
[‖]*mohammed.berrada@gmail.com*
[**]*mostafa.harti@usmba.ac.ma*
[††]*Gayo.Diallo@u-bordeaux.fr*

Published 23 July 2020

Abstract. Recent years have seen Social Web becoming a global phenomenon, which is being increasingly important in our daily lives. Millions of users are chatting on the Web and social networks and expressing their feelings and opinions about the latest outbreaks, symptoms, illnesses and new drugs. These opinions contain a large amount of data, which are destined to become a major source of information for business intelligence, as they are largely informative and therefore interesting to be dealt with in a decision-making process, in order to evaluate and improve the performance of health system. However, this source of information is currently underutilised. This work describes an approach to creating an analytical health framework that allows the integration and multi-dimensional analysis of available health data, with particular attention to socially generated data, using Semantic Web (SW) technologies and multi-agent systems.

Keywords: Social Web; business intelligence; multi-dimensional analysis; Semantic Web; multi-agent system.

1. Introduction

The health system faces very important challenges to improve the overall performance of the system. Different communities are interested in this subject from

different angles, ranging from technical questions to organisational aspects. An important aspect of this area of research is the integration of Social Web data into the system, particularly due to the rapid and growing development of many Social Web sites and online discussion communities. These social media can be general social networks like Twitter (500 million users worldwide) or other dedicated networks like PatientsLikeMe (growing, currently has more than 187,000 members and covers more than 500 patients) (PatientsLikeMe, 2004).

Millions of users voluntarily share information about their illness, treatment and experience on the Web. This leads to a large amount of data that we refer to as "social health data". They complement those available within the electronic health records (EHRs). The latter constitutes an electronic version of the medical history and conditions of a patient that are collected and stored by health practitioners (for example, clinicians). Social health data express too emotional, psychological attitudes, opinions and comments from many cases of experiences, practices and other behaviours related to health. EHR data are protected by HIPAA and HITECH privacy regulations (The U.S. Department of Health & Human Services, 2003) and locked into different EHR systems, and therefore difficult to share. This makes social health data a unique opportunity to examine healthcare from the patient's perspective to identify health problems and contribute then to improving the quality and performance of the overall health system. Social health data represents an unprecedented potential for innovation for a health system. Some examples of their potential benefits are the following:

— identify the most worrying health problems;
— perform epidemiological and pharmacovigilance studies;
— identify risk factors for disease;
— identify how patients perceive certain treatments and practices;
— monitor the effectiveness of treatments;
— identify health-related trends;
— reveal patients' attitudes towards health.

However, integrating and analysing social data is not a trivial task due to the scalability, complexity and heterogeneity of this kind of data. The following questions arise in that context:

— How to semantically integrate socially generated data?
— How to analyse these data in the absence of a relevant model?
— How to render the results of this analysis?

The objective of this study is to demonstrate that it is possible and feasible to build promising alternative solutions to traditional health system framework to improve its performance. In our view, taking into account "social data" can provide effective decision-support systems to help practitioners and researchers make optimal and effective decisions in dynamic and complex environments. Our approach consists in extracting data from several social networks, aggregating them and developing a

semantic model to meet the demands of high-level users. In addition, we show how adding an analytical component can help operators in understanding these social data.

The rest of the paper is organised as follows. Section 2 presents the preliminary concepts and Sec. 3 overviews related works on integrating the healthcare data. It raises the main research issues too. In Sec. 4, we present the semantic approach that we used to integrate social data, and in Sec. 5, the designed and implemented prototype architecture is reported. Then we conclude and outline the future work in Sec. 6.

2. Preliminary Concepts

In this section, we introduce the concepts that we will use in this paper.

Linked Data (LD) (Heath and Bizer, 2011; Bizer *et al.*, 2011) is an initiative of the W3C (World Wide Web Consortium)[1] aiming to promote the publication of structured data on the Web, not in the form of data silos isolated from each other, but linking them together to form a global information network. It relies on Web standards, such as the traditional HTTP and Uniform Resource Identifier (URI) — but rather than using these standards only to facilitate navigation by humans. LD expands them to share information equally between machines. This makes it possible to automatically query data, regardless of where they are stored, without having to duplicate them. Tim Berners-Lee has defined four pillars to support the "Linked Data" initiative (Shadbolt *et al.*, 2006):

(1) Use unique URIs to identify things.
(2) Use HTTP URIs that exist on the Web (URLs).
(3) Provide usable information, readable by humans and by machines, using open formats such as Resource Description Framework (RDF) or SPARQL through the URI.
(4) Mesh the initial URI by associating external URI addresses to improve the discovery of other information on the Web.

Resource Description Framework (Klyne, 2004) is a language for describing resources. It has a graphical model rendering and an XML serialisation. It allows to formally describe Web resources and their metadata, in order to allow automatic processing of such descriptions. Developed by the W3C, RDF is the basic language for the Semantic Web (SW). There are other RDF syntaxes which appeared later, seeking to make its reading more understandable; this is the case, for example, of Notation3 (or N3). An RDF document is a set of triplets. Each triplet is constituted of a subject, a predicate and an object:

— The "subject" represents the resource to be described.
— The "predicate" represents a type of property applicable to this resource.
— The "object" represents a datum or other resource: it is the value of the property.

[1] http://www.w3c.org.

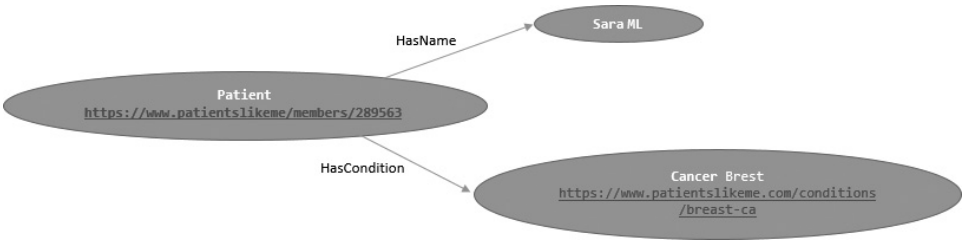

Fig. 1. Example of an RDF triple.

For example, a patient P1 could be described using his name by a triple as follows: ⟨P1, **hasName**, "Sara ML"⟩, ⟨P1, **hasCondition**, "Breast Cancer"⟩, the patient is related to Sara ML by name and to Breast Cancer by condition, as shown in Fig. 1.

