Some contributions on large scale heterogeneous data and knowledge integration : application to the healthcare domain

Gayo Diallo

Habilitation à Diriger des Recherches (H.D.R)
Université de Bordeaux

BPH Centre INSERM 1219 - Team ERIAS: Equipe de Recherche en Informatique de Santé
École Doctorale SP2

Discipline : Santé Publique - Informatique de Santé

Some contributions on large scale heterogeneous data and knowledge integration

application to the healthcare domain

Par : **Gayo DIALLO**

Committee:

- **Rapporteur :** Eero Hyvonen, Professor, Aalto University, Finland
- **Rapporteure :** Florence Sèdes, Professor, Univ. Paul Sabatier, France
- **Rapporteur :** Fernando Ferri, Research Director, IRPPS - CNR Italy
- **Examinateur :** Guy Melançon, Professor, Univ. Bordeaux, France
- **Examinateur :** Richard Chbeir, Professor, Univ. Pau, France

Date of Defence : 18 june 2019

Remerciements

I would like to sincerely thank here the members of my Jury, my rapporteurs who accepted the difficult work of rapporteurs, my examiners for the time they took to read this manuscript and to enrich the exchanges of the HDR defence.

Many thanks to all the people with whom I have been able to collaborate throughout these years, and for many of them I still continue to collaborate (PhD thesis years, postdoc, MCF). A thank you to the members of ERIAS who welcomed me and facilitated my insertion at the University of Bordeaux. I do not forget the colleagues of the LaBRI who have been welcoming me since 2014. May my colleagues from INSERM 1219 find here the expression of my best feelings for the good atmosphere of the Lab.

My family has been supporting my long working hours, my absences for missions, etc. for many years. May they find here the expression of all my gratitude.

The defence of an HDR marks a new course. I am delighted to count you among those who will share these new adventures.

Gayo DIALLO

Forword

A short journey into a multidisciplinary research associating ICT and Health

I am sitting on my modest knowledge to explore the depth of my ignorance.
[Cerno Samba Mombeya (1765-1852)]

I defended my PhD thesis in December 2006 at Laboratory TIMC IMAG Grenoble (France) on a topic related to Ontology-based homogeneous data access. The addressed questions in this thesis fall into the broad area of Semantic Web (SW) technologies applied to the health domain. At that time, it was clear to me that I wanted to pursue a career in the academic field. The SW was relatively in its infancy with very few solutions in the market. From the lessons learned from my PhD, the Life Sciences domain was one of the promising application field of this new vision brought by SW. I had approached or been approached by some teams for a possible PostDoc as this is a tradition. I ended up joining the UK and City University of London, working on the EU-funded FP6 project, Semantic Web Browser for Life Sciences (SeaLife). My main duty was the development of the background knowledge that will be used within the project and to design a framework for the evaluation of the final implemented prototypes. SeaLife was for me a fantastic opportunity to pursue into the SW field while applying it on use cases related to life sciences and biomedical domain. The SeaLife Browser aimed to offer a new dimension of context-based information integration by dynamically providing contextual information related to the Ontology terms that have been identified.

To me, it was a great opportunity to contribute to a multidisciplinary project which aims at providing easy access to disseminated information and resources in the life sciences' online databases. I started my PostDoc in March 2007 with a one year fixed contract, which I extended for six more months before to chose to go back in France. Beside leading the design and development of the background knowledge in the form of an Ontology for the project, I had in charge introducing some Semantic Web based technologies into the development of the National electronic Library of Infection (NeLI) in UK [1]. NeLI is a UK-based digital library bringing together the best available online evidence-based, quality-tagged resources on the investigation, treatment, prevention and control of infectious diseases. Given the variety of users of this type of portal, one of the challenges was personalising of provided information according to users' profile.

I came back in France in October 2008 for a second PostDoc at the Laboratory of Applied Computer Science, LISI/ENSMA at Futuroscope Poitiers, working within the Data Engineering team (Pr. Yamine Aït Ameur). I had the chance then to work in a context which mix personalising information and access to data at the semantic level. This was possible, thanks to the Ontology-Based Database, OntoDB, which was being developed at that time at LISI. I had in charge the supervision of a MSc student and participate in the supervision of the PhD thesis of Dilek Tapucu who focused on information personalising. With the postdoctoral position at LISI I had the opportunity to observe the different

1. http ://www.neli.org.uk/IntegratedCRD.nsf/NeLI_Home1 ?OpenForm

long-term academic job opportunities that could be offered and to better be prepared for the recruitment as Associate Professor in France.

In September 2009 I did accepted the offer for a position "Computer Science and Health" at the Bordeaux Institute of Public Health, Epidemiology and Development (ISPED)/ University of Bordeaux Segalen (which was the Health related University at that time before the reunification of 2014 which created the University of Bordeaux). With a research activity at INSERM U897 within a young team which was being set up in Health Informatics (2 permanent staff at that time). I was really interested in the challenge to set up a multidisciplinary team in a field which was being renewed in France. It associates both academics with computer sciences background and health professionals (medical doctors). And I have the chance to coordinate currently. The team investigates on providing methods and tools for healthcare data integration, in order to facilitate their secondary use in different aspects of Public Health. Since 2014, with the desire to be close to colleagues in Computer Science (section 27 of the National Steering Committee for University) I am an associate member of the Bordeaux Laboratory in Computer Sciences (LaBRI), the main IT Lab in the Bordeaux campus. One of the challenges from my personal perspective but also at the multidisciplinary team level is to maintain a balance between methodological and theoretical research as in Section 27 and applied research as required in health informatics and public health.

After this brief historical introduction of the beginning of my academic career, let's report more about the scientific aspect of my activities.

Content

Première partie

Research Work Synthesis

Chapitre 1

Overview my research work

Contents

1.1 Introduction

The recent years have seen an increase in the availability of digital data as never before. This is thanks to the development of new information communication and technologies (ICT) and medias from Web technologies : we have entered the era of Big Data for some time now. The situation is particularly noticeable in the health and life sciences sector. From the health perspective, data are routinely produced and **electronic health records (EHRs)** in developing countries particularly, have become the standard for clinical practice. These data collected by healthcare professionals provides access to a patient follow-up history more easily when the context is the same service or to a lesser extent in the same hospital. However, when it comes to having an integrated view of these data for the purpose of a global perspective from several departments/hospitals, the task becomes more complex. One of the main reason is the implicit meaning that these data convey. It often depends on common knowledge shared by health professionals that may not be explicitly expressed. These leads to the necessity of formal metadata that could be used to describe semantically the data.

From a more public health perspective, where the activities relies on data production and analysis, enabling an efficient use of the widespread available data is of major interest. Recently, developed guidelines for improved data availability and reusability-entitled FAIR Principles (Findability, Accessibility, Interoperability, and Reusability) have been proposed to address the issue of reusability of scholarly available data (Wilkinson et al.,

2016). This initiative addresses a part of the problem. An extension has been suggested in (Holub et al., 2018) which would facilitate access to complete provenance information, which is a prerequisite for reproducibility and meaningful integration of the data.

This data reuse issue has been identified since 2007 by the American Medical Informatics Association which has emphasised the value of **secondary use of medical data** (Safran et al., 2007). Further, recently in France, the report of Cedric Villany on Artificial Intelligence, entitled "Donner du sens à l'intelligence artificielle : pour une stratégie nationale et européenne" [1] dated last March 2018 has identified the unlocking of health related data at the local and national level as a powerful leverage for a more personalised medicine. More specifically, AI tools and methods can be used on *"the real-time follow-up of the patient and the logs he/her produces (monitoring of his or her physiological state, description of his or her symptoms, interactions with his environment,...)"*. It could be seen in this context that data are to be **harvested from various sources**, could be **heterogeneous in nature**, and are not available only within particular in-house information systems.

Whether at the individual or at the population level, the data that are of interest and that are processed, as can be seen, are heterogeneous in nature. Those from EHRs can be structured on the basis of a pre-identified data schema. However, a majority on them is **unstructured**. This is the case of free texts. In addition, the pedagogical practice of reading and evaluating the scientific quality of **medical literature** is inherent to the modern medical practice. Both the free text data from EHRs and those from scientific literature need to be handled properly and in an automated way, thanks to the use of approaches from **Information Extraction (IE)** and **Natural Language Processing (NLP)**.

Similarly, there is a need of texts processing approaches for a type of data that has become predominant nowadays : those generated by the population's activity on the Web. Indeed, **social networks** (twitter, web forums, etc.) have become common, and constitute an alternative that generates interesting data. It is estimated that about 43% of the world's population uses social networks, although there is a disparity between so-called resource-limited countries (such as Africa and some South East Asian countries) where there are still barriers, and developed countries. On social networks, users express themselves more freely in familiar, poorly supported or even degraded language. This poses real challenges when it comes to processing and integrating them with more traditional data to answer research questions in the healthcare domain.

In **resource limited settings countries** specifically, enabling healthcare related policies or public health decision making is usually limited by the lack of accurate data. Fortunately, new opportunities are emerging with the very high penetration rate of mobile technologies in these countries, particularly in Africa. Africa has become the world's fastest-growing mobile phone market according to the International Telecommunication Union. As of consequence, a kind of emerging data, which are not health data per-se are **Call Detail Records (CDRs)**. They are collected passively by mobile network operators in the course of providing their daily service (Jones et al., 2018). This type of voluminous and passively collected data are increasingly being used in research along with other forms of big data. They have proven their potential for instance with an unprecedented level of detail, the effect that mass gatherings can have on the spreading of some epidemics (Finger et al., 2016).

1. http ://www.enseignementsup-recherche.gouv.fr/cid128577/rapport-de-cedric-villani-donner-un-sens-a-l-intelligence-artificielle-ia.html

Finally, with the emerging and development of **Semantic Web (SW) technologies** (T. Berners-Lee et al., 2001) and **Linked Data** initiative and principles (S. T. Berners-Lee, 2011), another kind of data has been popularised. These are rather metadata that are available in the form of **Knowledge Organisation Systems (KOS)** (Mazzocchi, 2018) that could include terminologies, structured vocabularies or formal ontologies. They provide the shared understanding and knowledge to encode the data. They provide not only the knowledge to encode data within hospital information systems but also the model for many of the data sources and knowledge bases that could be available online on the Web. They could be used to ensure the **(semantic) interoperability** of the data they encode by the means of identifying and establishing a set of correspondences, so-called alignment between them.

These different data are key drivers of enabling a **digital health**. My research work attempts to leverage and integrate, in a **multidisciplinary context**, these type of data and knowledge at **large scale** for tackling issues in the healthcare domain. The designed and implemented methods are based on the elicitation (through knowledge formalisation) and exploitation of semantics encoded in KOS. More specifically, I address the challenges posed in this context through the following research topics, some identified as sub-fields of AI : i) **Knowledge Engineering in the Healthcare domain** and Large Scale Ontology Matching ; ii) **Large Scale Biomedical TextMining and Semantic Information Retrieval** ; and iii) **ICT for Development as digital Public Health enabler**.

This work has been carried out within the framework of several international, national and local multidisciplinary projects with numerous collaborations on applications in the field of life sciences and public health : pharmacovigilance and drugs monitoring, infectious diseases and Alzheimer disease and related syndromes, herbal-based medicine, etc. One defended PhD thesis has contributed to this work together with two others in progress. Similarly, several research internships at Master's or final engineering degree and research engineers have participated in this work. And a PostDoc has just started under an Horizon 2020 Marie Curie MCSA framework.

1.1.1 Structure of the manuscript

The rest of the manuscript will be structured as follows. First of all in the first chapter I give an overview of the work I have done and the context of the projects in which they are included with junior colleagues who contributed either as PhD students, Master (MSc) internships or engineers. Then, in order to keep an overall coherence of the work reported, I detail in the next 3 chapters the work according to the following structure.

— the second chapter focuses on the contributions related to large scale biomedical textmining and semantic information navigation and retrieval. I will detail the work conducted on large scale biomedical literature mining (several million of documents) for the purpose of **automated semantic indexing when only partial information is available** ; ii) then I will focus on the work dedicated to **medical web forums content mining** with the purpose of drugs misuse identification in the context of post-marketing drug monitoring ; an overview of **semantic browsing** of infectious diseases related resources will be the last part of the chapter ;

— the third chapter is dedicated to the contributions on knowledge integration with the K-Ware knowledge warehouse framework and large scale **(biomedical) ontology matching** on different sides : i) large scale biomedical ontology matching, ii) users involvement in the matching process and iii) multilingualism issue in ontology

matching ; the issue on handling heterogeneous KOS will be described thereafter with the K-Ware proposal for jointly handling different KOS and their alignment.
— the fourth chapter addresses the issue related to the exploitation of ICT and semantic technologies in order to contribute to overcoming some public health related issues in countries with limited settings (low and middle income countries). A first focus will be made on the integration and analysis of large CDR data to enable early public health decision making, while the second focus will be on the design and development of **knowledge graph** (Paulheim, 2016) about medicinal plant (WATRIMed).
— in the last part of the description of the research work, I finish by drawing the conclusion and outlining my research statement as perspective to this work.

An extended curriculum vitae completes the document in the second part this manuscript. As an overview, an extract of some noticeable events and scientific involvements throughout my career since the PhD thesis defence in December 2016 is depicted in Figure 1.1.

1.2 Brief Reminder about my Master and PhD thesis work

1.2.1 Overall Objective

Information systems are very often designed to operate independently and not for integration purposes. The objective of my early research work (MSc and PhD Thesis) was overall to explore methods and techniques for bringing together information retrieval (IR) and databases technologies in order to semantically access to information disseminated in several sources. Indeed, having specific information at one's disposal for good decision-making may require access to and cross-checking of information from several sources of information. The work of combining information from the data sources is then done manually. In my work, I was interested in proposing methods and techniques for semantic access to information based on ontologies and technologies developed within the framework of the Semantic Web.

1.2.2 Thesis Research Work

My thesis work focused on the proposal of a unified approach to the management of structured and textual data based on the use of Semantic Web technologies and different types of ontologies. This work led to the proposal of an ontology-based integration architecture (Diallo, 2006 ; Diallo, Simonet et al., 2006b).

In this architecture, ontologies are used both for the representation of the global schema and source schemas (Calì et al., 2002) as well as for the semantic characterisation (indexing) of the content of sources through *speciality* terminologies and structured vocabularies. To do so, I proposed and implemented a hybrid representation approach for textual sources (Diallo, Simonet et al., 2006a) that provides a pivotal representation of these sources. As semantic characterisation may require the use of several external sources of knowledge, which provide different points of view, I was interested in formalising a joint and dynamic management approach for several ontologies that are described in the Ontology Web Language (OWL)(McGuinness, Van Harmelen et al., 2004) and Resource

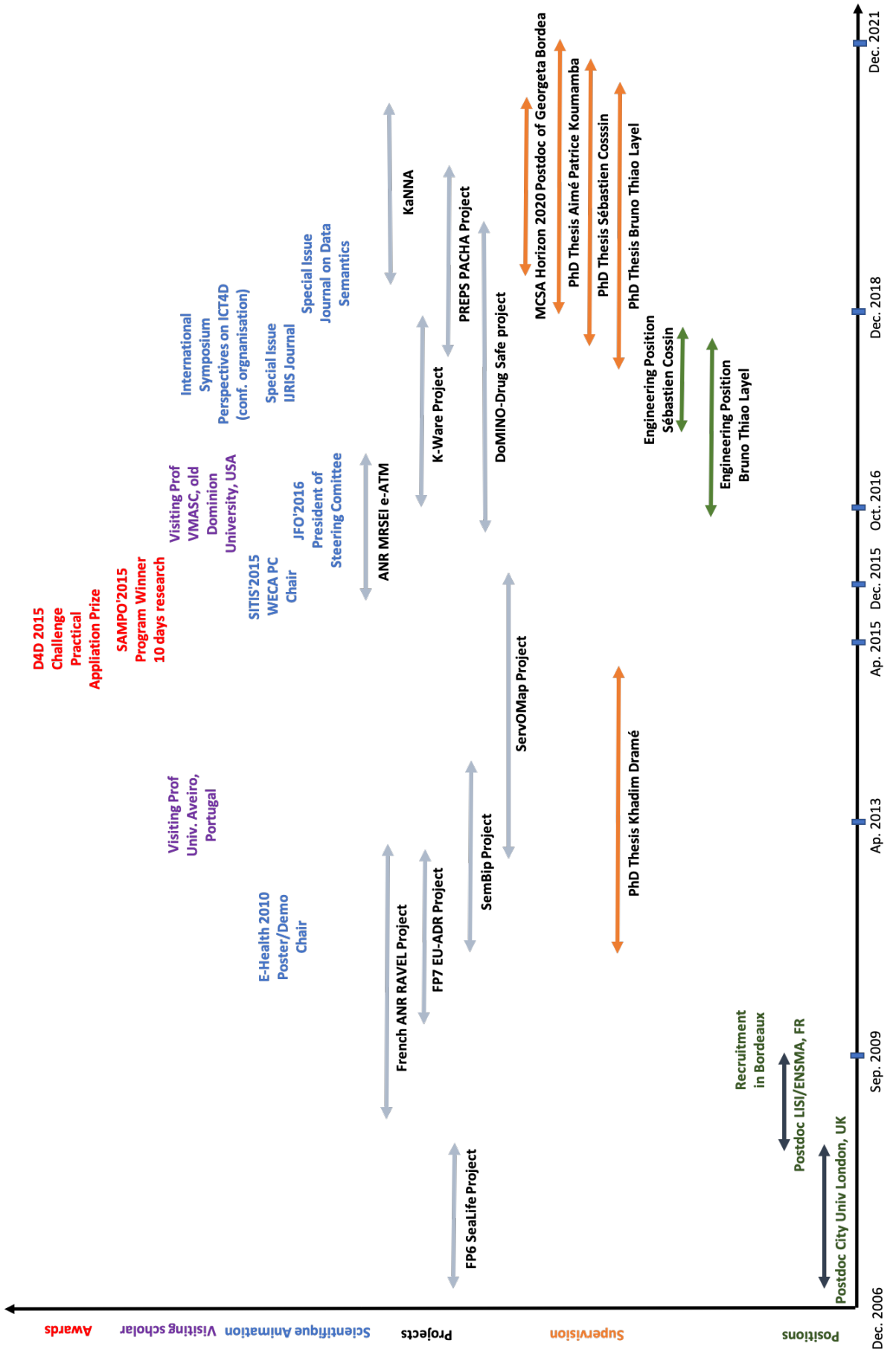

FIGURE 1.1 – Timeline about some main events and professional activities since my PhD thesis defence

Description Framework (RDF)(Lassila et al., 1998) languages, standards of the World Wide Web Consortium (W3C) for ontologies and metadata description. The approach is based on vector representation model from the Information Retrieval (IR) domain, for indexing and searching ontologies.

The application context of this work was the French Rhône Alpes regional project Bases de Connaissances Coeur Cerveau (BC^3). It aimed at extracting and representing knowledge about two vital organs, brain and heart, from the anatomical and functional perspectives. The project involved several teams from the Rhône-Alpes region [2] and the Neuropsychology Centre of the Pitié Salpêtrière Hospital (Paris).

One of the main challenges for the specific case of the brain was to understand the mechanisms in humans in order to improve the management of brain damaged patients. This understanding means, in particular, building a knowledge base (KB) that makes available anatomical and functional knowledge about the brain, as well as the links between anatomical areas of the brain and cognitive functions. This KB was based on the anatomical and functional brain Ontology I developed.

It was at this time that I began a multidisciplinary context research work, with a privileged application domain in healthcare.

1.3 Postdoctoral Research Work

1.3.1 Semantic Information Retrieval of Infectious Diseases Resources

As a follow-up to my thesis work on homogeneous access to heterogeneous information sources, I was interested in my work as a postdoctoral researcher at City University (C.U.) in London in a domain specific access and navigation at Web-scale information and services, in our case the field of infectious diseases. This was part of the FP6 European SeaLife project (Schroeder et al., 2006)). The problem of IE and therefore the identification and extraction of relevant terms from the domain on a web page being visited by a particular user is an important issue in this context. The proposed approach was the development and use of a structured vocabulary, built with experts in the field (Diallo, Kostkova et al., 2008) to support text mining and Web pages content annotation as well as the description of the different users for personalising information (clinical practitioners, microbiologists, clinicians, etc.) (Kostkova et al., 2008). This model is based on the Conceptual Open Hypermedia Service (COHSE) system (Goble et al., 2001), a precursor to current semantic browsers, which is one of the systems on which the SeaLife engine is based, for semantic navigation of the NeLI portal (portal on infectious diseases information and selected scientific publications in UK) developed and maintained by the City eHealth Research Centre group at C.U. where I was working. During my postdoctoral stay at C.U., I also coordinated the design of the user-centered SeaLife semantic browsers evaluation framework as part of the Work Package of the project dedicated to the evaluation (Oliver et al., 2009).

A description of the work on the annotation of information and the semantic information retrieval and browsing will be the subject of section 2.4 of chapter 2.

2. They include LAAS-CNRS laboratory of Univ. Lyon 1 (coord.), TIMC-IMAG Grenoble

1.3.2 Personalising Information in Ontology-based Databases

Subsequently, and following my City U. Postdoc, my postdoctoral research at the Laboratory of Applied Computer Science (LISI/ENSMA) - Futuroscope Poitiers (11 months) was an opportunity to work more specifically on users profiling (Zayani et al., 2017) in the context of databases. Indeed, I took an interest in Ontology Based Databases (ODBD) with the OntoDB framework which was being developed at LISI (Dehainsala et al., 2007). I contributed in the design of a more generic and persistent model for personalising information through user profiling and published in (Tapucu, Diallo, Ait-Aimeur et al., 2009 ; Tapucu, Diallo, Jean et al., 2009).

Preferences in the databases domain context used to be defined at logical model, particularly on the values of the columns of the different tables. Depending on the type of metrics, two ways of expressing preferences have been proposed : i) the qualitative approach which allows the user to define (relative) preferences between the tuples of the database (Chomicki, 2003 ; Kiessling, 2002) ; the quantitative approach which makes it possible to define scoring functions in order to calculate a numerical score, also known as absolute preference, for each tuple (Agrawal et Wimmers, 2000).

Our designed model defines user preferences at the semantic level (Tapucu, Diallo, Ait-Aimeur et al., 2009) and is based on OntoDB. ODBDs have the advantage of storing in the same infrastructure both data and ontologies that clarify their semantics. To this end, The OntoDB model has been extended with the preference model together with its associated OntoQL language (Jean et al., 2015) with a specific *PREFERRING* clause after the *SELECT* clause of OntoQL. Two types of preferences were distinguished : i) the interpreted preferences which are preferences associated with an evaluation procedure or of interpretation ; ii) uninterpreted preferences which correspond to a set of instances of classes or properties from an Ontology that are considered preferred.

The customisation preference model that has been proposed and implemented is independent of any logic data model. The ability to link it to any Ontology model makes it completely flexible. It offers the advantage of storing the defined preference model by offering the possibility of its reuse. In that sense, this work were more generic than the one we conducted in the previous context of the SeaLife browser.

1.4 Capitalising on these Previous Experiences

Since I joined in Septembre 2009 the INSERM BioStatistics and Epidemiology research center, now known as Bordeaux Population Health (BPH) INSERM 1219 research center, the main topic of my research work, in the frame of the work of the ERIAS research group, focuses on how semantics can leverage handling at large scale heterogeneous knowledge and data in the healthcare domain with a main purpose to facilitate their secondary use (epidemiological studies, clinical research, statistics analysis, etc.).

1.4.1 Knowledge Engineering in the Healthcare Domain

The need for a common understanding of the entities in a field of interest has led to the widespread adoption of ontologies as a means of representing knowledge (Gruber, 1995 ; Hoehndorf et al., 2015). This is particularly the case in the biomedical and healthcare domains where the desire to categorise things is a tradition that goes back many centuries. Further, there is an increasing consensus that medical knowledge representation (KR)

should use shareable standards for enabling computational support of data management. The converging of tools and methods is opposed to the richness of domains, concepts, and especially domain terms in a multitude of languages. A broad account for health knowledge representation should therefore be able to formalise knowledge using the following KR assets as basic building blocks :

1. Concepts (aka types, repeatables), i.e. language-independent entities of meaning that normally extend to classes of individual things ;
2. Individuals, i.e. tangible, non-repeatable entities ;
3. Terminologies, i.e. units of human language, denoting entities from a) or b).

Glued together by standardised languages (e.g. RDF, SKOS (Miles et al., 2005), OWL (McGuinness, Van Harmelen et al., 2004)) and principles (e.g., linked data(S. T. Berners-Lee, 2011), Open Biological and Biomedical Ontology (OBO) Foundry [3]), the resulting propositions ideally result in shareable, interoperable, and computable KR assets. And that is the purpose of an Ontology, defined as a formal, explicit specification of a shared conceptualisation (Guarino, 1998).

The work pursued on this topic has been carried out successively i) in the context of the FP6 SeaLife European project (the NeLI Ontology) ; ii) in the Semantic BioblioDem Portal project, as part of the PhD thesis of Khadim Dramé (KD) (Nov. 2011 - December 2014), which I co-supervised with my colleague Fleur Mougin ; iii) the French ANR MR-SEI e-ATM and the West African Herbal Based Traditional Medicine Knowledge Graph (WATRIMed) project that I coordinate and which involves the ERIAS research group, the Medical University of Graz in Austria (Pr Stefan Schulz), Nazi Boni University of Bobo Dioulasso in Burkina Faso (Dr Michel Some) and the collaboration with the West African Health Organisation (WAHO, Dr Koffi Busia).

Knowledge Organisation Systems Engineering in the Healthcare Domain

Building a KOS, in particular Ontology, is not a trivial task. In contrast to the database domain, there is no commonly agreed methodology. The process, which usually involves domain experts, must lead to a clear, coherent, easy to use and extensible artefact (Guarino, 1998). Ontology, terminology and other KOS building projects have been undertaken for many decades (see (Gruber, 1995 ; Uschold et Gruninger, 2004) for example). Some methodologies start from scratch while others proceed by learning from texts or reuse existing vocabularies or ontologies. Some (semi-)automated building methodology from texts emerged (for instance (Biebow et Szulman, 1999 ; Dramé, Diallo et al., 2014)), as the majority of any organisation's knowledge is encoded within the textual documents. The general idea is to identify the most important concepts within the texts as well as hierarchical and transversal relationships using NLP and IE tools, sometimes using external knowledge resources.

In the biomedical domain, the Unified Medical Language System (UMLS) (Bodenreider, 2004) which gather more than 160 KOS in the field is widely reused. As of illustration, Hahn and Schulz reused knowledge contained in the UMLS to develop a formal model with a more detailed level of conceptualisation (Hahn et Schulz, 2004). While UMLS and the MeSH (Medical Subject Headings) thesaurus are the input resources for building an ontology of cardiovascular diseases in (Gedzelman et al., 2005).

3. http ://www.obofoundry.org/

Identifying and establishing correspondences (mappings) between the domain ontologies with **foundational (upper level) ontologies** (Arp, Smith et Spear, 2015), like the Basic Foundational Ontology (BFO) (Arp et Smith, 2008) is of importance. Foundational ontologies, which consists in very general categories that are common for a large set of domains, can enable the interoperability between different domain ontologies. It assists for instance in the organisation and integration of biomedical information and guarantees its interoperability with other biomedical ontologies. I will briefly report the use of an upper level Ontology in the context of our WATRIMed ontological model (Some et al., 2019).

Once designed, the ontology or KOS has to be implemented in a specific language. Editor tools constitute help for this task when all choices have already been made. Protégé [4] is one of the tool widely used for Ontology building. The RDF/RDFS (Lassila et al., 1998) and the Web Ontology Language (OWL) (McGuinness, Van Harmelen et al., 2004) are currently the most representative. Similarly, the more recent Simple Knowledge Organisation System (SKOS) (Miles et al., 2005) which is based on RDFS plays nowadays a key role in semantic description for lightweight KOS (terminologies, vocabularies and other concepts schemes). Such KOS lack the formal, ontological distinctions made in many ontologies used in their pursuit of describing the nature of entities in the world and information about the world. KOS, on the other hand, often simply describe looser conceptual descriptions of the "relatedness" of terms within a domain or how a particular domain is organised by its members, whether or not such an organisation has much bearing on reality. Such KOS are, however, extremely valuable in information retrieval, indexing, information navigation, etc. This is why it has been used for the NeLI Ontology described afterwards.

In the following sections I report the results achieved in this area in three application domains.

The National electronic Library of Infection Ontology. As briefly presented previously, the National electronic Library of Infection (NeLI) provides a single access point to the best available evidence on all aspects of infection [5]. The family of NeLI projects include the National Resource for Infection Control (NRIC) [6], Bugs and Drugs on the Web [7] and other infection-related projects. Funded till recently by the UK's Department of Health and the Health Protection Agency, NeLI provides the main source of online medical evidence for a wide spectrum of users - clinicians, family doctors (GPs), environmental health officers, infection control nurses, and other public health professionals.

The objective of the NeLI Ontology (presented in (Diallo, Kostkova et al., 2008)) that we developed in the context of the FP7 SeaLife project (Schroeder et al., 2006) during the period 2007-2008 is to cover the wide spectrum of infection, including clinical microbiology, clinical infectious disease, infection control, public health, surveillance and populations affected. The infection domain includes the investigation, diagnosis, treatment and control of infectious diseases at a clinical and public health level. From a functional requirement point of view, there is no need to provide clinical diagnostic and decision making support as the library consists of the best available evidence interpreted by doctors rather than providing a clinical decision support system or linking to patient electronic

4. http ://protege.stanford.edu
5. www.neli.org.uk
6. www.nric.org.uk
7. www.antibioticresistance.org.uk

records. Therefore, formal reasoning for this particular need was not our primary goal. Instead, the Ontology is used for navigation, search and quick access to customised resources for groups of users. The following objectives were identified :

- **Semantic annotation of infection resources :** in order to support a user-customisable search ;
- **Personalising of the NeLI portal :** a "one size fit for all" approach for searching and ranking discovered knowledge on the Internet does not cater for the diverse variety of users and user groups with different preferences, information needs and priorities ;
- **Browsing of Infection domain related resources.** the browsing issue is motivated by the development the SeaLife project.

From implementation point of view, NeLI reuses the Medical Subject Headings (MeSH) thesaurus (Lowe et Barnett, 1994) and combines the OWL language with SKOS which offers builtin predicates for preferred (*skos :prefLabel*) and alternative (*skos :altLabel*) naming of entities. It has been used as background knowledge for two of the Semantic Web Browsers of the SeaLife project and its usefulness in accessing relevant information evaluated (Oliver et al., 2009). I will briefly describe in chapter 2 an illustration of semantic information retrieval and navigation with the Corese-NeLI semantic web browser which uses the NeLI Ontology as a background knowledge.

The Alzheimer Disease and related Illness Ontology. This work was as part of the **PhD thesis of Khadim Dramé** and the Semantic BioblioDem Portal project (SemBip), funded by the Fondation de France/Plan Alzheimer (2011-2013). The SemBiP project, that I coordinated with my colleague Fleur Mougin, is designed to offer a semantic access to scientific resources on Alzheimer's disease (AD) and related disorders. In particular leveraging the French BiblioDem newsletter and cumulative scientific publications database [8]. Its objectives were :

- providing a multilingual (mainly English and French) knowledge resource dedicated to the description of Alzheimer's disease and related illnesses supporting both content indexing and educational purpose. The evolution of this vocabulary has to be managed ;
- providing a semantic portal enabling semantic search and browsing of approved available evidence about Alzheimer's disease and related disorders ;
- taking into account the specificity and variety of users. According to the user profile, adapted search functionalities should be offered ;
- allowing multilingual access to both the original abstract (and the full paper if freely available) and their critical analysis in order to improve decision making. The multilingual aspects are very important for supporting both indexing and cross-lingual information retrieval in the portal (original articles abstract in English and comments in French).

We were therefore interested in providing semantic access to AD related resources to French clinicians and researchers for the purpose of clinical epidemiology with an ultimate goal to facilitate an evidence based practice. In France, the number of people suffering from AD is estimated at 860,000. Every year, 220,000 new cases are identified in this country. Nearly 350,000 people benefit from care for the long-term complaint of Alzheimer's-type illness and related disorders.

8. http ://sites.isped.u-bordeaux2.fr/bibliodem/

The solution was to design and develop an AD domain Ontology from existing resources and using it for indexing AD related resources and information retrieval in the French and English multilingual context. The work on the knowledge modelling and representation has led to **the proposal of a mixed approach** combining **NLP** techniques and the **reuse of existing semantic resources** (the UMLS) for the construction of a KOS in the form of a termino-ontological resource (TOR) (Dramé, Diallo et al., 2014). On the one hand, NLP offers robust tools to extract ontological knowledge from texts corpora with good accuracy. On the other hand, the reuse of existing TORs reduces the process of building ontologies. These two approaches are therefore complementary and the idea was to combine their advantages. In order to address the issue of multilingualism (original scientific articles mainly in English and their critical analysis available in French), we relied on methods from **parallel corpora alignment** that are widely used particularly for the acquisition and the enrichment of multilingual terminologies. Therefore, a **language alignment method** based on parallel corpora have been designed and implemented to map terms from different languages (Drame et al., 2012).

This methodology was applied to a case study on the critical reading of articles in the field of AD with the collaboration of the Epidemiology and Neuropsychology of Cerebral Ageing team of the INSERM U897 Epidemiology and Biostatistics Centre in Bordeaux. It made it possible to implement a bilingual Ontology of AD, OntoAD (Dramé, Diallo et al., 2014), used to support the SemBip semantic portal, for managing resources on the AD domain. OntoAD includes 5,899 concepts linked by 7,499 taxonomic relationships and 10,889 non-taxonomic relationships and is freely available on the online BioPortal repository [9]. BioPortal is a repository which gathers biomedical KOS (Noy et al., 2009).

WATRIMED : West African Herbal-Based Traditional Medicine Knowledge Graph. Medical KR formalisms are especially challenged in domains that occupy a rather marginal position in the KR ecosystem. A typical example is herbal based traditional medicine (TM), which has a significant foothold in large areas of the planet, and which is intertwined with the most diverse cultural heritage of population groups, with large influence on the health status of a population. E.g., the Worl Health Organisation (WHO) estimates that as much as 80% of the population uses TM in some form, and herbal-based Traditional Medicine (HTM) in particular.

From the experience on knowledge engineering I acquired, in collaboration with B. Kamsu Foguem (LGP ENIT Tarbes, France) and as part of my interest on bringing ICT for societal development, in particular for the healthcare domain, we had carried out a first conceptual work on the field of TM which used Conceptual Graphs (Sowa, 2008) to model and reason about African Traditional Medicine (ATM) (Kamsu-Foguem et al., 2013). Thereafter, following an attempt to set an Horizon 2020 proposal in 2017, which brought together partners from 4 European countries and 5 African countries that I coordinated and which was dedicated to bringing Semantic Web and mobile technologies to leverage ATM practices and contribute to increasing safety usage, we have initiate the design and development of the West African Herbal-based Traditional Medicine (WATRIMed) Knowledge Graph (KG) (Some et al., 2019). The work relies on the reuse of materials provided by the supra-national West African Health Organisation (WAHO) and other freely available knowledge bases within the Linked Open Data cloud or online on the Web. Similarly to previous attempts of digitising Chinese TM (Li et al., 2018) and more general African TM (Lô et al., 2017), using a state-of-the art, flexible and shareable knowledge

9. https ://bioportal.bioontology.org/ontologies/ONTOAD

representation approach, this effort aims at bringing West African TM to the digital world so as to help preserving the TM secular knowledge and contribute to the safely usage of TM thanks to enabling among others phytovigilance (pharmacovigilance about plants) activity for safety purpose. But also to provide an elaborated formal modelling about medicinal plants. As far as we know, there is no formal model available on the topic. However, an intensive effort has been conducted in the context of Traditional Chinese Medicine (TCM)(Li et al., 2018 ; Z. Wu et al., 2008) which leads to the integration of TCM to the upcoming eleven revision of International Classification of Disease (ICD) [10] of WHO.

For WATRIMed, a core model, which has been fed with data from the WAHO pharmacopoeia and linked to publicly available external knowledge bases about herbs information, recipes and chemical ingredients has been provided (e.g., DBpedia, PubChem, etc.). Further, the core model has been mapped to an upper level ontology, BioTopLite2 (BTL2), a light version of the BioTop Upper Level Ontology for the Life Sciences (Schulz et Boeker, 2013). As of result, the graph contains currently about 30,000 facts modelled with an Ontology of 1,129 concepts, 75 properties. WATRIMed is freely available online to the community, with a SPARQL endpoint [11]. It demonstrates that it is possible to bring advanced technologies in the context of knowledge modelling in countries with very limited settings.

I have mentioned here some semantic resources in specific project contexts. This type of approach has become common and results in multiple KOS, sometimes for related areas. However, in many cases, it is important to be able to integrate them and therefore identify first possible correspondences (called alignment) between their entities.

1.4.2 Semantic Knowledge Integration and Large-Scale Ontology Alignment

It is commonly accepted that efficient decision-making, particularly in the life sciences and healthcare domain, is based on the access to increasingly voluminous data, stored in many various sources (Gini, Ryan et al., 2013 ; Gini, Schuemie et al., 2016). These sources are often heterogeneous in many respects including their storage formats, coding vocabularies (metadata associated with data), languages and formalisms used. The emergence of the Semantic Web and the Web of Data has offered new possibilities for identifying and establishing (automatically) mappings between these heterogeneous data (Shvaiko et Euzenat, 2013). The offered solutions are based on identifying and accessing the KOS on which the data sources are based. These KOS are of various natures : domain ontologies for describing and querying data sources, speciality terminologies for indexing and annotation (semantic characterisation) of data sources. Several challenges are posed in this context :

— How to handle and manage jointly this set of KOS with different levels of formalism ?
— How to identify links between their entities and qualifying them semantically ?
— How to scale up in the presence of increasingly large KOS ? Indeed, in the biomedical and healthcare domains in particular, there are increasing resources with several hundred thousand entities (e.g, the SNOMED CT for example, a resource used in the healthcare domain has over 311,000 entities (Lee et al., 2014)).

10. https ://icd.who.int/en/
11. www.watrimed.org

The links between KOS entities can be of various kinds including equivalence, generality and/or specificity (Shvaiko et Euzenat, 2013). In this respects, an Ontology mapping or more generally a KOS mapping (referred also as match or correspondence) between entities of two ontologies O_1, O_2 is represented as a 4-tuple $< e_1,\ e_2,\ r,\ c >$ where $e_1,\ e_2$ are entities from $O_1,\ O_2$ respectively ; $r \in \sqsubseteq, \sqsupseteq, \equiv$ is a semantic relation ; and c is a confidence value usually within the interval [0...1].