SPARQL is a protocol and query language (Prud'hommeaux, 2008) for querying RDF graphs, developed by the W3C. It is considered one of the key technologies of the Semantic Web. In 2008, version 1.0 became an official W3C recommendation. Version 1.1, that saves and merges data from different sources, has become a recommendation since 2013. The SPARQL query evaluation mechanism is based on sub-graph matching: RDF triples are interpreted as oriented nodes and graph edges, and the query graph is matched to the data graph, instantiating the variables in the graph. The selection criterion is expressed as a graph pattern in the WHERE clause, made up of a Basic Graph Pattern (BGP). The operator "." represents the conjunction of graphical models. SPARQL supports aggregation functions and the GROUP BY clause, which are relevant for Online Analytical Processing (OLAP).

Multi-agent systems (MASs) refer to a more or less extensive set of actors who communicate with each other (Mansour, 2007, p. 39). The whole of this community aims at the accomplishment of a precise task, where each one has specific objective and offers. Here, a service means that each agent is able to perform certain tasks autonomously and communicates the results obtained to a receiver actor (human or software). A software agent is a classic programme called "smart". An intelligent agent is supposed to have the following intrinsic characteristics (Ferber, 1995): Intuitiveness, an agent must be able to take initiatives and to perform the actions assigned to him; Reactivity, an agent must be attentive to the actions of his environment and act accordingly; and Sociability, an agent must be able to communicate with other agents and/or users. Moreover, the agents can be mobile and possess the characteristic of moving autonomously through an acceptor network to perform various tasks (Wooldridge, 2009).

3. Related Works

In the following, we review previous researches in frameworks for multi-dimensional modelling, agents-based semantic data management and integration of Social Web data for health system.

3.1. *Frameworks for multi-dimensional modelling and analysis*

Multi-dimensional modelling aims to organise data in such a way that OLAP applications are efficient and effective. Most current frameworks are based on the RDF Data Cube (QB) vocabulary, which was specifically designed to publish data cubes on the Web. QB does not provide a mechanism to represent several levels on one dimension and the relationships between the levels of the schema. In response to this question, Kämpgen *et al.* (2012) introduced QB-like, an extension of the QB model, in order to represent statistical data in a multi-dimensional model. They showed how to perform OLAP analyses on data published in QB using the SPARQL query language. However, their solution does not support the hierarchical structure at several levels, and multiple hierarchies in a dimension.

Etcheverry and Vaisman (2012a) introduced a new RDF vocabulary called Open Cube (OC) for multi-dimensional modelling of RDF data. OC provides a set of classes and properties to model the different structures of the multi-dimensional model (dimensions, attributes and measures), including hierarchical relationships between the dimension attributes. From the RDF collections described using OC, different OLAP manipulations can be performed directly via queries expressed in SPARQL. Although this solution is based on the multi-dimensional modelling of the RDF data and allows expressing the OLAP operations in terms of SPARQL, its main limitation is the reuse of the data already published in QB, which is standardised. The work in Etcheverry and Vaisman (2012b) introduces the QB4OLAP vocabulary (Fig. 2). The latter extends and remains compatible with QB to support the multi-dimensional modelling of Linked Data. In Etcheverry *et al.* (2014), an

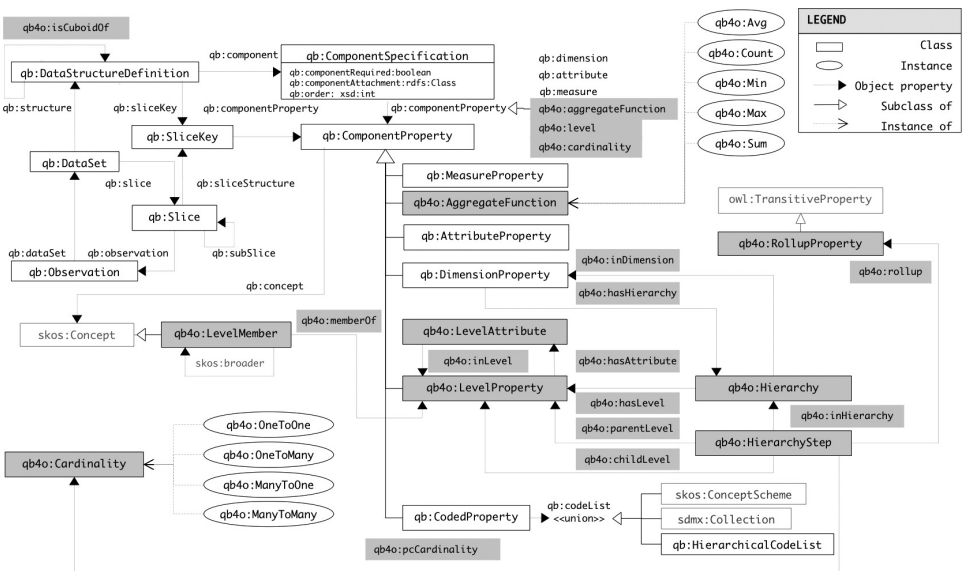

Fig. 2. The QB4OLAP vocabulary (Etcheverry and Vaisman, 2012b).

Table 1. Comparison of vocabularies.

MD concept	QB-like	OC	QB4OLAP
Dimensions	✓	✓	✓
Levels	✓	✓	✓
Level members	✓	✓	✓
Relationship between level members and levels	✓	✓	✓
Roll-up relations between levels	✓	✓	✓
Roll-up relations between level members	✓	✓	✓
Multiple hierarchies in a dimension		✓	✓
Measures	✓	✓	✓
Multiple measures in a cube	✓	✓	✓
Aggregation functions	✓	✓	✓
Ability to reuse data in QB	✓		✓

extension of QB4OLAP is proposed; it supports several hierarchies in a dimension and cardinalities between level members. In addition, mechanisms to transform an existing relational data warehouse to the QB4OLAP scheme have also been presented in this work.

We have summarised in Table 1 the vocabularies that allow the modelling of multi-dimensional data in the Semantic Web in order to perform an OLAP analysis. The vocabulary QB-like (Kämpgen *et al.*, 2012) does not have sufficient capacity to manage OLAP. OC is a modelling language that supports multiple hierarchies in a dimension, however the data already published in QB cannot be reused by OC. QB4OLAP, an extended version of QB, offers more features to support OLAP.