The work I have explored in this context started in 2010-2011 and is part of the general problem of interoperability and Ontology integration(Pinto et J. P. Martins, 2001) and jointly handling multiple KOS for an easy management of their content in order to facilitate their reuse and understand their different interlinks. It is a natural continuation of the work I started in my thesis as part of the joint and dynamic management of several KOS, which I had set aside for a while according to research opportunities offered by my successive post-docs. The FP7 EU-ADR (Early Detection of Adverse Drug Reaction) project (2009-2012)(Avillach, Dufour et al., 2012 ; Avillach, Mougin et al., 2009 ; Oliveira et al., 2013), my research group was participating in, gave me a good opportunity to figure out how present the heterogeneity issue of KOS is in the context of public health when it comes to conducting multi-centre studies.

The EU-ADR project as a multiple KOS alignment use case.

The EU-ADR project worked to develop an innovative computerised system to detect adverse drug reactions (ADRs), supplementing spontaneous reporting systems which are considered limited. The project exploited clinical data from electronic healthcare records (EHRs) of over 30 million patients from several European countries (The Netherlands, Denmark, United Kingdom, and Italy). In this project a variety of textmining, epidemiological and other computational techniques were used to analyse the EHRs in order to detect *signals* (combinations of drugs and suspected adverse events that warrant further investigation). In EU-ADR, we exploited data from eight European healthcare databases and Electronic Health Records for the detection of these signals. As these databases use different KOS, specifically medical terminologies (the WHO ICD-9 and ICD-10 respectively for Italy and Denmark), Read Codes for United Kingdom, International Classification of Primary Care (ICPC for The Netherlands)) to encode their data according to the database's country, mappings were needed to translate a query posed to the global system into queries understandable for the different actual data sources. Performing manual mappings between all the mentioned resources is hardly feasible within a reasonable time. Hence the need for methods and tools for the (semi)automated alignment of these KOS.

This case of EU-ADR is not isolated in the broad area of healthcare, in particular public health, we can also mention the TECSAN RAVEL French ANR funded project (2012-2014) (Diallo, Grabar et al., 2012) where we faced the issue on handling complex queries over complex patients (patients with a long follow-up) from EHR. The observation is that the need of aligning various KOS occurs quite often in various research field (semantic data integration domain (Calvanese et al., 2018), the semantic browsing of information (Kostkova et al., 2008) and web services composition(Claro et al., 2006 ; Hao et Y. Zhang, 2007)).

An extensive work has been conducted in the KOS alignment domain (please see (Kalfoglou et Schorlemmer, 2003 ; Otero-Cerdeira et al., 2015 ; Shvaiko et Euzenat, 2013), for an extensive review). However, until 2011 no or little attention has been paid to the alignment of large KOS, as it could be often the need in the biomedical and healthcare domain. And this is my focus. Aligning large KOS, which involves several

thousand hundreds entities presents several challenges. Let recall in the following section some important one.

Challenges on large scale KOS alignment

- **Reducing the search complexity :** traditional systems and algorithms that perform alignment tasks on small size ontologies used to compute nXm similarities between respectively the n and m entities of the two input ontologies. This makes them prohibitively inefficient for scaling up. Various techniques and strategies have been adopted to tackle this issue. The most common are : i) the divide and conquer strategy relying on clustering and modularisation (Shvaiko et Euzenat, 2013) ; ii) information retrieval based strategy which relies on indexing techniques (Diallo, 2014). The chapter3 will detail the ServOMap approach that we have proposed and which follows the latter strategy.
- **Evaluating and validating the alignment results :** while manual validation is possible for relatively small size ontologies alignment tasks, it is not reasonably feasible for large ontologies containing several hundred of thousand entities (like in the biomedical domain).
- **User involvement and human interaction :** it has been proven that there is a limit, in terms of alignment quality, that is difficult to overcome by solely fully automated Ontology alignment systems (Paulheim et al., 2013). The need to incorporate the user into the alignment process emerged a bit earlier(Falconer et Noy, 2011). The user can review the proposed correspondence candidates and validate those that are correct and to propose additional correspondences that the system would not have found.
- **Reducing time complexity :** if we assume that n is the size of each input ontology and t the time needed for the matching process of a system, the matching process in a Cartesian product fashion has a time complexity of $O(n^2Xt)$. This is considerable in the case of large ontologies.

Given the number of alignment systems that have emerged, it was necessary to find a framework to assess their effectiveness.

The Ontology Alignment Evaluation Initiative

The Ontology Alignment Evaluation Initiative (OAEI) [12] is an international coordinated initiative launched in 2004 and aiming at helping to improve the work on Ontology alignment. Its goals are :

- assessing strengths and weaknesses of alignment systems ;
- comparing performance of techniques ;
- increasing communication among algorithm developers ;
- improving evaluation techniques ;

The means to achieve these goals are the organisation of a yearly evaluation event since 2004 and the publication of the tests and results of the event for further analysis. OAEI provides datasets and a framework in order to perform a controlled experimental evaluation of the different alignment systems. Several tracks and challenges, subdivided in two categories, have been introduced over the years. As an illustration, the track *LargeBio* has been introduced in 2011 and dedicated to computing alignment between large healthcare

12. http ://oaei.ontologymatching.org/

related ontologies (SNOMED CT (Lee et al., 2014), the National Cancer Institute thesaurus (NCI) (Golbeck et al., 2003), and the Foundational Model of Anatomy (FMA) (Rosse et Mejino, 2008)). As noted by the organisers, these ontologies are semantically rich and contain tens of thousands of classes. The UMLS Metathesaurus(Bodenreider, 2004) has been selected as the basis for the track reference alignments. Other tracks include *Multifarm* (Meilicke et al., 2012) dedicated to multilingual ontologies alignment and the *Interactive* track (Dragisic et al., 2016) which considers the user involvement during the alignment process. Similarly, for the alignment at instance level and link discovery, several datasets have been introduced, like IIMB which is an OWL-based dataset that is automatically generated by introducing a set of controlled transformations in an initial OWL assertion component (Abox). With the recent growing popularity of knowledge graphs (Paulheim, 2016), a new track dedicated to both aligning schemata and instances is envisioned for the next edition of OAEI which will be held in October 2019 in the context on the International Workshop on Ontology Matching in Auckland (New Zealand) in conjunction with the International Semantic Web Conference [13].

Large Scale (Biomedical) KOS Alignment with ServOMap

I summarise in the following section, my main contributions on knowledge integration and large scale Ontology alignment domain, particularly for biomedical ontologies. Several MSc students from 2012 and more recently a collaboration with colleagues from University Bobo Dioulasso and University Tunis El Manar have contributed to this work in the context of the ServOMap project which resulted in several published articles including (Ba et Diallo, 2013 ; Diallo, 2014). The K-Ware knowledge warehouse framework, a natural contribution to this work, and which is introduced below and will be described in Section 3.2 is another contribution as part of the **PhD thesis of Bruno Thiao Layel** which has started in mid 2018 jointly with the Synapse Medicine start-up [14], a spin-off of the BPH - ERIAS research group, and that I am co-supervising.

The ServO Ontology Server. In order to properly handle heterogeneous KOS and to enable their alignment, the first realisation was the ServO Ontologies server proposal (Ba et Diallo, 2012 ; Diallo, 2011). It offers a dynamic way to index and query different heterogeneous KOS within the same infrastructure that allows them to be integrated in order to perform a number of meta-operations. These meta-operations include computing alignment between KOS or ; following the Swoogle (Finin et al., 2004) and the Watson system (d'Aquin et Motta, 2011), performing the retrieval or selection of specific KOS or their entities that are appropriate for a specific task (e.g. semantic indexing). ServO is based on a generic metamodel capable of abstracting itself from the heterogeneity due to the different representation languages used by the KOS themselves. Later on, in the context of the MSc internship of Zakaria Mambone, a fully decentralised version based on an algorithm which relies on a peer-to-peer infrastructure has been proposed (Mambone et al., 2015).

ServOMap : Scaling Up in Biomedical Ontology Alignment. In the area of large-scale Ontology alignment, the work I pursued relies on the exploitation of the ServO Ontology Server (Ba et Diallo, 2012 ; Diallo, 2011). In the context of increasing availability

13. http ://iswc2019.semanticweb.org/
14. http ://synapse-medicine.com/

of data and possibility associated metadata, represented as KOS, identifying matches between theses KOS and their underlying data is one of the key challenges to ensure semantic interoperability where needed.

The first avenue explored, as part of the MSc internship of Mouhamadou Ba, then Amal Kammoun, was the use of techniques from IR and algorithms developed in the context of ServO, considering each Ontology to be aligned as a **corpus of semantic documents**. The Ontology indexing approach implemented in this way has made it possible to meet the challenge of **reducing the search space** for alignment and to support large biomedical ontologies containing **several hundred of thousand entities** in a very short time (Ba et Diallo, 2013). Two similarity algorithms are implemented : a terminological approach that translates each entity to be aligned into a **virtual semantic document**, and a contextual approach that is based on Machine Learning (ML) strategy, and which takes into account the topological structure of the ontologies being aligned. The idea is to improve the Recall of the matching process by identifying contextual based candidate mappings while preserving the Precision (Diallo, 2014 ; Kammoun et Diallo, 2013).

Subsequently, the challenge was to take into account the user, who is an expert in the field, in the alignment process in order to overcome the limits in terms of performance of a fully automated approach. This is one of the key identified issue for improving the quality of the provided alignment and being able to align large Ontologies (Shvaiko et Euzenat, 2013). We have thus designed and implemented, as part of the MSc internship of Mahamadi Sawadogo and in collaboration with Michel Some (ESI, Univ Bobo Dioulasso), an approach in which the user can intervene in the mapping process either by suggesting initial matches, or by choosing similarity calculation metrics and finally by validating the candidate matches generated by the system (Savadogo et al., 2016). This has substantially improved the performance of our traditional approach in terms of F-score (reaching 20%).

The different algorithms have been implemented as part of the ServOMap system, whose development started in 2012, following that of the ServO Ontology server. It is a large-scale Ontology alignment system that can automatically process ontologies containing hundreds of thousands of entities in a short time. It has successively participated in the 2012, 2013 and 2015 editions of the OAEI challenge [15] and ranked among the top three systems for tasks dedicated to the alignment of large ontologies. The 2015 version of the system, ServOMBI, realised as part of the final year engineering degree of Nouha Kheder, introduces an algorithm inspired by (Ait-Kaci et Amir, 2017) for **bit-score binary indexing of the taxonomy** of each Ontology being processed in order to optimise the path and concepts look-up as part of the contextual similarity calculation. In addition, a candidate mappings selection and filtering strategy based on the principle of the **Stable Marriage Problem** (McVitie et Wilson, 1971) was proposed.

In the same topic on Ontology alignment, more recently and in collaboration with Marouen Kachroudi and Sadok Ben Yahia from University Tunis El Manar we did focused on cross-lingual Ontology alignment (aligning two Ontologies expressed in different natural languages) by bringing together ServOMap algorithms and those from the CLONA system (Kachroudi, Diallo et Ben Yahia, 2016). The resulting system, Kepler, use partition based techniques (Kachroudi, Zghal et al., 2013) coupled with input ontologies indexing strategy and a language translation engine to perform the alignment task between KOS in a multilingual context. The system has participated in the two last editions of OAEI with Top ranking in aligning same ontologies expressed in different natural

15. http ://oaei.ontologymatching.org/

languages (Kachroudi, Diallo et Ben Yahia, 2018 ; Kachroudi, Diallo et Yahia, 2017) with the MultiFarm dataset (Meilicke et al., 2012).

K-Ware : Handling Jointly heterogeneous KOS and their alignments. In the continuation of the work on large scale Ontology alignment and those on ServO in particular, and in the context of the **PhD thesis of Bruno Thiao Layel** (who was previously a research engineer and now one of my co-supervised doctoral students, together with Guillaume Blin (Bordeaux Laboratory in Computer Science - LaBRI) and Vianney Jouhet (ERIAS INSERM 1219 and CHU de Bordeaux) we are currently investigating in the design and the implementation of an approach for the transparent management of semantic metadata expressed by KOS (Ontologies and Terminologies) as well as their alignments and which is based on a generic metamodel (Thiao-Layel et al., 2016, 2018) with an application in the heterogeneous drug data management domain. The challenge is to be able to handle various heterogeneous KOS with different formalisms and languages (RDF, OWL and SKOS) and integrate them in the same infrastructure. The purpose is to preserve the original semantics of these resources while providing a mechanism for accessing the data they annotate. The advantage is to be able to provide a KOS manipulation and visualisation engine exclusively based on the provided metamodel with a preservation of the differentiation between formal ontologies described in Description Logics based languages(Baader et Nutt, 2003), such as OWL, and other structured vocabularies and terminologies. Indeed, for instance while it is possible to perform computational reasoning on the former, it is not suitable for the latter.

KOS are often used as external knowledge background for various NLP and IR tasks. They bring semantics that could contextualise unstructured information.

1.4.3 Contribution on Large Scale Biomedical TextMining and Semantic Information Retrieval

The growing availability of digital data both at individual and population level opens up new opportunities to carry out relatively more comprehensive studies in the field of health, and public health in particular (Krumholz, 2014). In this context, while healthcare professionals produce abundant textual information (e.g. discharge summaries) in their daily clinical practice, there are various other sources, available in unstructured textual format, valuable to be included in healthcare related studies. Thus, scientific literature, social networks data and web forums have become interesting data sources, which could complement the traditional data retrieved from traditional EHR.

Handling these unstructured data properly requires the use of NLP methods and tools and IR techniques. Information Extraction including Named Entity Recognition (NER), ontological entities mention, account among the major activities in the field. I report in the following sections the different approaches applied on documents harvested from **large set of different text genres** such as **biomedical scientific publications**, **health portals**, and **online communities**.

Biomedical Literature Mining at Large Scale

The research work pursued in the Biomedical Literature Mining domain has been performed as part of the PhD thesis of Khadim Dramé. More specifically, we focused on the large-scale automatic classification of free text documents. The challenge is to

be able to annotate biomedical scientific articles in an automated way by classifying a large collection of documents (more than 12 million) when only a partial information is available (e.g. for scientific literature, only the title and abstract). The idea is to use solely this partial available information to determine the relevant descriptors of the content of the complete documents. To do so, we proposed two approaches :

— a method based on the exploitation of the k nearest neighbours (kNN) algorithm ($kNN\_Classif$) (Dramé, Mougin et al., 2016). The idea is to rely on a dictionary which is available in the form of a KOS, as a descriptors provider whose entries are used to describe the documents being classified. We explored different techniques for calculating similarity between documents for nearest neighbours identification and relied on Machine Learning algorithms, such as Naives Bayes (Langley et al., 1992) and Random Forest (Kocev et al., 2007).

— a method based on Explicit Semantic Analysis (ESA) (Gabrilovich et Markovitch, 2009). ESA is an approach for representing textual documents in a semantic way. Documents are represented in a conceptual space constituted of explicit concepts automatically extracted from a given knowledge base. In our case, we used Wikipedia [16] as knowledge base and its entries serves as conceptual space into which documents are projected.

The identification of entities mention within texts relies on an implemented NLP algorithm which retrieve the most specific concepts (from the dictionary being used) into the textual corpora (Dramé, Diallo et al., 2014).

These approaches have been evaluated on a large dataset of several million of documents provided in the context of the international large scale biomedical semantic indexing and question answering (BioASQ) challenge [17]. BioASQ provides a framework and a large collection of annotated documents from the online PubMed database [18]. The documents are annotated using the descriptors from the MeSH thesaurus (Lowe et Barnett, 1994). In contrast to the ESA approach, $kNN\_Classif$ proved to be efficient to identify suitable descriptors when only a partial information is available, when it is trained on a large set of documents with results comparable to similar approaches.

Healthcare Related Web Forums and Social Networks Mining for Drug Misuse Detection

Drugs misuse can be defined in several ways. First, the failure to match the main purpose of the drug's use with the reason for the individual's use corresponds to a misuse of the drug and involves risks (for example, taking doliprane to sleep). In this misuse pattern, drugs are often used as narcotic substances, which can lead to serious health problems, including addiction. Then, misuse can be the overuse or underuse of a drug in relation to the recommended doses per hour or per day, the duration of non-compliance with a treatment, or the taking and combination of several drugs reported as harmful if taken simultaneously (the case of drug-drug interaction).

Drugs misuse is a source of iatrogenic pathologies and a significant additional cost to the health budget. The French High Committee on Public Health [19] considers as iatrogenic "the undesirable or negative consequences on the state of individual or collective health of any act or measure performed or prescribed by a qualified professional and which aims

16. http ://wikipedia.org
17. http ://bioasq.org/participate/challenges_year_2
18. https ://www.ncbi.nlm.nih.gov/pubmed/
19. https ://www.hcsp.fr/

to preserve, improve or restore health". The activity of monitoring and reducing drugs misuse reduces drug-related iatrogeny by decreasing the avoidable portion of this iatrogeny (Bates et al., 1999) (Gurwitz et al., 2003). However, the misuse of drugs in France is not well studied, and is mainly the subject of ad hoc studies. For example, the French Sentinel network has studied the misuse of antiemetics in patients with gastroenteritis (Roussel et al., 2013). Regional Pharmacovigilance Centres (CRPV) only study cases that are reported spontaneously by health professionals. They usually exploit the national notification database of the National Drug Safety Agency (ANSM) and those of the social security system.

In view of the impacts and costs that this may entail, dedicated systems to continuous monitoring, screening, quantification and monitoring of drug misuse therefore appears necessary. In this respect, the development of the use of social networks and health forums, which are nowadays communication channels more widespread, constitutes an opportunity to have an alternative source of information in near real time that could be exploited in order to identify very early signals and to identify new uses hitherto unknown. While some research works have been conducted to identify drugs side effects from this type of data source, the issue of misuse in a generic way has not been addressed. Some academic research work (Bigeard et al., 2018) or some commercial solutions, such as Detect [20] from the Kap Code company, attempts to overcome the issue but in a limited settings.

We have addressed this issue as part of the health informatics **PhD thesis of Sébastien Cossin** (2018-2021) that I co-supervise with Vianney Jouhet and in the context of two projects : the regional PEPS project DIMMER (2016-2017, funded by the CNRS and IDEX Bordeaux that I coordinated for the Team ERIAS and conducted in collaboration with three other teams [21], and the DOMINO project (coordinated with Frantz Thiessard) as part of the DRUGS Systematised Assessment in real-liFe Environment (DRUG-SAFE) platform, which has been launched in 2015 and constituted of six research groups [22].

Posts from online forums are characterised by short texts, often expressed in informal, even slang, language, which make their processing very challenging, particularly in the French context. The approach we followed is based on the computation of intersection between knowledge about drugs as declared in official Summary of Product Characteristics (SPC) for marketed drugs and what is expressed in the forum posts by users. The SPC indicates among others the list of indications (clinical sign, pathology or situation affecting a patient), possible side effects, dosage, etc. for each drug. The detail of the approach we developed is as follows.