3.2. *Agent-based semantic data management*

The management of data on the Web consists in a series of tasks to be performed, so it is necessary to use a cooperative system so that these different tasks can be carried out in a coherent way. From this point of view, it is natural to introduce the notion of agent. A software agent is a conventional programme called intelligent. Multi-agent systems refer to a more or less extensive set of actors that communicate with one another (Adadi *et al.*, 2014). This set aims at the realisation of a well-defined task, where each agent has a specific objective, offers services to perform some tasks, autonomously, and communicates the results obtained to a receiving actor (human or software).

Various research works have been done to manage data on the Web using multi-agent systems. The InfoSleuth Project (Bayardo Jr. *et al.*, 1997) extends the capabilities of the Carnot Technologies developed at Microelectronics and Computer Technology Corporation (MCC) in dynamically changing environments, which specialised in integrating heterogeneous information bases. InfoSleuth is accomplished by collaborative agents, and Java is used as a common agent wrapper. The aim is at automating the collection and analysis of dynamic data distributed over the Web, where each agent is an autonomous process that specialises in a particular

service. A group of agents collaborate with each other to accomplish complex tasks of collecting and analysing dynamic data.

Ontologies are also used to impose a uniform view on data from different sources. Gandon *et al.* (2002) proposed an innovative approach to the management of an organisational memory combining ontology engineering, Semantic Web and multi-agent systems in an integrated solution. The keystone of Gandon *et al.* (2002) is a common and shared ontology (Fensel, 2001) ensuring the coherence of memory and communications between agents. Chafik and Kazar (2009) present a Semantic Web services discovery architecture using agent technology and ontologies. This architecture integrates software components and exploits a domain ontology that is used in the discovery phase of Web services. It facilitates the automatic discovery of services since it allows to refine the search process that matches a request and offers of services. In Fatima *et al.* (2014), an approach for Semantic web service using domain ontological retro-engineering to generate semantic links is presented. This approach consists of two phases: the extraction of relevant information and the analysis of the extracted information using ontology of the domain using a similarity criterion. In Li *et al.* (2014), an agent-based architecture is proposed for the management of data sources in a dynamic environment (Fig. 3) based on Bayardo Jr. *et al.* (1997), and it allows managing related data (RDF, RDFS, etc.) in a virtual way, meaning that it does not load the data into a local data store. In addition, it integrates LD sources while maintaining their local autonomy. This system has the ability to monitor, manage and query data sources on LD by using SPARQL queries.

In our approach, we adjust this agent-based architecture for the selection and retrieval of open social data.

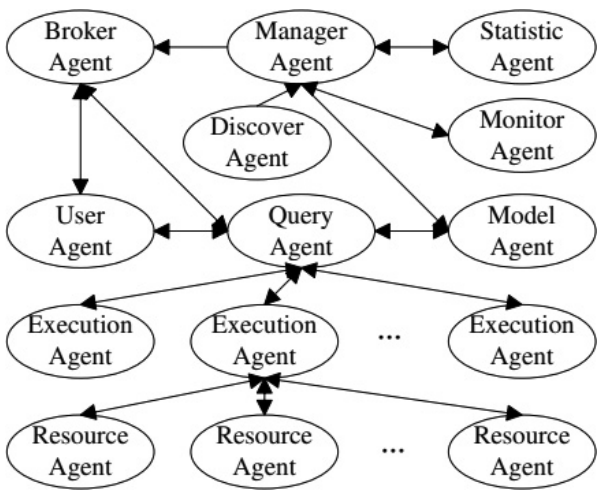

Fig. 3. LDMS architecture (Li *et al.*, 2014).

4. Proposed Approach

The health domain is a highly complex and dynamic domain where decision-makers often rely on decision support systems to analyse and compare the actors involved (clinical trials, procedures, etc.). In recent years, a large amount of data is scattered among many social networking sites and online discussion communities, where the patient experiences (diseases, treatments, side effects, etc.) are shared. These experiences can serve as important data sources to provide collective intelligence and awareness of health problems in real time as well as help evaluate and improve health system performance. The challenges of using Social Web data include the data volume being large, but distributed and in a highly unstructured form. Appropriate data collection, cleaning and aggregation efforts are needed to transform these data for meaningful use. This work describes an approach to creating an analytical health framework that allows the integration and multi-dimensional analysis of health data sources, with particular attention to socially generated data. We first created a health knowledge base in which data from multiple open sources are included. Data from these sources are integrated and linked via Semantic Web technologies. The main contributions of this work can be summarised as follows:

— The development of a health data model. This model supports data characteristics from many different sources. The model focusses on health and on patient-generated data, such as conditions, treatments and related information, with a focus on integrating health data from Social Web. At the implementation level, the data is stored as RDF triplets.
— Providing a process of automatic data integration and linking using multi-agent systems. Data is collected automatically from multiple sources and transformed into an RDF format.
— The development of an analytical service allowing a multi-dimensional analysis of Social Web data focussed on medical conditions, treatments and symptoms.

4.1. *Integration of Social Web data for healthcare*

Semantic Web technologies have been widely used as a framework for the integration of public data, to create links between distributed resources in heterogeneous data sources. Semantic Web principles require the use of URIs to identify resources, RDF/RDFs to represent information and generally the use of SPARQL to access information. Sheth and Ramakrishnan (2003) examine the viability of the Semantic Web for data integration, while in Harth and Gil (2014), a scenario of integration and querying of geospatial data with Semantic Web technologies is described. Specia and Motta (2007) integrate folksonomies into a social tagging system with an ontology. In the health field, the study reported in Ae Chun and MacKellar (2012) proposes a preliminary model of semantic integration of different sources of health data that can help annotate social health blogs. The work in Ji *et al.* (2017) uses an integrated semantic model to create a machine-readable encoding of the content

semantics of various open health data sources, especially social data sources. Tofferi *et al.* (2004) study an approach to extract information from the Social Web for personalising health. They pointed out that the available data sources do not provide APIs for integration with third-party applications.