1. the design and development of a health web forums scrapping approach in order to build a corpus for latter analysis. It has been applied on a popular French health forums [23] which allowed the acquisition of a forum posts corpora. These posts have been included in a semantic platform thanks to the use of the Content Exchange and Semantic Interoperability Between Social Networks (SIOC) model [24]. This work has been part of the engineering contract of Thimothée Jourdes and

20. https ://www.kapcode.fr/detect/

21. Pr. Richard Chbeir from LIUPPA research Lab of Pau, Romain Bourqui LaBRI Bordeaux, Antoine Pariente BPH INSERM 1219 and Benjamin Renoust from Unité Mixte Internationale CNRS UMI 3527 JFLI Japan and National Institute of Informatics (NII)

22. The BPH INSERM 1219 teams : Pharmacoepidemiology, Epidemiology of Ageing, IETO, BioStatistics and SISTM, ERIAS ; in addition SESSTIM and ORS PACA Marseille UMR 1252. The project is coordinated by Pr Antoine Pariente (BPH INSERM 1219)

23. www.doctissimo.fr/

24. https ://www.w3.org/2008/09/msnws/papers/sioc.html

Bruno Thiao-Layel I was supervising. A dataset of about 5 million forum posts has been constituted and semantised and stored in a graph based backend.

2. a semantic annotation of unstructured data (forum posts) and identification of diseases/symptoms on one hand and drugs mention in the other hand. The drugs mention identification relies on the Romedi resource developed as part of the DoMINO project by Sébastien Cossin as part of his PhD thesis. Romedi is an open data source about French drugs on the semantic web (Cossin, Lebrun et al., 2019).
3. metrics definition and indicators algorithms for misuse detection from large web forum posts. These metrics are defined as part of the MSc internship of Louis Biliet.
4. a visual analytics platform as a proof of concept dedicated to pharmacovigilance experts. It is accessible at http ://domino-project.fr :8888/. The platform is intended to be used by the ANSM and different CRPV in France.

Semantic Information Browsing and Retrieval : application to infectious disease

Lets elaborate more on the work done in the context of the SeaLife project and the Corese-NeLI Semantic Web Browser which has been designed and implemented with the purpose to enable semantic navigation of the NeLI infectious disease portal. This work is part of a collaboration with Khaled Khelif and colleagues from INRIA Sofia Antipolis, who were partners of SeaLife.

The sheer volume of resources available online makes it increasingly harder for users to find specific information and make quality judgements. This problem is of particular concern to the life sciences, where sharing and making data available on the Web is widely accepted. This motivated the SeaLife project with the aim at providing a Semantic Web Browsers (SWB) for Life Sciences. SWB in our context stands for any browser which complies with the following characteristics :

— uses at least one KOS (Mazzocchi, 2018), either a structured vocabulary or an Ontology, to support the browsing ;
— is able to identify and highlight *useful* terms in the content of the portal being visited ;
— enables semantic interpretation of these Web pages and adds semantic hyperlinks to their highlighted terms ;
— gathers additional information from the highlighted terms, which may involve access to external data and services (e.g., European Bioinformatics Institute portal or PubMed [25]) which are referred as targets.

The CORESE-NeLI engine (Diallo, Khelif et al., 2008) that has been designed supports the navigation of a portal by the use of a KOS. The browser can perform : a) a semantic search and b) semantic browsing of the NeLI portal. It bases its semantic search on semantic annotations generated from Web pages using the NeLI Ontology, and using the relationships in the KOS (i.e., narrower, broader, related to) to retrieve annotated pages related to the user query. For semantic browsing, CORESE-NeLI can identify and highlight, in a Web page being visited, terms retrieved from a structured vocabulary. From the highlighted terms, it can then create links to related pages within the portal, enabling semantic browsing. The main component of the SWB are :

25. https ://www.ncbi.nlm.nih.gov/pubmed/

— a queries generator : which uses keywords entered by the user or extracted automatically from web pages being visited. It generates various SPARQL queries in different formats to query pre-identified search.
— a profile manager : which implements the approach proposed in (Mrabet et al., 2007) to recommend pages according to the detected user's profile. It relies on semantic annotations and navigation logs for learning model profiles and classifying the different users.
— a reasoning mechanism : which consists of a set of rules allowing, (i) the building of a navigation scenario, and (ii) the use of advanced functionalities of the Corese search engine (Corby et al., 2004), such as performing approximate search.
— ontology-based annotations : which describes (i) knowledge embedded in web pages ; (ii) metadata on web pages including the creation date and the source ; (iii) learnt users' profiles, and (iv) information on relevant external sources.

The Corese-NeLI semantic engine has been evaluated using the classical NeLI portal as control, alongside the other SeaLife SWB in two settings (Oliver et al., 2009) : remotely online (for 3 months) and workshop based setting. Our findings suggest that the Corese-NeLI browser reduces the time taken for users to find information or perform tasks and that they prefer semantic links when available. In contrast, it does not shortens the pathway taken to find information or perform tasks.

Beside the research work conducted in the above mentioned axes, I am investing some efforts in contributing to the exploitation of information and communication technologies (ICT) for contributing to solve public health issues in countries with limited settings. The work on the WATRIMed knowledge graph is part of this effort.

1.4.4 Enabling Digital Public Health in Limited Settings Countries

While the development of ICT has increased substantially the availability of data nowadays, particularly in developing countries, low and middle income countries as defined by the United Nation Development Program (UNDP) [26], with limited settings are still struggling with the acquisition and analysis of valuable data for making efficient decision to tackle various issues such as those related to Public Health. In this context, a new emerging field, ICT for development (ICT4D) (Heeks, 2017) is attempting to overcome these limitations. It refers to the application of ICT toward social, economic, and political development, with a particular emphasis on helping poor and marginalised people and communities. One of the major idea is to rely on alternative data to those from official statistics systems, such as mobile based data referred as Call Detail Records.

Call Detail Records

Call Detail Records (CDRs) are (meta)data collected by mobile network operators in the course of providing their daily service (Jones et al., 2018). This type of voluminous and passively collected data are increasingly being used in research along with other forms of big data. In the context of countries with limited settings, like Africa, where collecting traditional data is a very challenging task, CDRs could be used as complementary data for tackling some issues. Indeed, with more than half a billion of mobile phone users in 2015, mobile has emerged as the platform of choice for creating, distributing and

26. https ://www.undp.org/content/undp/en/home.html

consuming innovative digital solutions and services in Africa. Several factors are driving this trend, including the expansion of advanced mobile networks, lower selling prices, growing adoption of smart devices, convenience of accessing real-time, feature-rich content and services on the go, and underdevelopment of alternative technologies, notably fixed-line connectivity, in the region. Africa is the first continent to have more mobile phone subscribers than fixed-line subscribers. According to the International Telecommunication Union [27] survey, Africa has become the world's fastest-growing mobile phone market. This open up a great opportunity to generate these emerging CDRs data type with potential for public good.

Large CDR Data Analysis from the Data for Development Challenge.

I coordinated the participation of a consortium of 4 research groups (Team ERIAS INSERM 1219 Univ. Bordeaux, ComNet Lab Aalto University Finland, LIPPA Univ. Tunis El Manar of Tunisia and VMASC Laboratory of Old Dominion University USA) for a one year project in the context of D4D - Data For Development challenge (2014-2015), a project that won the practical application prize award in 2015 (Mutafungwa et al., 2015).

Data for Development (D4D) is an innovation challenge open on ICT Big Data for the purposes of societal development. In 2013, Sonatel and the Orange Group Telecom Operator made anonymous data, extracted from the mobile network in Senegal, available to international research laboratories, as well as contextual data. Three dataset, constituted of CDR about phone calls and text exchanges between more than 9 million of Orange mobile operator customers in Senegal between January 1, 2013 to December 31, 2013 were provided to participants. The datasets are : i) one year of site-to-site traffic for 1,666 sites (mobile antennas) on a hourly basis ; ii) fine-grained mobility data (site level) on a rolling 2-week basis for about 300,000 randomly sampled users ; One year of coarse-grained (123 arrondissement level) mobility data at individual level for about 150,000 randomly sampled users.

Our work in this context falls in the area of ICT for Development in the healthcare sector, in particular the *integration and analysis of large CDR* with alternative data to enable decision making. As the decision of the establishment of a healthcare infrastructure in an area depends on many parameters taken into account by health authorities, we investigate whether data from the use of mobile phones could feed this reflection. In order to do this, we chose two diseases that require rapid hospitalisation for their care : myocardial infarction and stroke. The objective of the study was to integrate CDR data and other relevant alternative data in order to show the areas of a country in which the absence of a nearest hospital can result in death or serious sequelae.

In the approach we proposed, the mobile phone antenna coverage was estimated by the use of Voronoi diagrams (Baert et Seme, 2004). The real population density in each antenna area was estimated with the mobile population density. A total of 40 hospitals located across the 14 regions of Senegal where considered for the study. The maximum distance around each hospital was estimated to be reached in 90 minutes or three hours (corresponding to the time limit for the two diseases considered). The numbers of expected cases for the two diseases were estimated with the incidence rates of stroke and Myocardial infarction in the population, and the number of people in each antenna area.

As a result, from the expected 13,508 strokes each year, only 462 (3.42%) will occur

27. https ://www.itu.int/

too far from a hospital in the country to be able to have the thrombolysis treatment, because 96% of the population can reach a hospital in less than 3 hours. By cons, from the expected 24,315 myocardial infarctions, 4,241(17.4%) will occur too far from a hospital to be able to have the balloon treatment because they cannot reach the hospital in less than 90 minutes. These findings suggest that when large set of CDR are integrated with other contextual data and knowledge, they can contribute to the production of indicators that can be used for decision-making in resource-limited settings.

In the next chapters I will detail some of the contributions and research work that have been outlined above.

Chapitre 2

Large Scale Biomedical TextMining and Semantic IR

Contents

The work which contributed mainly to this chapter has been addressed in the following publications :

G. Diallo, K. Khelif, O. Corby, P. Kostkova, Gemma Madle. Semantic Browsing of a Domain Specific Resources : The Corese-NeLI Framework. Proceeding WI-IAT '08 Proceedings of the 2008 IEEE/-WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Volume 03 Pages 50-54.

H. Oliver, G. Diallo, E. de Quincey, D. Alexopoulou, B. Habermann, P. Kostkova, M. Schroeder, S. Jupp, K. Khelif, R. Stevens, G. Jawaheer, G. Madle. A user-centred evaluation framework for the Sealife semantic web browsers. BMC Bioinformatics. Suppl 10 :S14 2009

E. de Quincey, H. Oliver, P. Kostokova, G. Jawaheer, G. Madle, G. Diallo, D. Alexopoulou, M. Shroeder, B. Habermann, K. Khelif, S. Jupp, R. Stevens. Mining for Pattern of Semantic Link Usage : Do Domain Users Actually Like Semantic Browing ? International Workshop on Intelligent Web Intereaction (IWI'09), Milano, Italy. 2009

K. Dramé, F. Mougin G. Diallo. Large scale biomedical texts classification : a kNN and an ESA-based approaches. J. Biomedical Semantics. Vol :7(40), 2016. DOI Information : 10.1186/s13326-016-0073-1

K. Dramé, F. Mougin, G. Diallo. A k-nearest neighbor based method for improving large scale biomedical document indexing. The 6th International Symposium on Semantic Mining in Biomedicine (SMBM), 6th-7th October, 2014, Aveiro, Portugal

K. Dramé, F. Mougin, G. Diallo. Query expansion using external resources for improving information retrieval in the biomedical domain. CEUR WS Lab Proceedings for CLEF 2014. Vol-1180 :189-194

S. Cossin, V. Jouhet, F. Mougin, G. Diallo, F. Thiessard. IAM at CLEF eHealth 2018 : Concept Annotation and Coding in French Death Certificates. Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum CEUR-WS. Vol-2125.

S. Cossin, L. Lebrun, G. Lobre, R. Loustau, V. Jouhet, R. Griffier, F. Mougin, G. Diallo, F. Thiessard : Romedi : an open data source about French drugs on the semantic web. MEDINFO'2019, 26-30 Aug 2019, Lyon, France

In Preparation

S. Cossin, L. Biliet, F. Lalane, R. Bourqui, V. Jouhet, F. Mougin, G. Diallo, F. Thiessard. Drugs misuse detection from social networks : case of French Health Forums. To be submitted to Journal of Medical Internet Research.

2.1 Introduction

One of the main challenges when dealing with data in the healthcare domain is that about 80% of them remain unstructured and untapped after it is created (e.g., text, image, signal, etc.) (Kong, 2019). Among these unstructured data, free texts type of data are very pregnant, from those available from EHRs, including clinical notes, discharge summaries, to biomedical literature and social networks based data which are exploding exponentially with the development of ICT technologies (Weber et al., 2014). These data are expressed in natural language, which makes their automated processing increasingly difficult (for languages other than English in particular (Névéol et al., 2018)) and very challenging to efficiently access to useful information.

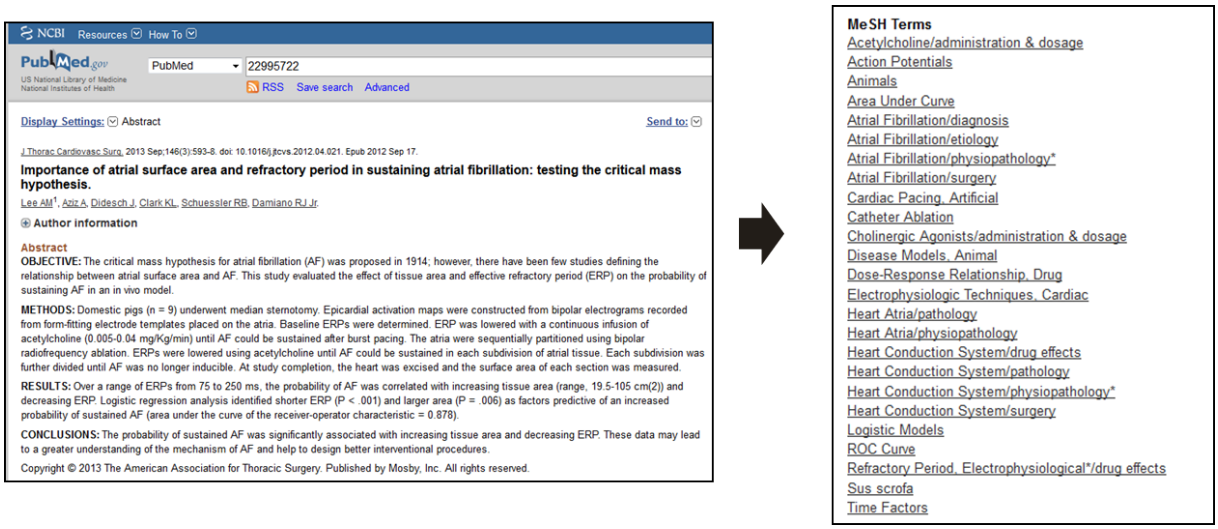

FIGURE 2.1 – Illustration of the problematic of biomedical literature mining in our context.

In this chapter I do address the issue of extracted relevant information from textual data using IR and NLP techniques. I will present the challenge related to efficiently handling large biomedical literature data (several millions of documents) in order to enable access to recent findings reported in scientific literature easily. Then, in the context of the Drug Misuse Identification project I will focus on how we handle data (expressed in natural language, sometimes in a poor, informal or degraded vocabulary) from online social networks and web forums in particular to detect drug misuses. The last presented work is conducted earlier in the context of the SeaLife European project with the issue of semantic browsing of a domain specific resource and semantic IR using an Ontology as knowledge background.

2.2 Large scale biomedical literature mining using kNN

In this section I describe briefly the work conducted as part of the **PhD thesis of Khadim Dramé**. This work is done in the context of the Semantic BioblioDem portal (SemBip) project (2012-2014) in collaboration with the team Epidemiology of Ageing of the INSERM 897 (DRAMÉ, DIALLO et al., 2014). The main purpose of the reported study is to automatically annotate and classify a large set of documents from the online PubMed databse when only a partial information is available (title and abstract) and approximate the results with the manually annotated classes from the MeSH thesaurus (LOWE et BARNETT, 1994), the MeSH headings, provided by PubMed indexers. The problematic could be seen as the task of annotating a whole document when only a partial information is available. It is illustrated in Figure 2.1 where from the abstract of the article with id 22995722, a set of MeSH descriptors (concepts) which describe the whole corresponding document are proposed. The idea is to perform such a task in an automated way on several million of documents.

2.2.1 Brief Overview of Related Work

In order to facilitate accessing textual content of biomedical scientific literature, a possible solution is to bring a suitable representation over the textual documents. Controlled and structured vocabularies, such as the Medical Subject Heading (MeSH®) thesaurus, are widely used to assist in this task (Trieschnigg et al., 2009) and consequently facilitate access to useful information (Crespo Azcárate et al., 2013 ; Díaz-Galiano et al., 2009) thanks to semantics brought to index the documents.

The Medical Text Indexer (MTI) tool (Aronson, Mork et al., 2004) is one of the first attempt to index biomedical documents (MEDLINE citations) using controlled vocabularies. To map biomedical text to concepts from the UMLS Metathesaurus - a system that includes and unifies more than 160 biomedical terminologies - the MTI tool uses the well-known MetaMap concept mapper (Aronson, 2001) and combines its results with the PubMed Related Citations algorithm (Lin et Wilbur, 2007). More recently, the MTI was extended with various filtering techniques and ML algorithms in order to improve its performance (Mork et al., 2013). Ruch has designed a data independent hybrid system using MeSH for automatically classifying biomedical texts (Ruch, 2006). The first module is based on regular expressions to map texts to concepts while the second is based on a Vector Space Model (Salton et al., 1975) considering the vocabulary concepts as documents and documents as queries. Then, the rankers of the two components are merged to produce a final ranked list of concepts with their corresponding relevance scores. The results showed that this method achieved good performances, comparable to ML-based approaches. One limitation of this system is that it may return MeSH concepts which match partially the text (Trieschnigg et al., 2009).

Machine learning based classification

ML-based approaches are an alternative proposed to deal with biomedical texts classification. The idea is to learn a model from a training set constituted of already annotated documents and then to use this model to classify new documents. Trieschnigg et al. have shown in a comparative study of six MESH-based classification systems that the kNN-based method outperforms the others, including the MTI (Trieschnigg et al., 2009). In their work, the kNN classifier uses a language model (Ponte et Croft, 1998) to retrieve documents which are similar to a given one. The relevance of MeSH descriptors is the sum of the retrieval scores of documents annotated by these descriptors among the neighbouring documents. A similar kNN-based approach has been proposed in (Huang et al., 2011). A language model is used to retrieve the neighbours of a given document. Then, a learning-to-rank model (T.-Y. Liu, 2007) is used to compute relevance scores and consequently to rank candidate labels collected from these document neighbours. The number of labels to classify a document in this case is set to 25. Experiments conducted on two small standard datasets (respectively 200 and 1,000 documents) showed that it achieves better performances than the MTI tool.

Multilabel Classification

Classifying biomedical documents in which each document of the considered dataset is assigned one or several categories (also called "labels") can be assimilated as a **multi-label classification task**. Multi-label classification (MLC) has been subject to various studies and especially for text classification purposes (Tsoumakas et al., 2009). The dif-

ferent approaches can be classified into two main categories : the problem transformation approach (Read et al., 2011) and the algorithm adaptation approach (Spyromitros et al., 2008 ; M.-L. Zhang et Zhou, 2007). The problem transformation approach splits up a multi-label learning problem into a set of single-label classification problems whereas the algorithm adaptation approach adjusts learning algorithms to perform MLC.

KNN-based approach for Multilabel Classification (MLC) has been proven efficient for MLC in terms of simplicity, time complexity, computation cost and performance (Spyromitros et al., 2008). Zhang and Zhou (M.-L. Zhang et Zhou, 2007) proposed a Machine Learning-KNN (for Multi-Label kNN) method which extends the traditional kNN algorithm and uses the maximum a posteriori principle to determine relevant labels of an unseen instance. For a given instance t, the ML-KNN identifies its neighbours and estimates respectively the probabilities that t has and has not a label l based on the training set, for each label l. Then, it combines these probabilities with the number of neighbours of t having l as a category to compute the confidence score of l.

Spyromitros et al. propose a similar approach, Binary Relevance KNN (BR-KNN), and two further extensions (Spyromitros et al., 2008). BR-KNN is an adaptation of the kNN algorithm using a BR method which trains a binary classifier for each label. Confidence scores for each label are computed using the number of neighbours among the k neighbours that include this label. In (Madjarov et al., 2012), an experimental comparison of several multi-label learning methods is presented. In this work, different approaches were investigated using various evaluation measures and datasets from different application domains. The experiments showed that the best performing method is based on the Random Forest classifier (Kocev et al., 2007).

Other recent works address MLC with large number of labels (Bi et Kwok, 2013). Indeed, in many applications, the number of labels used to categorise instances is commonly very large. For example, in the biomedical domain, the MeSH thesaurus consisting of thousands descriptors (27,149 in the 2014 version) is often used to classify textual documents. The large number of descriptors can affect the effectiveness and performance of multi-label models. To address this issue, a label selection based on randomised sampling is performed (Bi et Kwok, 2013).

Semantic and Deep Learning for document indexing

Over the years there have been various learning methods that have been proposed with attempts based on the shallow semantic features, such as latent semantic indexing (LSI) (Deerwester et al., 1990), Latent Dirichlet Allocation (LDA) (Blei et al., 2003) and probabilistic Latent Semantic Indexing (pLSI) (Hofmann, 1999).

In the last years deep learning has arisen the focus of many researchers. Some have tried Deep Learning approaches for semantic indexing as it could be used to represent more semantic information other than frequency. Thus, Wu et al. have designed a deep structure with restricted Boltzmann machines (RBMs) to compute semantic representation of documents (H. Wu et al., 2014). A Deep Learning for Multi-label Classification approach improved this work by proposing a feature space which can make labels less interdependent and easier to model and predict at inference time. Convolutional Neural Networks (CNN) has been used for Multi-label Text Classification in (J. Liu et al., 2017), while a Recurrent CNN has been used in (Lai et al., 2015) for texts classification. More recently, a Boltzmann CNN (B-CNN) framework has been proposed in (Y. Yan et al., 2018) for biomedical semantic indexing.

In the following, I detail the $kNN\_Classif$ approach for large biomedical scientific publications annotation when only partial information is available.

2.2.2 $kNN\_Classif$: large biomedical literature classification when partial information is available

Nearest neighbours' retrieval and candidate labels collecting

For a given document, represented by a **vector of unigrams** its k most similar documents are retrieved (Figure 2.2). The TF.IDF weighting scheme is used to determine the weights of different terms in the documents. Then, the cosine similarity between documents is computed (Salton et al., 1975). Once the k nearest documents of a target document are retrieved, the set of labels assigned to them are used for training the classifiers (in the training step) or as candidates for classifying the document (in the classification step). Labels, which are the instances here, are first represented by a set of attributes. Thereafter, ML algorithms are used to build models which are then used to rank candidate labels for annotating a given document. For ranking the labels, different learning algorithms are explored.

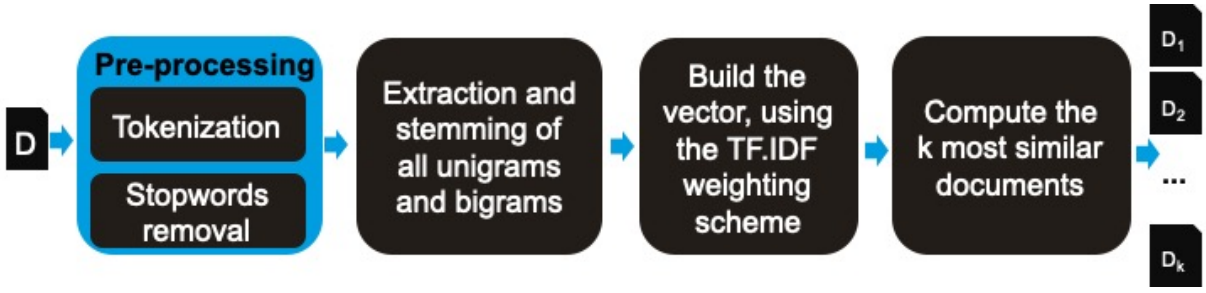

Figure 2.2 – Workflow of the identification of the k most similar documents.

The purpose of the nearest neighbours' retrieval is to retrieve the k most similar documents (KNN) for each document. As commonly done, a collection of documents previously annotated are used. Like the PubMed Related Citations approach (Lin et Wilbur, 2007), we consider that two documents are similar if they address the same topics. The cosine similarity measure is chosen for that purpose. It is commonly used in text classification and information retrieval with the VSM (Salton et al., 1975). The documents are first segmented into sentences and tokens, while stop words are removed. From these pre-processed texts, all unigrams are extracted and normalised according to a stemming technique (Porter, 1997). Then, the cosine measure enables to compute similarity between documents, which are represented by vectors of unigrams. Formally, let $C = d_1, d_2, ..., d_n$, a collection of n documents, $T = t_1, t_2, ..., t_m$, the set of terms appearing in the documents of the collection and the documents d_i and d_j being represented respectively by the weighted vectors $d_i = (w1_i, w2_i, \ldots, wm_i)$ and $d_j = (w1_j, w2_j, \ldots, wm_j)$, their cosine similarity is defined by :

$$Sim(d_1, d_2) = \frac{\sum_{k=1}^{m} w_k^i . w_k^j}{\sqrt{\sum_k^m (w_k^i)^2}\sqrt{\sum_k^m (w_k^i)^2}}$$

For a given document, once its kNN are retrieved, all labels assigned to these retrieved documents are gathered in order to constitute a set of candidate labels likely to annotate it (Figure 2.3. Since this process can be seen as a classification problem, we use ML

techniques to rank these candidate labels. Thus, classical classifiers are used to build classification models which are then exploited to determine the relevant labels for annotating any unseen document. For that purpose, candidate labels are used as training instances (in the training step) or instances to be classified (in the classification step).

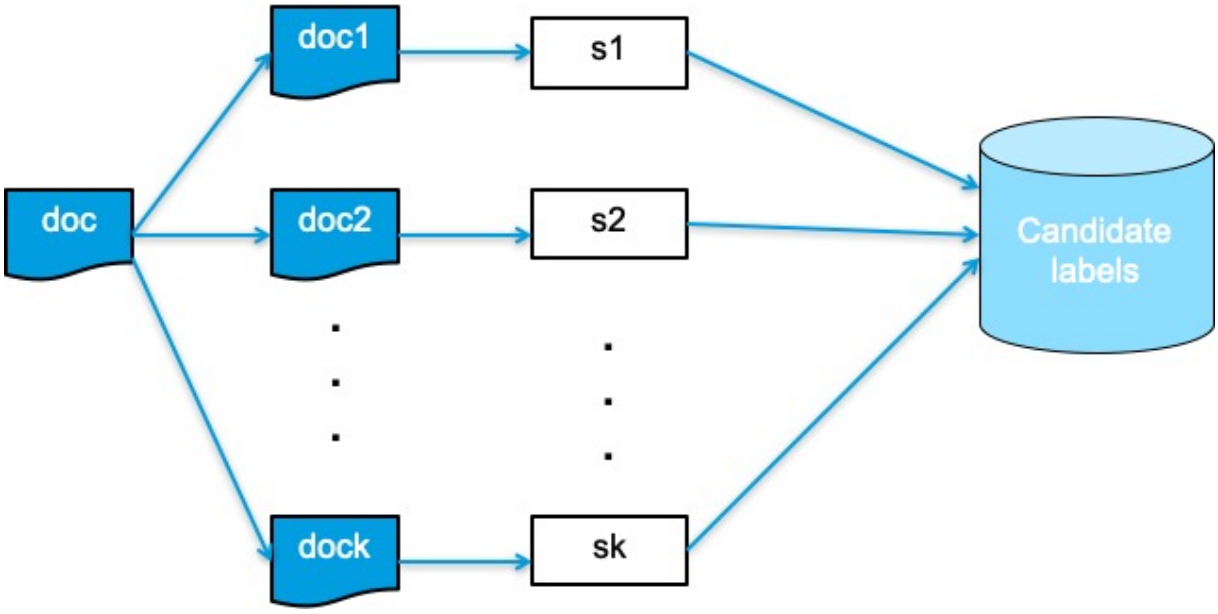

FIGURE 2.3 – Retrieving the candidate labels from the k most similar documents.

Feature extraction

For indexing biomedical related literature, we have previously seen that the MeSH thesaurus is commonly used. The latter is composed of a set of descriptors (also called main headings) organised into a hierarchical structure. Each descriptor includes synonyms and related terms, which are known as its entry terms.

To determine the relevance of a candidate label, it is represented by a vector of features. In the training step, its class is set to 1 if the label is assigned to the target document and 0 otherwise. In the classification step, the model uses the label features to determine its class. According to previous related works ((SPYROMITROS et al., 2008)) we have defined six features. Formally, let $L = l_j, j = 1, \ldots, n$ be the candidate labels set of a new document d, and $V = \{d_i\}, i = 1, ..., k$ its k nearest neighbours, the values of these attributes for the label l_j are respectively defined as :

$f_1(l_j) = \frac{1}{k} \sum_1^n assigned(l_j, d_i)$ and $f_2(l_j) = \frac{1}{k} \sum_{l_j \in d_i}^n sim(d, d_i)$

— **Feature 1** : for each candidate label, the number of neighbour documents to which it is assigned is used as a feature. This value represents an important clue to determine the class of the label. Moreover, in the classical kNN-based approach, it is the only factor used to classify a new instance. In practice, a voting technique is used to assign the instance to the class that is the most common among its k nearest neighbours.

— **Feature 2** : for each candidate label, the similarity scores between the document to classify and its nearest neighbours annotated with this candidate label are summed and this sum is another feature. Since the distance between a document and each of its neighbours is not the same, we consider that the relevance of the labels assigned

to them for the target document is inversely proportional to this distance. In other words, the closer a document is to the target document, the more its associated labels are likely to be relevant for the latter. In (Trieschnigg et al., 2009), this is the only feature used to determine the relevance scores of candidate labels.

— **Feature 3** a flag indicating whether for each candidate label, all its constituent tokens appear in the title and abstract. This binary feature has been chosen because it captures disjoint terms (terms constituted of disjoint words) which are frequent in the biomedical texts.

— **Feature 4** : indicates whether for each label (known as descriptor here), one of its entries appears in the document. It is set to 1 in this case, to 0 otherwise.

— **Feature 5** : this feature indicates the frequency of the descriptor in the document.

— **Feature 6** : is used to verify whether a candidate label is contained in the document's title. The assumption is that if a label appears in the title, it is relevant for representing this document.

The relevance of each of these features is estimated using the information gain measure (Table 2.1). The first two features mainly permit to compute relevance scores of candidate labels.

Feature	**Description**	**Information gain**
Feature 1	Number of neighbours in which the label is assigned	0.16
Feature 2	Sum of similarity scores between the document and all the neighbours' document where the label appears	0.17
Feature 3	Check whether all constituted tokens of the label appear in the target document	0.01
Feature 4	Check whether one of the label entries appears in the target document	0.03
Feature 5	Frequency of the label if it is contained in the document	0.03
Feature 6	Check if the label is contained in the document title	0.02

Table 2.1 – Chosen features information gain for the $kNN\_Classif$ algorithm

Classifier building

To build the classifiers, a labelled training set consisting of a collection of documents with their manually associated labels is constituted. For each document in the training set, its nearest neighbours and their manually assigned labels are collected. Each label of this collected set is considered as an instance for the training. Thus, for each label, its different features, described previously, are computed. Thereafter, labels obtained from neighbours of the different documents of the training set are gathered to form the training data. Then, classifiers are built from this labelled training data. We have tested the following classification algorithms : Naive Bayes (NB) (Langley et al., 1992), Decision Trees (DT also known as C4.5 in our case (Quinlan, 1993)), Multilayer Perceptron (MLP) and Random Forest (RF) (Breiman, 2001). We have chosen these classifiers as they have yielded the best performances in the tests we have performed.

Selection of the optimal value of N for top scoring documents

Given a document to be classified, the candidate labels collected from its neighbours are represented as the training ones (see the previous section). The trained model is then used to estimate the relevance score of each candidate label. Indeed, the model computes, for each candidate label, its probabilities to be relevant or not. From these probability measures, the relevance score of each label is derived. Candidate labels are then ranked according to their corresponding scores and the N top-scoring ones are selected to annotate the document (indexing descriptors), where N is determined using three different strategies.

— Initially, N is set as the number of labels with a relevance score greater than or equal to a threshold arbitrarily set to 0.5. This strategy based only on the relevance score of the label according to the document is inspired by the original kNN algorithm.
— Then N is set as the average number of assigned labels collected from the neighbours. This strategy has been successfully used for extending the kNN-based method proposed in (Spyromitros et al., 2008).
— In the third strategy, we use the method described in (Mao et Lu, 2013). The principle is to compare the relevance scores of successive labels of a list of candidate labels ranked in descending order for determining the *cut off* condition enabling to discard the irrelevant or insignificant ones. This strategy is defined by the following formula : $\frac{S_i+1}{S_i} \geq \frac{i}{i+1+\alpha}$, where s_i is the relevance score of a label being at position i and α a constant whose optimal value is determined empirically.

We have used the Waikato Environment for Knowledge Analysis (WEKA) framework (Hall et al., 2009) for the implementation of these classifiers. It integrates various ML algorithms, including the four ones we have tested.

2.2.3 Evaluation of the $kNN\_Classif$ algorithm

Datasets and Experiments Settings

The first experiment of $kNN\_Classif$ is performed in the context of a dataset provided by the BioASQ challenge [1]. BioAsq is a challenge on large scale biomedical semantic indexing and question answering (QA). The challenges include tasks relevant to hierarchical text classification, machine learning, information retrieval, QA from texts and structured data. In the Large-scale semantic indexing, referred as Task 1a, the goal is to classify documents from the PubMed digital library [2] into concepts of the MeSH hierarchy. PubMed is the most widely used literature database. It comprises more than 27 million citations for biomedical literature from MEDLINE, life science journals, and online books. In the context of BioASQ new PubMed articles that are not yet annotated are collected on a daily basis. These articles are used as test sets for the evaluation of the participating systems. The performance evaluation is made according to the annotations that are available from the PubMed curators (Balikas et al., 2015).

The different experiments have been conducted using the computing facility of the Bordeaux Mésocentre, Avakas [3] at the time of evaluation. We used two computation nodes c6100, which provide 48 GB of RAM and 24 cores Intel Xeon X5675.

1. http ://www.bioasq.org
2. https ://www.ncbi.nlm.nih.gov/pubmed/
3. https ://redmine.mcia.univ-bordeaux.fr/

Evaluation Metrics

As indexing biomedical documents can be assimilated to a MLC problem, instead of one class label, each document is assigned a list of labels. Thus, measures that are traditionally used for evaluating indexing methods were adapted for the MLC context (Tsoumakas et al., 2009). Thus, the example-based precision (EBP) measures how many of the predicted labels are correct. It is expressed as follows : $EBP = \frac{1}{m}\sum_{i=1}^{m}\frac{|Y_i \cap Z - i|}{|Z_i|}$. The example-based recall (EBR) measures how many of the manually assigned labels have been retrieved. It is expressed as follows : $EBR = \frac{1}{m}\sum_{i=1}^{m}\frac{|Y_i \cap Z - i|}{|Y_i|}$. Since EBP and EBR evaluate partially the performance of a method, like traditional Precision and Recall do, the example-based f-measure (EBF) combines both measures for a global evaluation. It is expressed as follows : $EBF = \frac{1}{m}\sum_{i=1}^{m}\frac{|Y_i \cap Z - i|}{|Z_i|+|Y_i|}$. The accuracy (Acc) is also a complementary measure expressed as : $EBF = \frac{1}{m}\sum_{i=1}^{m}\frac{|Y_i \cap Z - i|}{Z_i \cup Y_i}$.

Results of the experiments

Table 2.2 depicts the results achieved of our best system for the bacth 3. In tests 2 and 5, our best system uses a Naïve Bayes classifier and selects only labels with a confidence score greater than or equal to 0.5 while in the others, the best system sets N to the average size of the sets of labels collected from the neighbours. In most cases, using this value for N yields better or similar results than the other strategy. In the challenge, we do not use the automatic *cut-off* method to fix the number of labels as described in (Mao et Lu, 2013) but in the second experiment, this technique is explored.

In the second experiment, we tested the combinations of various classifiers with different techniques defining the number of labels suitable to annotate a given document. The evaluation of configurations with the two best classifiers in our experiments, NB and RF, are presented in Table 2.3. The parameter k is empirically set to 25 using a cross-validation technique. When the minimal score threshold is used, the precision often increases significantly, mainly with the RF classifier but the recall is lower. Regarding the **average size strategy**, it yields a good recall but the precision decreases slightly. In this case, the results of both classifiers are similar but the RF one slightly outperforms the NB classifier. The best results are achieved with the cut-off method which balances both precision and recall, and yields the best F-measure. Except for the minimum threshold technique where the NB classifier results are better, the best F-measure is achieved with the RF classifier. The DT (C4.5 algorithm of Weka) and the Multilayer Perceptron (MLP) classifiers have also been tested but their results are less interesting. The former yields lower performances while the latter performs very slowly and gets results comparable to the RF ones. The MLP classifier requires more CPU and memory during the training process.

When the training set is raised from 20,000 to 50,000, the performances are slightly improved in two test sets (one of 1,000 documents and another of 2,000). Table 2.4 presents the results of the different classifiers in this larger training set. The value of α (constant used in the strategy based on label scores comparison - strategy c- for optimising N) also affects the classification performance. The lower the value of α, the higher the precision is but the lower the recall is and vice versa. In these experiments, we set α to 1.6 which yields the best results using cross-validation techniques. Furthermore, we note that when the classifiers are trained on this extended dataset, they yield similar performances but the RF classifier slightly outperforms the others. In terms of training time, NB, DT and

RF classifiers performed similarly with respectively 4, 6 and 9 minutes once data were represented in suitable format for the Weka tool (e.g. Attribute-Relation File Format (ARFF) format). The pre-processing step (retrieval of neighbours and computation of features values) however takes more times (1 hour and 43 minutes). Note that since we have different types of attributes (binary and numeric), we discretise the latter in nominal attributes. The MLP classifier is, meanwhile, very costly in terms of training time (23 hours).