4.2. *Benefits of using Semantic Web technologies and multi-agent systems*

We discuss here the benefits of using LD technologies and the multi-agent systems for the health field. The first benefit of Semantic Web technologies is syntactic interoperability, as these technologies provide a common technical infrastructure for exposing data in a uniform format, allowing easy access to data and facilitating data integration between sources. Thus, they facilitate the automatic and dynamic consumption of these datasets in new applications (decision support). Another advantage comes from the links that are established between datasets. While these links facilitate data integration primarily, they also serve to enrich a data source with additional knowledge. The agent technology provides a clear, modular structure, and therefore easy reuse and ideal maintenance. In addition, agents are able to communicate with each other without external intervention. The exchange of information is then simplified and the execution of the tasks is clarified, with each agent performing a well-defined set of tasks. In addition, the concept of mobility is a very important aspect.

Indeed, the agent can move where the information is and return to its original location. It will thus be able eventually to autonomously create remote databases, associated with other agents capable of providing consistent services to the user, for the collection of relevant information from the Web, but also through a local network, for example.

In the next sub-section, we describe in detail our approach for the multi-dimensional analysis of Social Web data in the context of health; we first begin with elaborating the design and analysis models of our framework.

4.3. *Design and analysis models*

4.3.1. *Use-case diagram*

A decision-maker is the main actor. The basic use cases that will be highlighted to assist in decision-making are as follows (Fig. 4):

— Extract data from a Social Web.
— View the data in cubes.
— Apply OLAP operations to the data cube.

4.3.2. *Class diagram*

In order to provide an analytical framework that meets the needs of decision-makers, we should understand the data that the framework needs to manage. From these

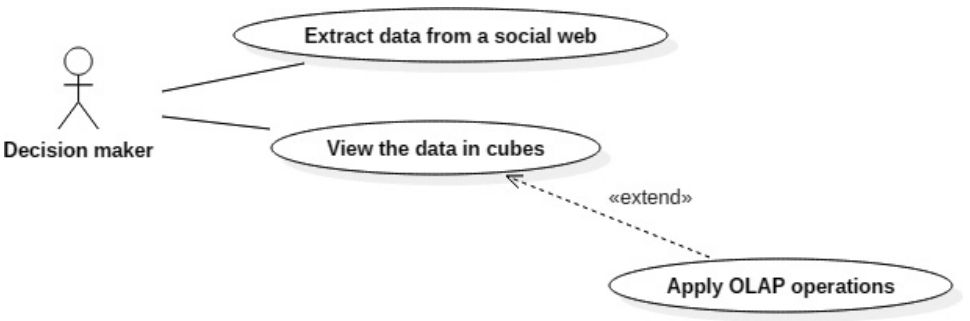

Fig. 4. Use-case diagram.

Table 2. Classes of the model.

Classes	Comments
Patient	Includes all people with specific diseases who share information about their illness, treatment and experience
Condition	Includes all diseases that can be shared by patients
Treatment	Includes all treatments that can be shared by patients
Tweets	Groups all tweets about treatment or illness that can be shared on Twitter by the patients
Symptom	Groups all the symptoms of each condition

needs, it is possible to derive the following central classes: patient data, medical condition, symptom, tweets and treatment. Figure 5 depicts a class diagram describing the concepts we need to model, as well as the relationships between them.

In this model, there are five main classes that are "Patient, Condition, Symptom, Treatment, Tweets" listed in Table 2. And also, a patient may have one or more diseases, as a certain illness may have several symptoms, the disease may also have a lot of treatments. Finally, a treatment can have several "tweets".

4.3.3. *Linked knowledge representation model*

Publicly available health data is hosted on a variety of sources, including, PatientsLikeMe, Twitter, etc., and these sources describe and provide access to data through different representations and platforms. To perform intelligent analysis across multiple data sources, we use a lightweight ontology to build an integrated knowledge base. Using the entities in the data model, we developed a lightweight ontology by structuring entities and relationships ontologically as a concept hierarchy as shown in Fig. 6. An ontology typically consists of classes, properties and the relationships between classes (for example, IS-A, PART-OF) and instances. In our model, classes and concepts are derived from central concepts that include condition, symptom and treatment. To detect the instances of different classes in the ontology we used UMLS (Bodenreider, 2004). If a health term matches with the one in

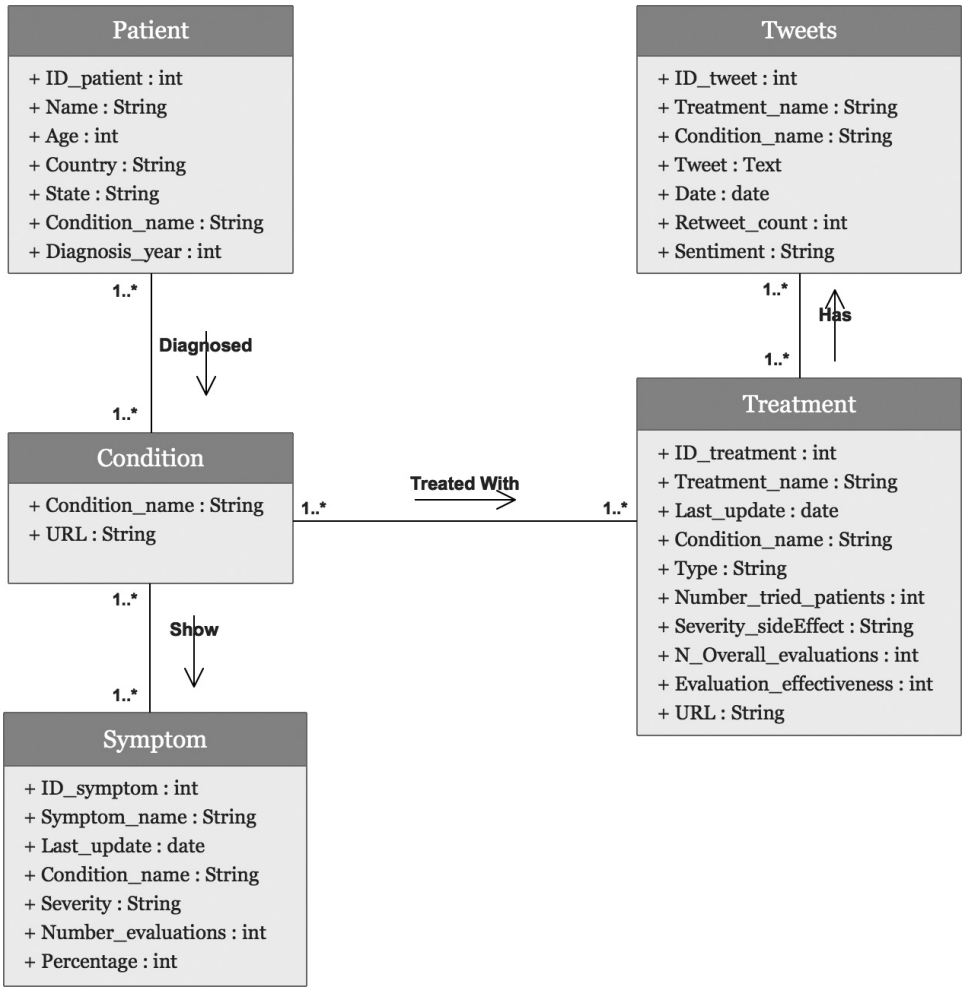

Fig. 5. Class diagram for social health data.

the UMLS, then the category of the term is traversed to classify it as one of the class terms.