Test	**# documents**	**System**	**EBP**	**EBR**	**EBF**
Test 1	2.961	$kNN\_Classif$	0.55	0.48	0.49
		BestBioAsq	0.59	0.62	0.58
Test 2	5.612	$kNN\_Classif$	0.52	0.5	0.48
		BestBioAsq	0.62	0.60	0.60
Test 3	2.698	$kNN\_Classif$	0.55	0.49	0.49
		BestBioAsq	0.64	0.63	0.62
Test 4	2.982	$kNN\_Classif$	0.49	0.55	0.49
		BestBioAsq	0.63	0.62	0.62
Test 5	2.697	$kNN\_Classif$	0.50	0.53	0.48
		BestBioAsq	0.64	0.61	0.61

TABLE 2.2 – Results of our $kNN\_Classif$ and the best systems participating in the BioASQ challenge on the different tests of the batch 3

.

Strategy	**Classifier**	**EBP**	**EBR**	**EBF**
a)	NB	0.58	0.49	0.49
	RF	0.74	0.34	0.43
b)	NB	0.51	0.54	0.51
	RF	0.52	0.54	0.52
c)	NB	0.56	0.52	0.51
	RF	0.61	0.52	0.53

TABLE 2.3 – Results of $kNN\_Classif$ according to the classifier and strategy used for fixing N : a) 0.5 as the minimal confidence score threshold, b) the average size of the sets of labels collected from the neighbours and c) the cut-off method. A training set of 20,000 documents is used.

Classifier	**EBP**	**EBR**	**EBF**	**Acc**
NB	0.59	0.54	0.54	0.39
RF	0.62	0.54	0.55	0.41
DT	0.63	0.52	0.54	0.39
MLP	0.64	0.46	0.51	0.36

TABLE 2.4 – Results of $kNN-Classif$ according to the classifier using the cut-off method with a training set of 50,000 documents.

Discussion

While textual classification has been widely investigated, few approaches are currently able to efficiently handle large collections of documents, in particular when *only a portion of the information* is available. This is a challenging task, particularly in the biomedical domain.

Our experiments show that our kNN-based approach is promising for biomedical documents classification in the context of a large collection. Our results confirm the findings presented in (Trieschnigg et al., 2009), where among the multiple classification systems, the kNN-based one yielded the best results. If we compare our method with the latter, we use more advanced features to determine the relevance of a candidate label. Indeed, (Trieschnigg et al., 2009) determine the relevance of a label by summing the retrieval scores of the k neighbour documents that are assigned to the label. In our method, this sum is only considered as one feature among others for determining the confidence scores of labels. While the results of our method do not outperform the extended (and improved) MTI system (Mork et al., 2013) which is currently used by the NLM curators, it gets promising results (0.49 against 0.56 of F-measure). A direct comparison with the method proposed in (Huang et al., 2011) is not simple since the authors used an older collection than the official datasets provided in the BioASQ challenge, which are more recent and annotated with descriptors of a recent MeSH thesaurus (2014 version). Similarly to their experiments, when our method is evaluated on 1,000 randomly selected documents, it outperforms this method (0.55 against 0.50 for the F-measure). But a comparison with their results in the first challenge of BioASQ (Mao et Lu, 2013) where they integrated the MTI outputs, their system performs better than ours (F-measure of 0.56 against 0.49). Compared with the two approaches proposed in (Zhu et al., 2013), one based on the MetaMap tool (Aronson et Lang, 2010) and another using IR techniques, our method gets better results (0.49 against 0.42 for the F-measure). Our approach outperforms also the hierarchical text categorisation approach proposed in Ribadas et al., 2013.

A more recent evaluation of our kNN-based approach using a larger dataset (50,000 documents) for training the classifiers shows that it provides better performances, comparable to the best methods which participated to the challenge (with an f-measure of 0.55). Moreover, unlike the extended MTI system (Mork et al., 2013), we do not use any specific filtering rules. This makes our approach generic and its reuse in other domains straightforward.

Finally, note that we have also used Explicit Semantic Analysis (ESA), which is a vector based representation of text (simple words, fragments of text, entire document) that uses a document corpus as a knowledge base representing. It is a (statistical) semantic-based approach (Gabrilovich et Markovitch, 2009) which represents documents in a conceptual space constituted of explicit concepts. In our case we used Wikipedia as conceptual space and we constituted a training set of 4,432,399 PubMed documents (titles and abstracts). It turned out that although ESA technique has shown interesting results in traditional texts classification, it yields very low performances, when it comes to classifying large set of documents when only a partial information is available, compared to basic methods using a simple correspondence between the text and the semantic resource inputs. Detailed results are provided in (Dramé, Mougin et al., 2016).

2.3 Mining Online Web Forum Data : application to Drugs Misuse Detection

The work reported in this section has been conducted in the context of the DIMMER (Détection de Mesusage de Médicaments à partir de réseaux sociaux et foras de santé) I coordinated with participation of Benjamin Renoust (JIFL NII Tokyo) and Richard Chbeir (LIAPA Pau) and Drug Misuses In NetwOrks (DoMINO) project which I co-coordinate with my colleague Frantz Thiessard. DoMINO is part of a larger project, the DRUGS Systematised Assessment in real-liFe Environment (DRUG-SAFE) platform. The study was the subject of Sébastien Cossin's engineering work, then part of the application of his PhD thesis, which I co-supervise with Vianney Jouhet. Similarly, Bruno Thiao Layel (as part of K-Ware platform) and Thimothée Jourde has intervened as engineers and Louis Bilet as part of his MSc internship. The final visualisation of the implemented tool is in collaboration with the BKB team of LaBRI (Fréderique Lalane as an engineer, Romain Bourqui and David Auber).

2.3.1 Context of Drugs Misuse Identification

In conventional medicine, drugs are at the heart of the system. According to the French public health law, they may be administered with a view to "restoring, correcting or modifying physiological functions by exercising a pharmacological, immunological or metabolic action". Before marketing a drug, a number of studies must be conducted to evaluate its efficacy and safety. These studies are mainly conducted in vitro and in animals (pre-clinical studies) and then in vivo in healthy subjects and patients. For a drug to be marketed, it must obtain a Marketing Authorisation (MA) that takes into account the results of these studies. The MA indicates the terms of use of the drug : indication(s), dosage, duration of treatment, precautions for use, contraindications, etc.

These terms of use are set out for each drug in the Summary of Product Characteristics (SPC). Most SPCs are available on the French National Agency of Drugs and Health Products (ANSM) website, as well as the patient information leaflet, or on the European Medicines Agency (EMA) website [4].

In France, pharmacovigilance is based on spontaneous reports made to regional pharmacovigilance centres (CRPV). However, there is a low reporting rate (Hazell et Shakir, 2006). Despite the low observed rate, there is a risk related to non-compliance with the terms, constituting a misuse that results from a failure to match the terms of use of the drug with the reason for the individual's taking it. It is in this context that Marisol Touraine, the former French Minister of Health, commissioned a report from three experts about real-life monitoring of drugs. The report was issued in 2017 (Begaud et al., 2017). It highlights in particular that "with the digital revolution and the increasing possibilities for information collection and analysis that it allows, the production and use of observational data is becoming a strategic objective for all health systems and renews the traditional approach to drug studies, historically focused on clinical trials".

Misuse identification is one of the activities in pharmacovigilance. In France, regulatory authorities in relation to public health and health products have provided a reference framework for the concept of misuse. Thus, the article R5121-152 of the Public Health Code define it as "an intentional and inappropriate use of a medicinal product or product

4. https ://www.ema.europa.eu/en

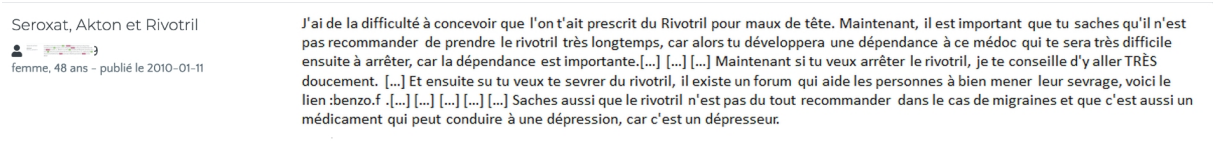

FIGURE 2.4 – Example of a forum post related to the Rivotril drug

not in accordance with the marketing authorisation or registration and recommendations for good practice" while the ANSM defines it as "an use not in accordance with the recommended conditions of use of the health product". In order to be able to combat this misuse and thus reduce the consequences on the health of individuals, it is necessary to have knowledge of these dangerous behaviours adopted by the population. This knowledge comes from analysing population data. According to (BEGAUD et al., 2017) such data can come from multiple sources : they can be extracted from patients' computerised records, or constitute a sub-product of the information used for the reimbursement of care ; it can be collected specifically, for example in the context of pharmacovigilance procedures, or for establishing health related registers or cohorts, or more occasionally as part of ad hoc studies ; they may also come from the web, social networks, connected objects, etc.

Web forums and social networks (like Twitter [5] and Facebook [6]) are nowadays digital mediums that are commonly used by people and therefore constitute an important alternative source of information. A web forum is an electronic space for interactive conversation between users. Speech is relatively free, unlike in front of a healthcare professional. However one the main challenge when dealing with data which come from such a source is the use of non standard language that is far from the specialised language in SPCs for instance.

In regards with the above context, the overall objective of the study we have conducted is to design a framework to **support drugs misuse identification in French web forums**. We have chosen as case for proof of concept the most representative one in France for the health related domain : Doctissimo [7], a French-speaking website dedicated to health and well-being. Our hypothesis is that by mining a large set of real life data from web forum posts we will be able to identify some signals that could indicate drug misuses. We proceed by scrapping forums to collect posts of interest which are then processed using NLP algorithms with the use of experts knowledge on drugs on the one hand and symptoms and diseases on the other hand. An example of a post related to the Rivotril drug is depicted on Figure 2.4.

2.3.2 Brief Overview of Related Work

Retrieving data from Web forums involves crawling for relevant pages and extracting data of interest. Thus, iRobot is an approach which involves an offline component that extracts the sitemap from sample pages and identifies the optimal path from one page to another to avoid duplication. An extension for the specific case of Web forums has been proposed later (Y. WANG et al., 2008 ; YANG et al., 2009). Thanks to the use of page-flipping links, it is able to recognise posts belonging to the same thread (or threads belonging to the same forum. A notable limitation of iRobot is its dependence to the

5. www.twitter.com
6. www.facebook.com
7. www.doctissimo.fr/

sampling pages. FoCUS is another tool which uses a learning strategy for online crawling. It learns regular expression patterns to extract the main features from a sample collection and uses these expressions for the crawling (Jiang et al., 2013). Vigi4Med Scraper is another more recent scrapping framework for Web Forum Structured Data Extraction and Semantic Representation (Audeh et al., 2017) developed in the context of the Agence National de Sécurité du Médicament funded project Vigi4Med. It allows detecting information related to posts, threads and authors and uses the SIOC semantic graphs to store extracted data

Regarding drug misuses, very limited work have addressed the issue of their detection in web forums or social networks, particularly in the French context. Bigeard et al. proposes an approach of detection and classification of drug misuses from the analysis of user-generated data in a French social media (Bigeard et al., 2018). To do so, messages from the Web forum are indexed using a builtin vocabulary which is enriched with terms adapted to those used in forums. Then a manual typology of misuses is performed before a language models for the automatic detection of misuses in forum messages. This work relies on a limited set of forum posts from the Doctissimo web forum.

The PREscription Drug abuse Online Surveillance and Epidemiology (PREDOSE) is a Semantic Web platform which is designed to facilitate the epidemiological study of prescription (and related) drug abuse practices using social media (Cameron et al., 2013). PREDOSE uses web forum posts and domain knowledge which has modelled in a manually created Drug Abuse Ontology (DAO) in order to facilitate the extraction of semantic information. It is currently in use in an online discovery support system, by prescription drug abuse researchers at the Center for Interventions, Treatment and Addictions Research (CITAR) at Wright State University (USA).

Some researchers have addressed the specific case of Adverse Drug Reaction (ADR) and social medias. Thus, Katsahian et al. (Katsahian et al., 2015) used the Net Scoring rating tool on a set of social network web sites in order to assess the capabilities of these tools to guide experts for selecting the most adapted social network web site to mine ADRs.

Regarding the twitter specific social network, non medical use of prescription medications is reported in (Kalyanam et al., 2017). They collected 11 million tweets that they filtered for three commonly abused prescription opioid analgesic drugs. Then, they used unsupervised machine learning on the collected tweets which contain mentions of one of the studied drugs in order to detect tweets about non-medical drug use.

Texts from web forums or social medias like twitter used to be short. Thus, traditional texts processing and topics uncovering, such as LDA tend to be not suitable. Therefore, approaches such as the Biterm Topic Model (BTM) (X. Yan et al., 2013) has been proposed. BTM learns the topics by directly modelling the generation of word co-occurrence patterns (i.e. biterms) in the whole corpus being analysed.

As far as we know there is no study which focuses on defining generic indicators and metrics at large scale, relying on approved knowledge like the one on SPCs, and using them for drug misuse identification in an automated way.

2.3.3 Summary of the Proposed Approach

The main idea we followed to enable drug misuse identification is to cross-link data collected from web forums with knowledge contained in the SPCs in order to identify possible differences. The overall workflow of the approach is depicted in figure 2.5 and

summarised as follows.

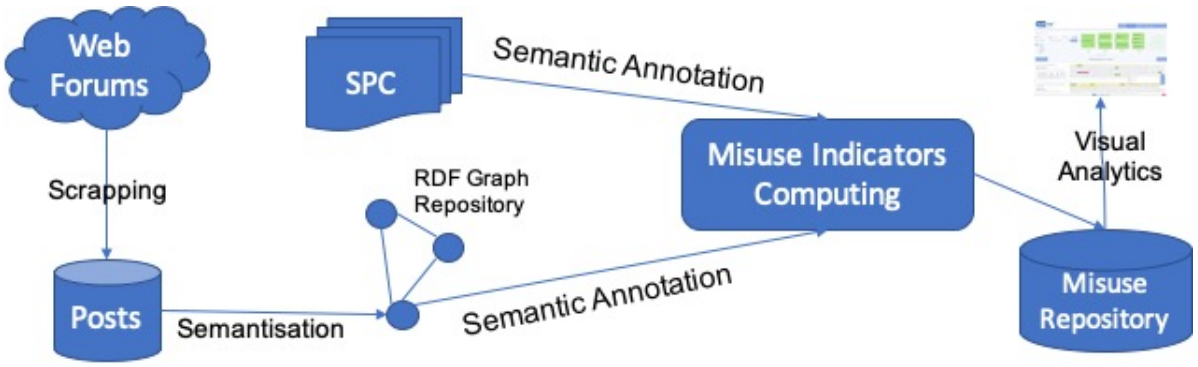

FIGURE 2.5 – Workflow of the Drugs Misuse Identification from Forum Posts.

Web Scrapping : extracting structured data from web forums. The scrapping strategy simulates the spontaneous behaviour of a user exploring a web forum similarly to (AUDEH et al., 2017). The first step is to extract the messages from the web pages, keeping the structure of the forum (categories, sub-categories, threads...). The publication date of each message and some information about the authors (year of birth and gender when available) are gathered.

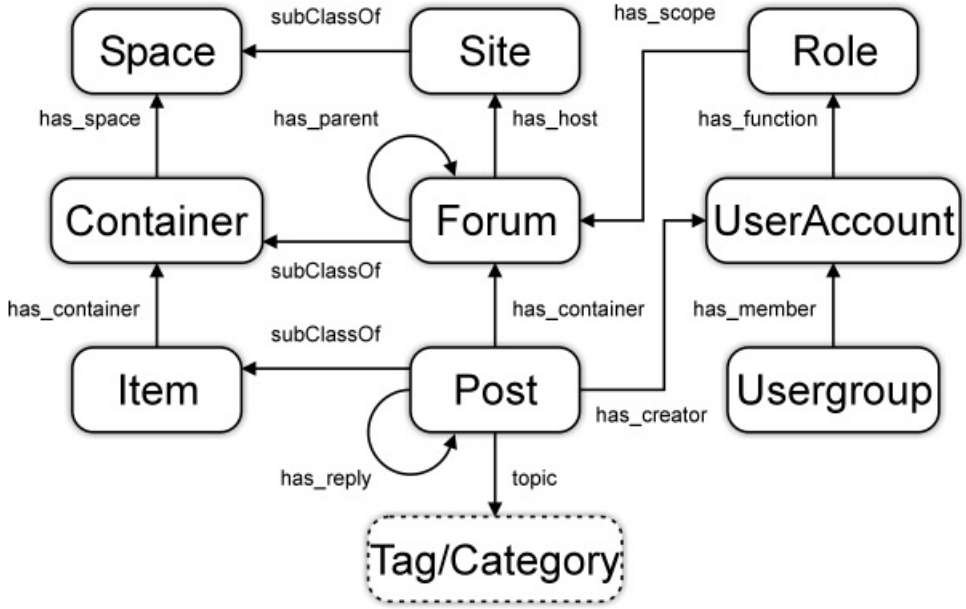

FIGURE 2.6 – Semantically-Interlinked Online Communities

Similarly to an indexing robot (crawler) like the robot of the Google search engine, Googlebot, we start from a root web page (in the case of healthcare related forum, the different categories) then we go through the hierarchy of the forum by following the links present on the pages until the leaf messages are reached. We use for that the jsoup library [8] which is Java HTML Parser which provides a very convenient API for extracting and manipulating data, using the best of DOM, CSS, and jquery-like methods.

8. https ://jsoup.org/

Semantic Graph based structuring of web forums data. In order to represent collected data from the actual web forums and allow them to interconnect easily, we need a flexible structure that can handle their heterogeneity. We have chosen an RDF graph based representation with the Semantically Interlinked Online Community (SIOC) Ontology (Breslin et al., 2006). The model is depicted in Figure 2.6. It offers the necessary entities to describe a web forum and its content : post, user account, content category, etc. The use of a semantically rich model allows integrating more easily data coming from various Web forums. The resulting data are stored in a graph based database, in our case the BlazeGraph triplestore [9].

Semantic Annotation of Web posts and Drugs related information. The purpose of a semantic annotation task of free text is the identification of key-phrases or terms in texts and linking them to concepts from a knowledge base in order to be qualified semantically. In our case, the forum posts are in French. Therefore dedicated NLP techniques were needed to process them. To do so, the approach used to process free texts in order to identify drugs or disease/symptoms mention in French texts relies on an early work of Khadim Dramé during his PhD thesis (Dramé, Diallo et al., 2014) and now adapted and improved by Sébastien Cossin as part of his PhD thesis (Cossin, Jouhet et al., 2018). A dictionary based approach is used, where entities to detect into the forum posts are represented into different dictionaries, the Anatomical Therapeutical Classification (ATC) for drugs and the ICD10 for diseases and symptoms. It takes as inputs a set of terms, normalizes them and stores them in a tree data structure for fast dictionary look-up. It handles abbreviations and a set of them were manually added for drug detection : “ac” for “acide”, “vit” for “vitamine”... The typo module for drug detection is a logistic regression trained on a manually created gold standard of 3,438 potential spelling errors of brand names and molecules (Cossin, Lebrun et al., 2019).

Computing misuse indicators. Two approaches have been designed to compute the possible misuses, a work done as part of the MSc internship of Louis Biliet.

— The first approach consists in the comparison of the list of indications, coded with the ICD10 vocabulary, which are associated with the drug listed in the official repository of the SPC forms with those collected from forum posts and which appear in the same sentence together with the drug of interest. This comparison uses a statistics-based approach and the traditional TF.IDF metric.

— The second approach relies solely on the forum posts without refereeing to the SPC forms and is based on the Vector Space Model (VSM)(Salton et al., 1975). To do so, the ATC is used as the vocabulary for drugs. For each drug, its list of associated indications (diseases and symptoms) is retrieved from the forum posts and constituted in a vector. From the ATC classification, the list of the siblings of the considered drug are retrieved as well and their associated indications from the forum posts constituted in a second vector. The cosine similarity of the two vectors is computed. As a remember the cosine similarity between two vectors is computed as follows :

$$sim(q,d) \;=\; v(\vec{q}).v(\vec{d}) \;=\; \frac{V(\vec{q}).V(\vec{d})}{\|V(\vec{q})\|.\|V(\vec{d})\|}$$

9. http ://www.blazegraph.com/

.

Visual Analytics and end users validation. The idea is to make available to pharmacovigilance experts a visual interface enabling the use of the platform in order to monitor drugs and validate identified misuses 2.7. The type of misuse identified for each drug referred in the forum posts and computed using the above mentioned algorithms could be explored using the hierarchical classification of the ATC.

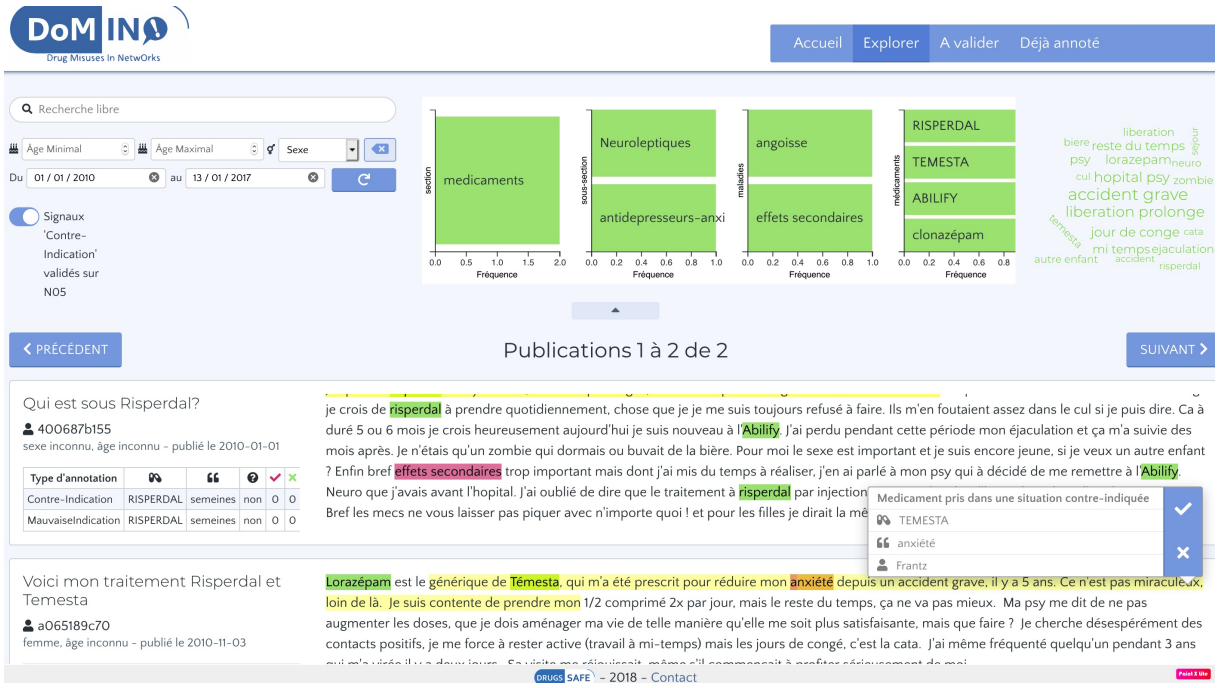

FIGURE 2.7 – GUI of DoMINO to explore and validate identified misuses

Main results and Implemented tool

The Web forum scrapping has been used on the French Doctissimo forum and collected 5,073,462 posts from 2010 to 2015. They are categorise into 364,619 discussion threads with 648,895 posts mentioning at least one drug. The forum posts have been RDFised and stored in the semantic repository, which accounts for 32,911,902 RDF triplets. The distribution of the most frequent drugs encountered in the posts is depicted in Figure 2.8, while the most frequent diseases and symptoms encoded with the ICD10 are depicted in Figure 2.9.

A web interface is provided which allows exploring the repository of the forum posts according to the possible identified misuses. The user has the possibility to confirm or to infirm a detected misuse indication 2.7 using the implemented metrics algorithm. A typical example is depicted on Figure 2.10.

Further information about the drugs identified within forum posts could be accessed thanks to a dedicated implemented Open Data source about French drugs (Romedi, figure 2.7 freely accessible online [10]), as part of the DoMINO project (COSSIN, LEBRUN et al., 2019). The set of detected misuses could be navigated hierarchically using the ATC classification (Figure 2.12).

10. http ://www.romedi.fr

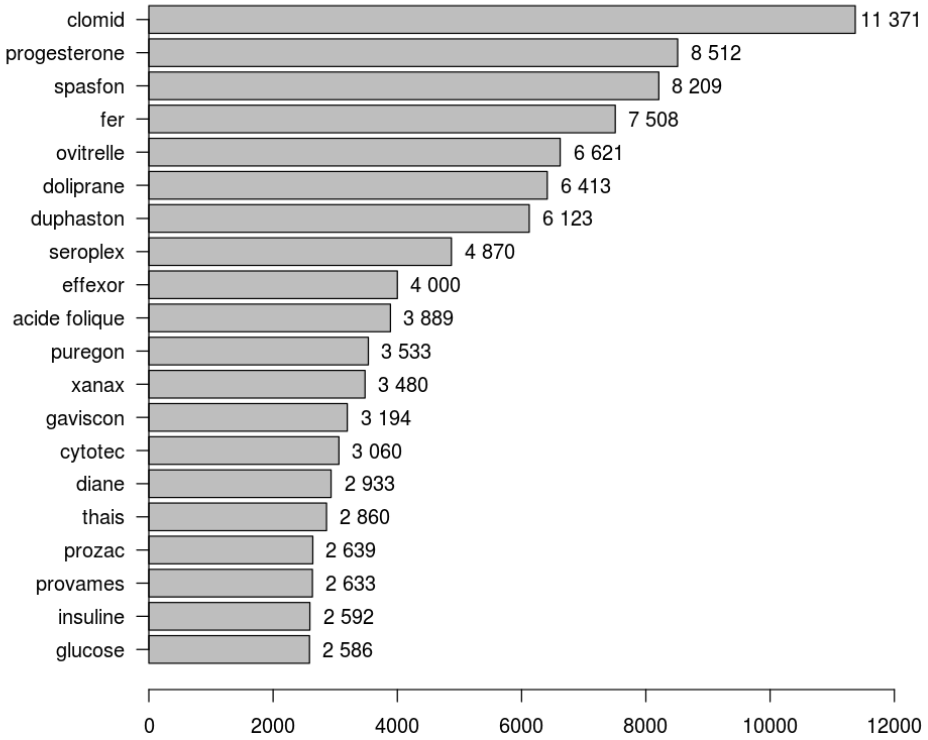

FIGURE 2.8 – Most frequent drugs encountered in the posts

The implemented tool relies currently on forum posts from a single health portal and dated between 2010 and 2015. As of next step, they will be updated and diversified. Furthermore, in the context of the funding project DoMINO/DurgSafe, the tool will be made available to different CRPV in France for domain experts usage testing.

2.4 Semantic Browsing of a Domain Specific Resources : the case of Infectious Disease

The work reported in this section has been conducted in the context of the EU-funded project SeaLife Semantic Web Browser for the Life Sciences during my post-doc at City University of London (UK). It has been realised with Patty Kostkova and Gemma Madle from the CeRC research group in collaboration with Khaled Khelif and Olivier Corby from INRIA Sophia Antipolis (France) and published in (DIALLO, KHELIF et al., 2008 ; OLIVER et al., 2009).

2.4.1 Introduction

One of the underlying reasons for the success of the Web is that it only uses one type of client, the Web browser, thereby providing a single access point to widely available content and services. However, it is becoming tougher to deal with the vast amount of data available in order to make efficient decisions.

Let's imagine a nurse consultant visiting the Internet and searching for information on healthcare associated infections. He/She must use specific search terms such as "MRSA" or

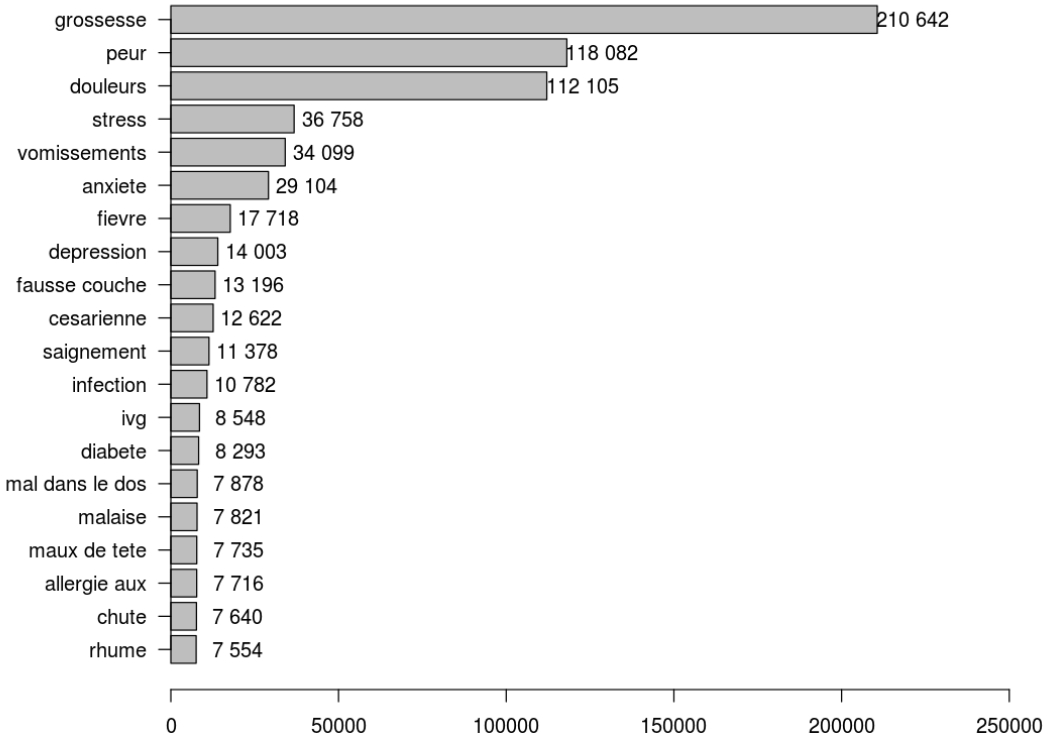

FIGURE 2.9 – Most frequent Diseases and Symptoms encountered in the posts

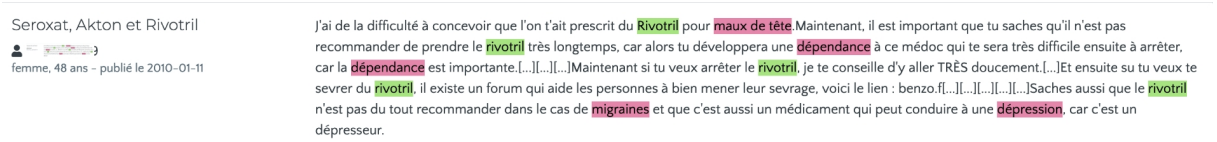

FIGURE 2.10 – Example of an annotated forum post with drug mention (Rivotril) and indications.

"Clostridium difficile", and the search engine will then return a list of documents indexed by these keywords. The user must manually filter through these documents to find the required information. The search engine will not be able to provide any indication of the quality or reliability of the documents retrieved, nor will it discriminate between the types of information, such as fact sheets, clinical guidelines, etc. It would be more useful to have a single interface to provide users with infectious diseases, their modes of transmission, treatment methods and also dynamic links to organisations such as the Health Protection Agency (HPA, www.hpa.org.uk) in the UK, which contains information on all aspects of infection and also offers authoritative guidelines.

We have designed and implemented the Corese-NeLI browser as part of the SeaLife project, which aimed to build a Semantic Web browser (SWB) for the Life Sciences (SCHROEDER et al., 2006). It is dedicated to the infectious disease domain and aims to improve the browsing experience of the user, with the ultimate goal of providing more relevant and supporting information in a timely manner. It has been designed on top of the National electronic Library of Infectious Diseases (NeLI) [11] portal with the NeLI On-

11. http ://www.neli.org.uk

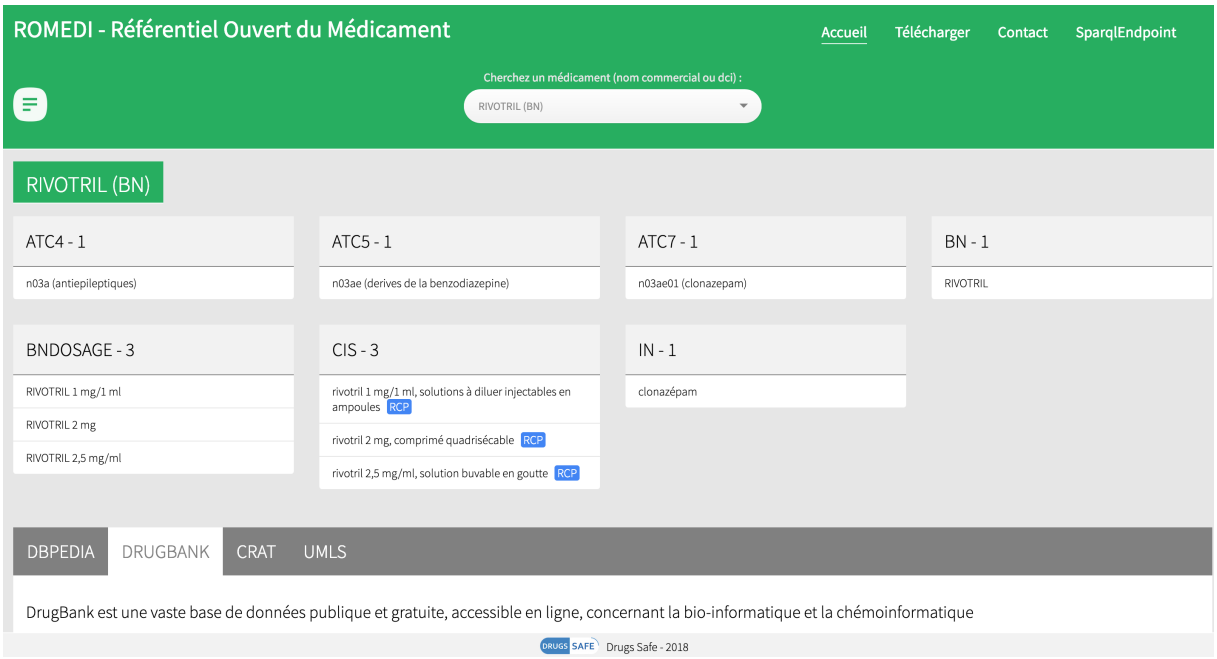

FIGURE 2.11 – Romedi, Référentiel Ouvert du Médicament with the information about Rivotril hilighted.

tology (DIALLO, KOSTKOVA et al., 2008) as a knowledge background. It follows the idea behind the Conceptual Open Hypermedia paradigm (GOBLE et al., 2001) which provides navigation between web resources, supported by a knowledge organisation system. It has been shown that vocabulary with weak semantics is sufficient for browsing according to (BECHHOFER et al., 2008).

2.4.2 Principle and Architecture

The architecture of the SWB is depicted in the Figure 2.13. It takes into account the variety of users by contextualising the information it provides. Context and customisation are some of the key factors for accurate, effective and relevant information access in Internet digital libraries and in the Semantic Web in general. Allan et al. (ALLAN et al., 2003) define contextual retrieval as a general framework combining search technologies and knowledge about a query and so called "user context" into a single framework in order to provide the most appropriate results for users' information needs. In our approach, the profile of the user is a record of user specific data that define users' weighted interests in a specific topic. The main components of the SWB are :

— **a queries generator :** this either uses keywords entered by the user or extracted automatically from the web pages. It generates several queries in different formats to query search engines (SPARQL queries in our case).
— **a profile manager :** it implements the approach proposed in (MRABET et al., 2007) to recommend pages according to the detected user's profile. It relies on semantic annotations and navigation logs for **learning model profiles** and classifying users.
— **a reasoning mechanism :** that consists of a set of rules allowing : (i) the building of a navigation scenario, and (ii) the use of advanced functionalities of the Corese search engine (CORBY et al., 2004) such as the approximate search.

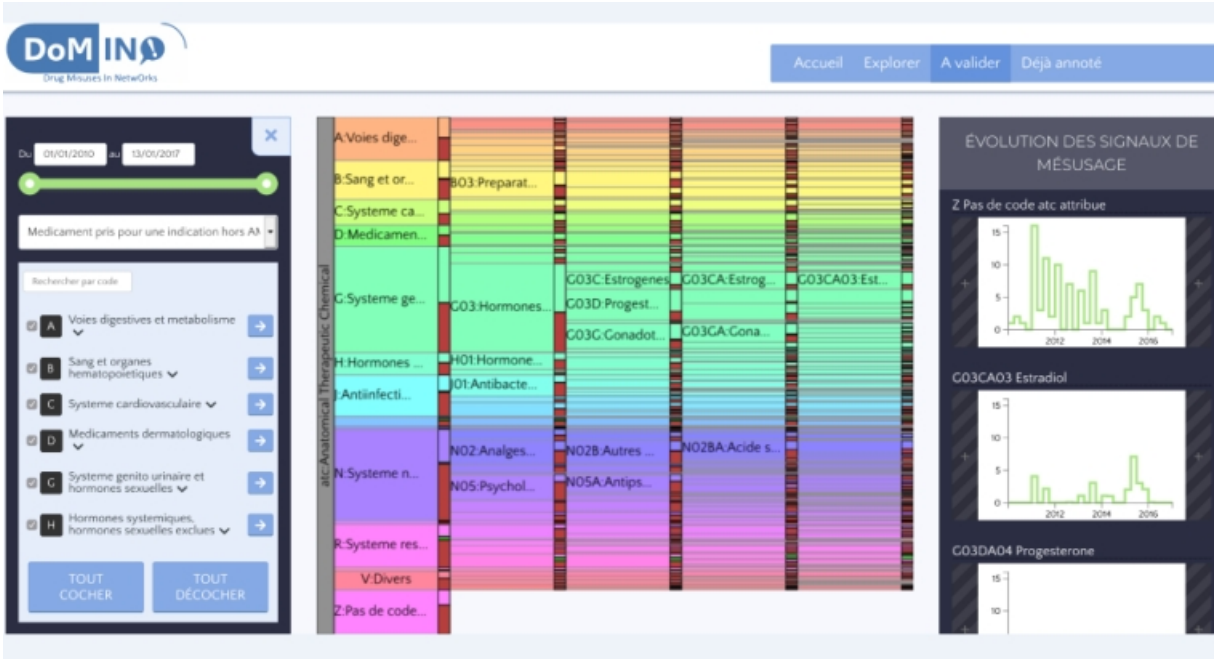

FIGURE 2.12 – Exploring visually the set of detected misuses from the forum posts

— **an annotation base :** which contains ontology-based annotations describing (i) knowledge embedded in web pages ; (ii) metadata on web pages including the creation date and the source ; (iii) learnt users' profiles, and (iv) information on external sources.