Extracted health entities from sources can have relationships among themselves. The relationships in the semantic model shown in Fig. 6 are used. We represented the extracted instances using the (RDF) triple standard representation of *subject*, *predicate* and*object*. The Condition class and Treatment class are the central concepts in the model. The Patient class has the *HasCondition* relationship to the Condition class. The Condition class has, respectively, the *HasSymptom* and the *HasTreatment* relationships with the Symptom class and the Treatment class. The Treatment class has the *HasSideEffect* relationship with the Side-Effects class. For instance, a patient's profile at PatientsLikeMe reports the condition, treatment

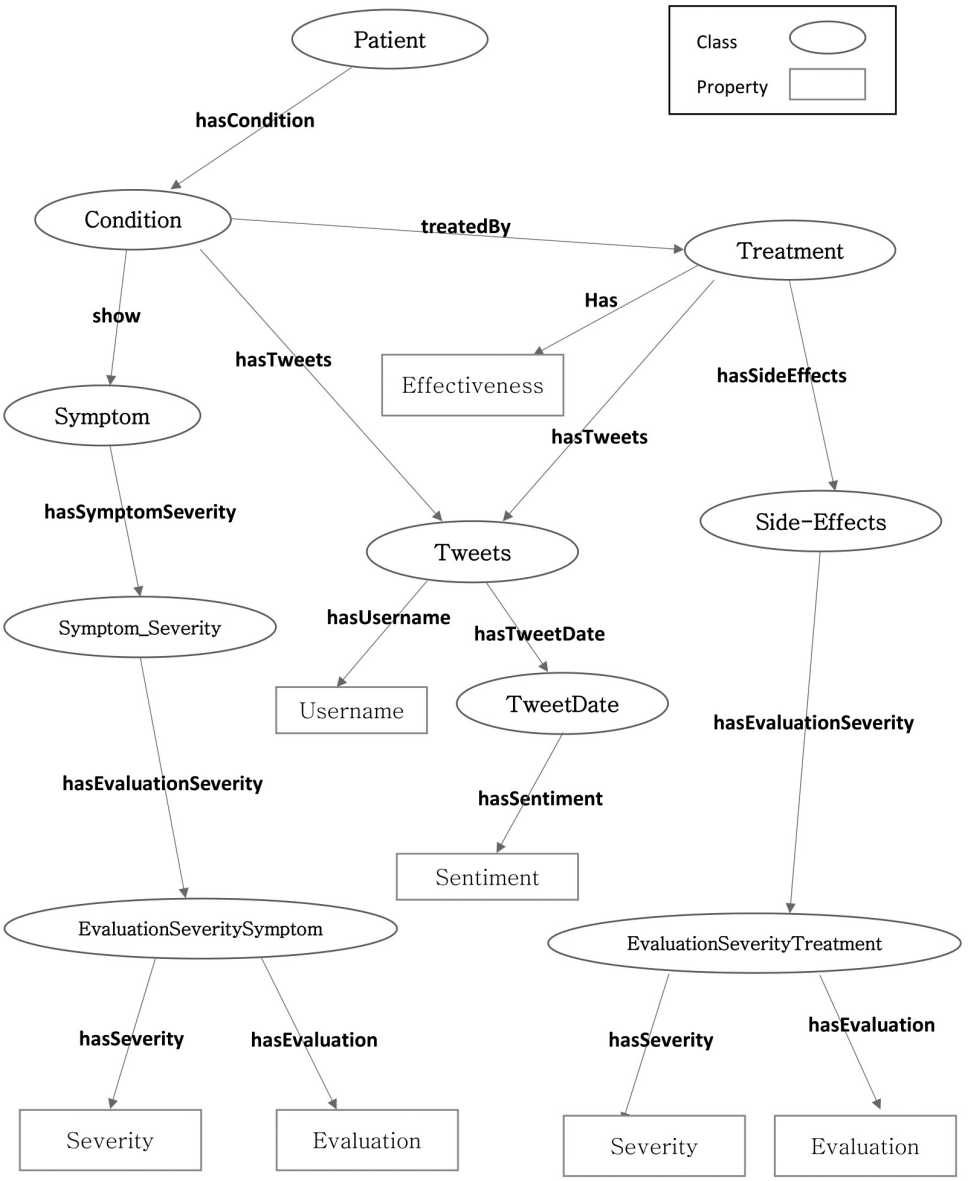

Fig. 6. Semantic model for social health entities.

procedures and prescription drugs taken. This particular patient's health experience can be described in the (RDF) model, whose snippet is shown in Fig. 7. In this example, the subject is the patient's profile webpage, predicates are "the patient has condition, has treatment ..." and the objects are identified as #385 and #279 in the PatientsLikeMe.com site.

```
<RDF xmlns =  "http://www.w3.org/1999/02/22-rdf-syntax-ns#" xmlns :rdf = http://www.w3.org/1999/02/22-rdf-syntax-ns#>
    <Description about = "https://www.patientslikeme.com/members/477077">
        <Patientsns:HasName>firefly84</Patientsns:HasName>
        <Patientsns:HasLocation>WI, United States</Patientsns:HasLocation>
        <Patientsns:HasGender>male</Patientsns:HasGender>
        <Patientsns:HasAge>33</Patientsns:HasAge>
        <Patientsns:HasCondition rdfressource ="https://www.patientslikeme.com/conditions/385-autonomic-neuropathy"/>
        <Patientsns:HasTreatment rdfressource ="https://www.patientslikeme.com/treatments/show/279"/>
         …
    </Description>
</RDF>
```

Fig. 7. Representation of a patient's health data, conditions and associated properties in RDF triple.

5. Analytical Health Framework

5.1. *Architecture*

Our proposed approach is based on an evolving architecture offering great flexibility and strong structuring based on a set of agents. The architecture shown in Fig. 8 extends the architecture that we presented in our previous work (El Hajjami *et al.*, 2017), where the different interactions between agents are presented. The prototype system mainly consists of the modules discussed in the following sub-sections.

5.1.1. *Data collection module*

Data collection is the obvious first step. It is about collecting data according to the needs of different sources of the Social Web in different formats. Data sources include PatientsLikeMe and Twitter. These types of data are essential and must comply

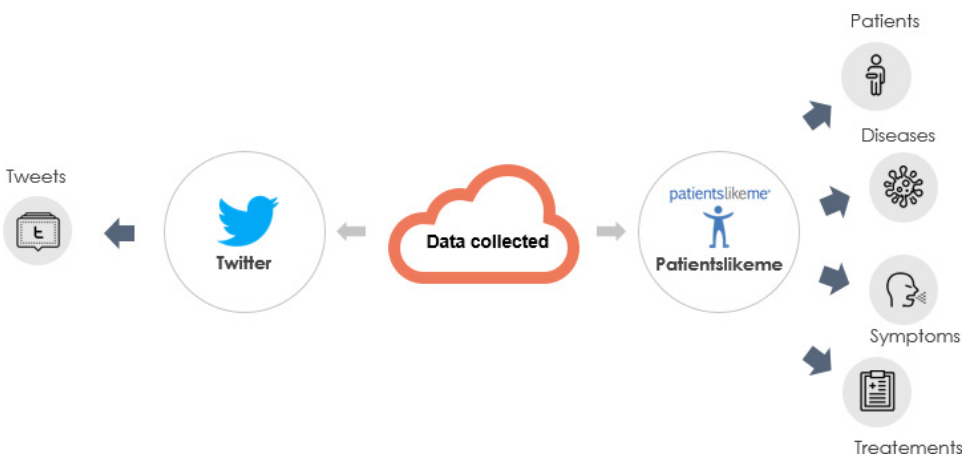

Fig. 8. Data collection.

with national and international laws and legislation. It is important to maintain patient confidentiality. Indeed, disidentification and anonymisation measures are used in the second phase to mask sensitive patient information. This component is managed by an agent whose role is to extract the data necessary for analysis in a virtual and automatic way using a Web indexing robot specific to each social network.

5.1.2. *Data transformation module*

Once the data is available, this module consists in filtering and classifying the data according to their structure and in carrying out the necessary transformations in order to perform a meaningful analysis; these tasks are carried out by agents.

More broadly, in the first step, filtering, enriching and appropriately processing the information collected are necessary to improve the quality of the data before the data transformation phase in a unified model. This module uses Algorithm 1 for matching terms and additional information specified in the UMLS Metathesaurus, which provides common vocabulary and semantics for several terms referring to the same concept, in order to recognise identical concepts; if we find that the medical term used by the user in his/her post is similar to at least one medical term in the list of synonyms, with a threshold more than the degree of thresholding that we have defined (for example, 70%), we will keep the information about the medical term used.

However, collected data may contain sensitive information, which makes it extremely important to take sufficient precautions when processing and storing the data. The HIPAA and other regulations govern several aspects of privacy, including the use of anonymisation, patient consent to the use of their data and non-discrimination. A data anonymisation approach is therefore used to guarantee the

Algorithm 1 Matching Term Algorithm

Require: $dataset, dict\_synonyms, threshold$
- Browse each *post* from the *dataset*;
- Search in the medical dictionary *dict_synonyms* the list of terminologies corresponds to the term (i.e. name of condition) used in the *post*;
- Get the different synonyms terms and add the terms in a list *condition_terms*;
- Calculate the similarity between the *post* and the list of medical synonyms *condition_terms*;
- Initialize the filter *treshold* at 70%;

for $post \in \lceil ataset$ **do**
  - Extract ($[dict\_synonyms]$, *post*)
  if $Similarit\_yratio \geq threshold$ **then**
    - Store the *post*;
  end if
end for

Table 3. Non-anonymised data of patient.

Patient ID	Name	Tel. number	Date of birth	Sex	Country	City	Condition name	Diagnosis year
1	Hind Alim	0645897562	18/07/1989	F	Morocco	Fez	Breast cancer	2016
2	Ahmed Samir	0789561423	02/01/1949	M	Canada	NL	Amyotrophic lateral sclerosis	2002
3	Julie M	0698158795	15/02/1994	F	France	Bor	Autism spectrum disorder	2007

Table 4. Anonymised data of patient.

Patient ID	Name	Tel. number	Date of birth	Sex	Country	City	Condition name	Diagnosis year
1	Patient (1)	*********	18/07/1989	F	Morocco	Fez	Breast cancer	2016
2	Patient (2)	*********	02/01/1949	M	Canada	NL	Amyotrophic lateral sclerosis	2002
3	Patient (3)	*********	15/02/1994	F	France	Bor	Autism spectrum disorder	2007

confidentiality of sensitive data. Our approach to data anonymisation is to replace sensitive data elements such as name, social security number and telephone number with an unidentifiable value.

It is not really an encryption technique, so the original value cannot be returned from the encrypted value. In Table 3, there are eight attributes and three examples of records. Our approach consists to replace sensitive data elements such as name, social security number and telephone number with unidentifiable values; in this example, we replace each of the values of the attribute "Name" by "Patient (i)" and all the values of the attribute "Telephone number" by a "*", as shown in the anonymised Table 4.

In the second step, the transformation of the data into a unified format is carried out using the model presented in Fig. 6, which shows the different classes and properties of our RDF model. It allows improving the quality of the data before the analysis or modelling phases.

5.1.3. *Data modelling module*

Once the data has been collected and transformed, the modelling process is managed by another agents whose role is to model a multi-dimensional data schema by observing the facts through dimensions and measurement axes. To do so, we use QB4OLAP structural metadata because of its expressiveness and as it is specialised for OLAP. Thus, the construction of the data cubes is done on the fly according to the expressed analysis of the user.

Figure 10 shows an example of conceptual schema of data cube using QB4OLAP. The social health fact contains a measure (TopTrendingtreatement) that represents

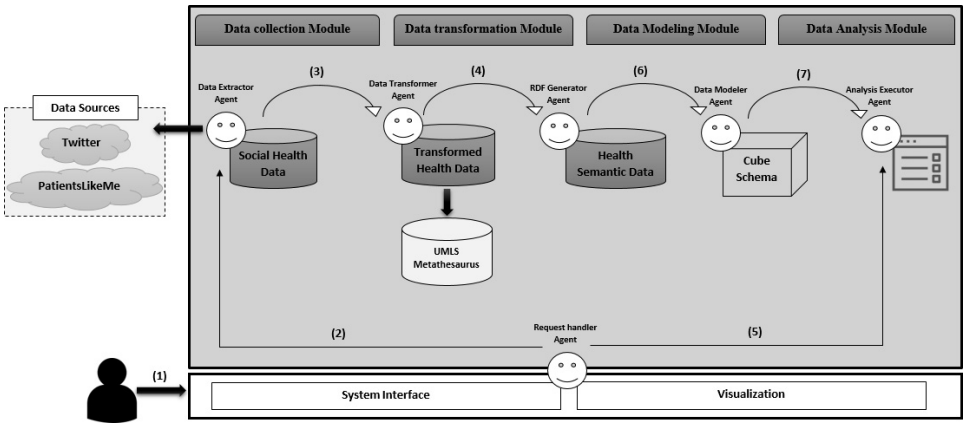

Fig. 9. System architecture.

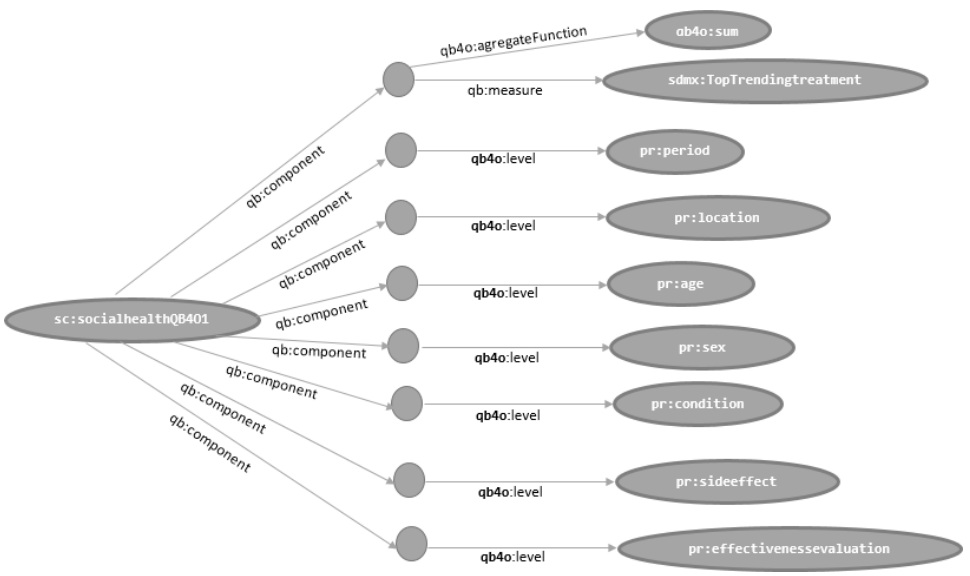

Fig. 10. Example of QB4OLAP representation of the social health data cube schema.

the top trending of treatment. This measure can be analysed according to the following analysis dimensions: age and sex of the patient, a period which represents the time of treatment tried and consists of two levels (month and year), location which consists of two levels (country and city), condition, side effect and the effectiveness evaluation.

5.1.4. *Data analysis module*

Data analysis contains analysis sub-modules to implement OLAP operations, and this implementation is carried out by agents to transform a user's request

(Roll Up, Drill Up, Drill Down, Slice, etc.) into SPARQL. The purpose is to retrieve the data and return the aggregated results to the user.

5.2. *Simple scenario*

We have instantiated six main agents offering services which allow the multi-dimensional analysis of the LOD on the Social Web.

- **Request Handler Agent.** This is the system driver. It acts as a proxy between the system and the user. It is responsible for overseeing agent migrations.
- **Data Extractor Agent.** Its role is to extract and collect the data according to the need of the analysis.
- **Data Transformer Agent.** Its role is to establish consistent semantics and reasoning capabilities between different data sources, in order to filter out the data collected.
- **RDF Generator Agent.** Its role consists in transforming the extracted data into a unified RDF format.
- **Data Modeler Agent.** It allows to model a multi-dimensional schema.
- **Analysis Executor Agent.** Its role is to translate into SPARQL queries the OLAP operators selected by the user.

A sample scenario is used to demonstrate how the agents of the described architecture interact with each other. When the system starts, a user submits a request through the user interface provided by the Request Handler Agent (1). After verification, the request is sent to the Data Extractor Agent (2), which extracts data according to the need for analysis. Data Transformer Agent (3) whose role is to filter the information retrieved, first uses additional information specified with UMLS concepts to derive data to identify different instances of terms representing the same concept and in the second step, it anonymises sensitive information. Once the data has been filtered out and processed, it transmits it sequentially to the RDF Generator Agent. The latter transforms the collected data into a unified RDF format (4) and transmits it to the Data Modeler agent (6), which deals with multi-dimensional modelling and generates a cube schema using the QB4OLAP vocabulary. Then, the Analysis Executor Agent translates the OLAP operators selected (5) into SPARQL queries and executes them on the multi-dimensional cube schema created (7). To resume, the Request Handler Agent displays the results according to the queries selected by the user (1). See Fig. 9.

6. Conclusion and Perspectives

The Social Web is considered as a major source of information for the health system as it constitutes a means of complementing patient's real-world data (de Lusignan *et al.*, 2015). In order to take advantage of it, the health system must be enriched by new approaches and methods to analyse and use data from outside the EHR data,

mainly on the Web and sensors. Including them in analyses allows multiple perspectives to decision-makers.

In this paper, we have described an approach that integrates and analyses health information from a variety of Social Web sources, mainly from PatientsLikeMe and Twitter. It is based on Semantic Web technologies and multi-agent systems. A semantic model for integrating health data has been defined for various data sources. A prototype has been developed for integrating and analysing the resulting semantic knowledge base related to health.

In future work, we aim at extending the approach to handle additional types of Social Web sources, and public datasets provided by government bodies. We plan after exploring all these data sources to build a semantic model that will be able to group all of these data together to generate knowledge. We also aim to implement these results in an extended platform that will contain numerous queries and analytical tools to better serve policy-makers and patients.

References

Adadi, N, M Berreda, D Chenouni and B Bounabat (2014). Multi-agent architecture for business modeling of web services composition based on WS2JADE framework. *International Review on Computers and Software*, 9, 1667–1674.

Ae Chun, S and B MacKellar (2012). Social health data integration using semantic Web. In *Proceedings of the 27th Annual ACM Symposium on Applied Computing*, pp. 392–397.

Bayardo, Jr. RJ, W Bohrer, R Brice, A Cichocki, J Fowler, A Helal, V Kashyap, T Ksiezyk, G Martin, M Nodine, M Rashid, M Rusinkiewicz, R Shea, C Unnikrishnan, A Unruh and D Woelk (1997). InfoSleuth: agent-based semantic integration of information in open and dynamic environments. *ACM SIGMOD Record*, 26(2), 195–206.

Bizer, C, T Heath and T Berners-Lee (2011). Linked data: The story so far. In *Semantic Services, Interoperability and Web Applications: Emerging Concepts*, pp. 205–227. Hershey, PA: IGI Global.

Bodenreider, O (2004). The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32(Suppl _1), D267–D270.

Chafik, B and O Kazar (2009). Une approche basée agent pour la découverte de services Web. In *Proceedings of the 2nd Conférence Internationale sur l'Informatique et ses Applications (CIIA'09)*, Saida, Algeria.

de Lusignan, S, L Crawford and N Munro (2015). Creating and using real-world evidence to answer questions about clinical effectiveness. *Journal of Innovation in Health Informatics*, 22(3), 368–373.

El Hajjami, S, M Berrada, M Harti and G Diallo (2017). Towards an agent-based approach for multidimensional analyses of semantic web data. In *Proceedings of the 2017 Intelligent Systems and Computer Vision (ISCV)*, pp. 1–6.

Etcheverry, L and AA Vaisman (2012a). Enhancing OLAP analysis with web cubes. In *ESWC 2012 — The Semantic Web: Research and Applications*, Lecture Notes in Computer Science, Vol. 7295, pp. 469–483. Heidelberg: Springer.

Etcheverry, L and AA Vaisman (2012b). QB4OLAP: a new vocabulary for OLAP cubes on the semantic web. In *Proceedings of the Third International Conference on Consuming Linked Data (COLD)*, Vol. 905, pp. 27–38.

Etcheverry, L, AA Vaisman and E Zimányi (2014). Modeling and querying data warehouses on the semantic web using QB4OLAP. In *DaWaK 2014: Data Warehousing and Knowledge Discovery*, Lecture Notes in Computer Science, Vol. 8646, pp. 45–56. Cham: Springer.

Fatima, B, H Abdelkader and B Djelloul (2014). Description et classification des services web sémantiques. *Nature & Technologie A: Sciences Fondamentales et Engineering*, 10, 41–47.

Fensel, D (2001). Ontologies. In *Ontologies*, pp. 11–18. Heidelberg: Springer.

Ferber, J (1995). *Les Systèmes Multi-agents: Vers Une Intelligence Collective*. Paris: Inter-Editions.

Gandon, F, R Dieng-Kuntz, O Corby and A Giboin (2002). Web sémantique et approche multi-agents pour la gestion d'une mémoire organisationnelle distribuée. In *13eme IC, Conférence Ingénierie des Connaissances*, Rouen, France.

Harth, A and Y Gil (2014). Geospatial data integration with linked data and provenance tracking. In *Proceedings of the W3C/OGC Linking Geospatial Data Workshop*, pp. 1–5.

Heath, T and C Bizer (2011). *Linked Data: Evolving the Web into a Global Data Space*, Synthesis Lectures on the Semantic Web: Theory and Technology. Mountain View, CA: Morgan & Claypool.

Ji, X, SA Chun, P Cappellari and J Geller (2017). Linking and using social media data for enhancing public health analytics. *Journal of Information Science*, 43(2), 221–245.

Kämpgen, B, S O'Riain and A Harth (2012). Interacting with statistical linked data via OLAP operations. In *ESWC 2012: Satellite Events*, Lecture Notes in Computer Science, Vol. 7540, pp. 87–101. Heidelberg: Springer.

Klyne, G (2004). Resource Description Framework (RDF): Concepts and abstract syntax. Available at http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/. Accessed on 5 March 2020.

Li, X, Z Niu and C Shi (2014). An agent-based linked data integration system. In *Proceedings of the Fifth International Conference on the Applications of Digital Information and Web Technologies (ICADIWT 2014)*, pp. 113–117.

Mansour, S (2007). Un modèle de gestion distribuée de groupes ouverts et dynamiques d'agents mobiles. Doctoral dissertation, Université de Pau et des Pays de l'Adour.

PatientsLikeMe (2004). Heal together, get answers, take charge. Available at https://www.patientslikeme.com/. Accessed on 5 March 2020.

Prud'hommeaux, E (2008). SPARQL query language for RDF. W3C Recommendation. Available at http://www.w3.org/TR/rdf-sparql-query/. Accessed on 5 March 2020.

Shadbolt, N, T Berners-Lee and W Hall (2006). The semantic web revisited. *IEEE Intelligent Systems*, 21(3), 96–101.

Sheth, AP and C Ramakrishnan (2003). Semantic (web) technology in action: Ontology driven information systems for search, integration, and analysis. *IEEE Data Engineering Bulletin*, 26(4), 40–48.

Specia, L and E Motta (2007). Integrating folksonomies with the semantic web. In *ESWC 2007 — The Semantic Web: Research and Applications*, Lecture Notes in Computer Science, Vol. 4519, pp. 624–639. Heidelberg: Springer.

The U.S. Department of Health & Human Services (2003). The HIPAA privacy rule. Available at https://www.hhs.gov/hipaa/for-professionals/privacy/index.html. Accessed on 5 March 2020.

Tofferi, JK, JL Jackson and PG O'Malley (2004). Treatment of fibromyalgia with cyclobenzaprine: A meta-analysis. *Arthritis Care & Research*, 51(1), 9–13.

Wooldridge, M (2009). *An Introduction to MultiAgent Systems*. Chichester: John Wiley & Sons.