The SWB supports the navigation of a Web portal using a structured vocabulary or a domain Ontology with two main functionalities : i) semantic search of a Web portal relying on semantic annotations generated from Web pages using the background knowledge ; ii) semantic browsing of a Web portal : the SWB offers the possibility to identify and highlight terms retrieved from a structured vocabulary within a visited Web page. From the highlighted terms, it can then create dynamic semantic links to related pages within the portal, thereby enabling the semantic browsing and offer the possibility to query external resources such as the PubMed database [12].

Generating Semantic Annotations : the NeLIAnnotator

NeLIAnnotator is an adaptation of the MeatAnnot system (KHELIF et al., 2007). It starts from a page URL and relies on Corese (CORBY et al., 2004), the GATE text processing tool (CUNNINGHAM et al., 2001) and the NeLI vocabulary (DIALLO, KOSTKOVA et al., 2008) to generate an RDF annotation describing the semantic content embedded in the web page. To extract terms, NeLIAnnotator uses the Tokeniser module of GATE and the TreeTagger (MÀRQUEZ et RODRÍGUEZ, 1998). The tokeniser splits text into tokens, such as numbers, punctuation and words, and the TreeTagger assigns a grammatical category (noun, verb...) to each token. After tokenising and tagging texts, the NeLIAnnotator uses an extraction window of four (four successive words are considered as a candidate term). If a candidate term already exists in the NeLI vocabulary, the NeLIAnnotator processes the following word ; otherwise it decreases the size of the window till it reaches zero. The NeLI Ontology is queried using SPARQL in order to compare candidate terms

12. https ://www.ncbi.nlm.nih.gov/pubmed/

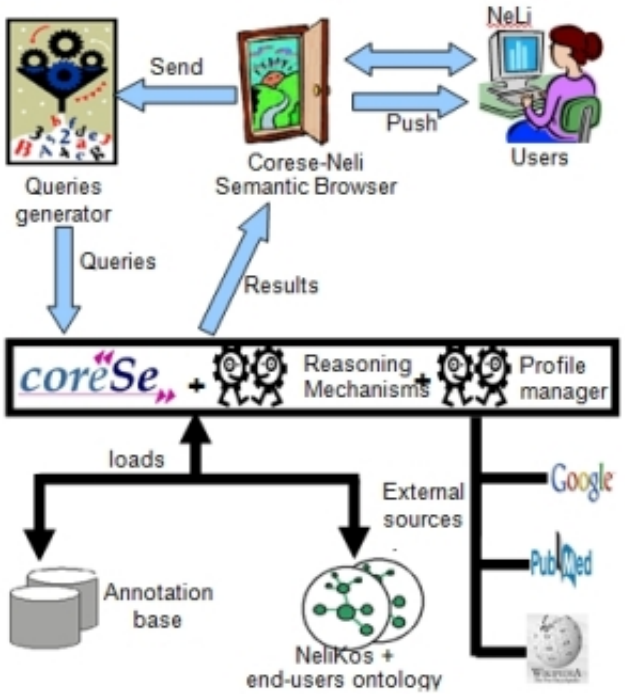

FIGURE 2.13 – Architecture of the Corese-NeLI Semantic Web browser

with entities of the Ontology.

Implemented browser

The Corese-NeLI SWB is provided as a firefox extension. Semantic browsing is enabled by navigating the graph of the NeLI vocabulary according to the user's actions. Figure 2.14 presents a screenshot of the interface for semantic browsing.

- Panel 1 : allows the user to query the NeLI portal using their own keywords. These keywords are mapped into the NeLI vocabulary concepts (w.r.t synonyms and linguistic variations). Suggestions of terms related to the user's entered keywords are also available (in this example the user is entering the term fever and the system proposes Dengue fever, Ebola Fever, etc.).
- Panel 2 : Gives a view on concepts related to the user query (in this example there are different kind of fevers and fever symptoms, etc.).
- Panel 3 : Stores the navigation history of the user in order to allow them to go back and find results from previous queries.
- Panel 4 : Gives the list of NeLI pages annotated by the chosen terms 2.15. This list can be refined by the source of the document (e.g. by international organisations or scientific journals) and by the date of publication. User can also have an idea about other terms annotating the selected page. By clicking a link the user is forwarded to the actual NeLI portal in order to have more information about the selected document.

Evaluation of the SWB

In order to measure the added value of the Corese-NeLI SWB, we have designed an end user centred evaluation framework (OLIVER et al., 2009). The study participants consisted of a set of traditional NeLI users separated in two groups ; those taking part

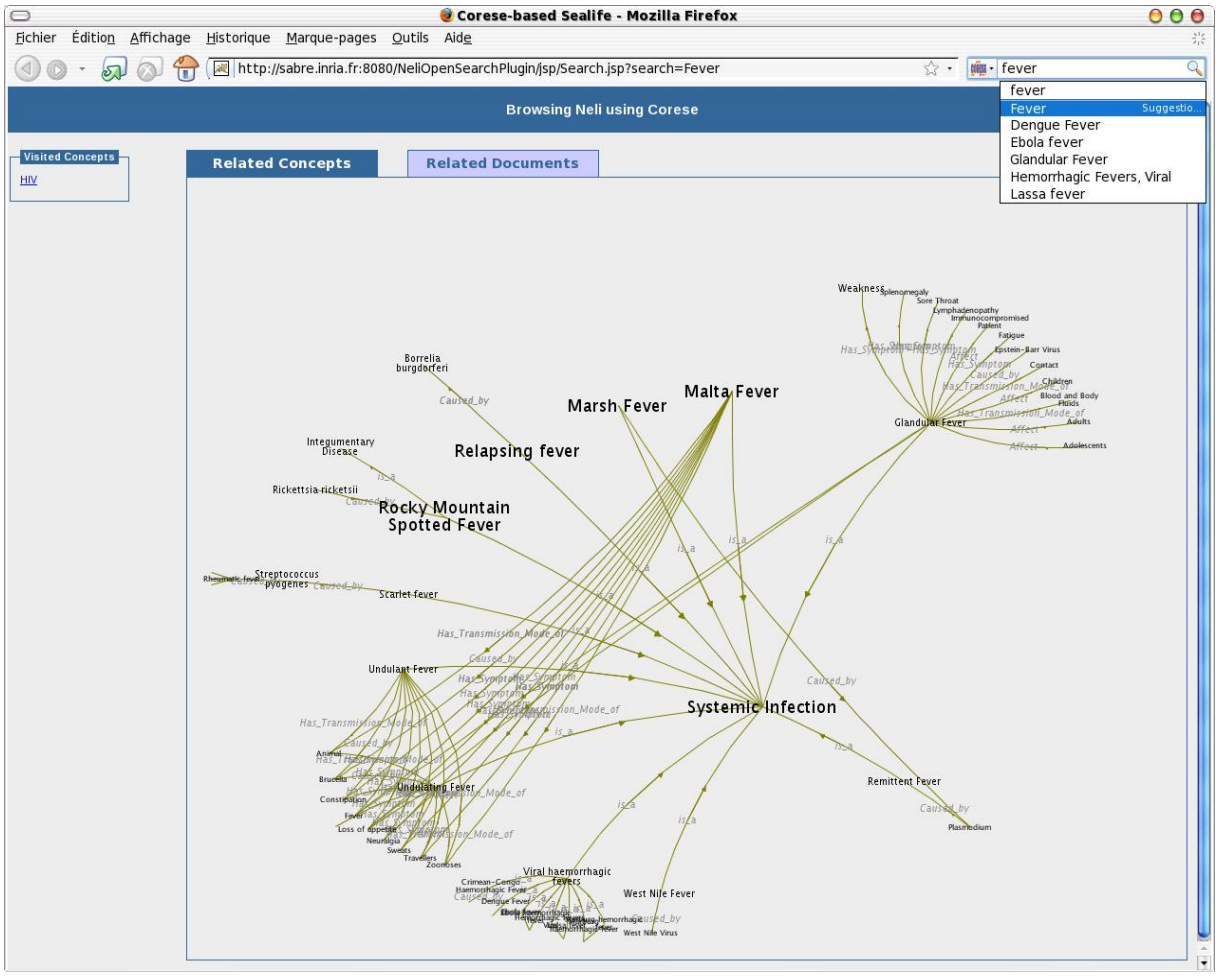

FIGURE 2.14 – Semantic browsing of the NeLI portal from the NeLI vocabulary

in a workshop-based evaluation and those taking part remotely online (for 3 months). In order to evaluate the Corese-NeLI SWB, the actual NeLI portal is used as the control system (baseline). Two settings were considered, an online setting with 4 participants and and workshop setting with 14 participants.

The following hypothesis was made to test the key purpose of the SWB.

— Mobility and travel within the system

H1 :The SWB reduces the time taken for users to find information or perform tasks.

H2 :The SWB shortens the pathway taken to find information or perform tasks.

H3 :Where semantic links are available, users will always follow them instead of non semantic links.

— User attitude and satisfaction

H4 :Users find the SWB easier to use than the control platform.

H5 :Where semantic links and ranking are available, users prefer them to non-semantic links and ranking.

H6 :Use of the SWB is intuitive : a) Users think the SWB helps them to find information or complete tasks. b) Users intuitively understand how to use the SWB to find such information or complete tasks.

Aim & objectives of the evaluation

FIGURE 2.15 – CORESE-NeLI Semantic Browser pane of related documents

— *Mobility & travel within the system :* the mobility and travel perspectives is assessed using the following objectives.
Objective 1 : time taken for users to find information or perform tasks
Objective 2 : pathway taken to find information or perform tasks
Objective 3 : use of semantic links compared with non-semantic links. Do users use semantic links ? Which semantic links are they using – hierarchical, semantic relationships, etc. What percentages of links are non-semantic and semantic ?
— *User attitude and satisfaction :* the user attitude and satisfaction will be assessed using the following objectives.
Objective 4 : user satisfaction with how easy the system is to use
Objective 5 : user attitudes to the availability of semantic links or ranking of results.
Objective 6 : user understanding of the semantic browser. This issue is subdivided into two questions ; does the user think that the SWB helps him find information or complete tasks ? Does the user understand how to use the semantic browser to find such information or complete such tasks ?

Objectives 1 (O_1), 2 (O_2), and 3 (O_3) evaluate the ease of travel of a user's journey through the system. Objectives 4 (O_4) and 5 (O_5) examine whether the reduction or increase in time taken and change (if any) in pathway is a positive or negative experience for the user. Objective 6 (O_6) examines their understanding of how the semantic browser works, i.e. the provision of material they may not have found otherwise, or known was relevant to their needs. Overall, these objectives evaluate the effectiveness of, and user satisfaction with the semantic browser, compared with the NeLI portal alone, thereby demonstrating the value of such a tool.

Table 2.5 give the confirmation or contradiction of original hypotheses concerning the evaluation of the Corese-NeLI SWB.

As of illustration for the evaluation of the objectives, for O_1, the time taken per task was calculated from the online evaluation logs using the difference between task page and question page loading times. It took in average 266 seconds to complete the tasks using the SWB against 387 seconds for the classical NeLI portal.

The detailed evaluation of the Corese-NeLi browser as well as the other implemented

Hypothese	Corese-NeLI SWB
H_1 :The SWB reduces the time taken for users to find information or perform tasks	Yes
H_2 :The SWB shortens the pathway taken to find information or perform tasks	No (target found by few users)
H_3 : Where semantic links are available, users will always follow them instead of non semantic links	No
H_4 :Users find the SWB easier to use than the control platform	No
H_5 :Where semantic links and ranking are available, users prefer them to non-semantic links and ranking	Yes
H6 :Use of the SWB is intuitive - a)Users think the SWB helps them to find information or complete tasks	No
H_6 :Use of the SWB is intuitive - b)Users intuitively understand how to use the SWB to find such information or complete tasks	No

TABLE 2.5 – Confirmation or contradiction of original hypotheses

SWB browsers in the context of SeaLife is reported in (OLIVER et al., 2009).

2.5 Conclusion

We have described in this chapter the work conducted in the context of large scale biomedical literature mining and semantic information retrieval.

We have followed a K-NN based approach, $kNN\_Classif$, for the multi-label classification of a large corpus of biomedical scientific articles (several millions) when only a partial information is available, a quite challenging task. The approach proven to be able to predict relevant labels with results close to similar approaches for such a task when a larger learning set is used.

I addition to the $kNN\_Classif$ algorithm that has been proposed, we have also used and Explicit Semantic Analysis approach for the same task. However, our findings suggest that this approach as exploited in our context has proved less conducive to multi-label annotation of a large corpus of textual documents when only partial information is available.

From the point of view of Web forum posts analysis for drugs misuse detection, the approach followed, which is more generic than those available in the literature, and implemented with a visualisation tool, made it possible to detect situations of inappropriate drugs usage. Further evaluation which necessitates more posts from additional web forums is needed, which is planed in the context of the Drug-Safe project.

The last part focused on semantic browsing of a domain specific resources, the infectious disease domain with the National electronic Library of Infection in the UK. The implemented approach includes a background knowledge based annotation of web pages textual content, which enables semantic links for navigation. The evaluation conducted in an online and workshops based settings showed that the semantic browser reduces time taken for users to find information or perform tasks, which is relevant for information retrieval. However using a semantic browser has proven to be less intuitive.

Chapitre 3

Knowledge Warehousing and Large Scale Ontology Matching

Contents

The first part of this chapter is dedicated to the ServOMap large scale Ontology matching system. It benefited with input from several MSc students and collaborations with colleagues from University of Bobo Dioulasso for User Involvement in Ontology matching and the University of Tunis El Manar regarding the multilingualism issue. The second part described in section (3.2) is part of the K-Ware Knowledge Warehouse framework in the context of the *PhD thesis of Bruno Thiao Layel* (mid 2018 - mid 2021) that I am co-supervising jointly with Guillaume Blin (LaBRI) and Vianney Jouhet (ERIAS and CHU Bordeaux). This work is a natural continuation of the work on ServoMap and is carried out under a French ANRT-CIFRE support in collaboration with the Synapse Medicine start-up [1], a spin-off of the BPH/ERIAS research group.

3.1 Large Scale Matching of Ontology with ServOMap

3.1.1 Introduction

Ontology matching has been an active research area especially since the advance of the Semantic Web (S. T. Berners-Lee, 2011). It is part of the more general topic of schema matching (Shvaiko et Euzenat, 2013), which aims at finding correspondences between entities in different conceptual schemas called *alignments*. Various methods and strategies are used to perform a matching task : from terminological and structural to semantic based computation of candidate mappings (please see ((Kalfoglou et Schorlemmer, 2003 ; Otero-Cerdeira et al., 2015 ; Shvaiko et Euzenat, 2013) for a complete survey). Despite the advances achieved in matching relatively of small size ontologies, the large

1. https ://synapse-medicine.com/

scale matching problem still presents thriving challenges to tackle, due to the complexity of such a task. Some of these challenges have been recalled in Section 1.4.2 : search space optimisation, computation times, users involvement and alignment validation.

In order to handle such a task, various approaches for Ontology matching have been proposed in the literature. They include clustering and blocking strategies for the reduction of the search space (Algergawy, Massmann et al., 2011 ; Aumueller et al., 2005 ; Hamdi et al., 2010 ; Hu et al., 2006 ; Lambrix et Q. Liu, 2009). These systems mainly rely on a partition-based block approach which consists in splitting the ontologies to be aligned into smaller parts. In contrast to these divide-and-conquer methods, Wang et al. (P. Wang et al., 2011) use two kinds of reduction anchors to match large ontologies and reduce time complexity. In the context of biomedical ontologies matching, some approaches rely on the use of one or several specialised background knowledge, such as the UMLS (Bodenreider, 2004). Thus, XMap is a scalable matcher that follows parallel processing strategy to enable the composition of basic ontology matchers (Djeddi et al., 2017). It uses synonyms provided by the UMLS Meta-thesaurus. Whereas, the AML system relies on three sources of background knowledge which can be used as pivots between the input ontologies (Faria, C. Martins et al., 2015 ; Faria, Pesquita et al., 2013) : the Uber Anatomy Ontology (Uberon) (Mungall et al., 2012), the Human Disease Ontology (DOID) (Kibbe et al., 2015) and the Medical Subject Headings (MeSH) (Lowe et Barnett, 1994). The LogMap system similarly uses the UMLS Lexicon for normalisation and handling spelling variants (Jiminez-Ruiz et al., 2011).

In this context, it is worth mentioning that ServOMap does not use any biomedical related background knowledge, which could limit the ability of matching system to some particular contexts. It does not rely on a partition based techniques nor on a partitioning algorithm that could impact the quality of the alignment. Instead, ServOMap relies on an indexing strategy for handling large biomedical ontologies by considering each input Ontology as a corpus of semantic documents. Before to detail the approach, let us give some preliminary definitions of the notions that will be of use in the remainder of the chapter.

Preliminary Definitions

Definition 1 (Ontology) : an ontology is a 5-tuple $O =< C^o, R, H_r, T, Lex >$ and $R = R_I \cup R_T \cup R_D$ such that :

- C^o is a set of concepts ;
- $R_I \subset C^o x C^o$ is a concepts taxonomy and $h = (c_1, c_2) \in R_1$ means C_1 is a subsumer of C_2, the *is-a* relation ;
- $R_T \subset C^o x C^o x L_T$ is a set of transversal relationships where L_T is a set of relations labels ;
- $R_D \subset C^o x C^o x L_D$ is a set of attributes where P is a set of xml primitive data types and L_D is a set of relations labels ;
- $H_r \subset RxR$ is a taxonomy of relationships on R_T and R_D ;
- T is a set of (multilingual) strings terms that are concept labels (synonym terms) ;
- $Lex : T \rightarrow C^o$ is a function which associates concepts with their labels.

R_T and R_D are, respectively, object and dataype properties in the sense of the RDF Semantic Web languages (Lassila et al., 1998). Constraints can be associated with these properties, for instance the notion of *functional property.*

Definition 2 (direct descendant concepts Sub_c^o) : Given an Ontology O and a concept c, the direct descendant concepts of c within O denoted Sub_c^o is the set :
$\{c_i \in C^o | c_i \neq c, c_i R_I c \; and \; (\nexists \in C^o, cR_I c_k \; and \; c_k R_I c_i\}$

Definition 3 (direct ancestor concepts Sup_c^o) : Given an Ontology O and a concept c, the direct ancestor concepts of c within O denoted Sup_c^o is the set :
$\{c_i \in C^o | c_i \neq c, c_i R_I c \; and \; (\nexists c_k \in C^o, c_i R_I c_k \; and \; c_k R_I c\}$

Definition 4 (sibling concepts Sib_c^o) : Given an Ontology O and a concept c, the sibling concepts of c within O denoted Sib_c^o is the set :
$\{c_i \in C^o | \exists y \in C^o, y \in Sup_c i^o and y \in Sup_c^o\}$
Now let us define the notion of *virtual document* that is subdivided into *direct virtual* and *extended virtual document*. The notion of virtual documents from the RDF graph has been previously used in (B. Chen et al., 2006)

Definition 5 (direct virtual document) : Given an ontology O and an entity e, a *direct virtual document* or $dVD(e)$ is constituted by the combination of its uniform resources identifier (uri), obtained by $ID(e)$, the local name ($locName(e)$), the labels in different languages extracted by the inverse of the function $Lex(Lex^-1)$ and the set of annotations associated with it. Formally,
$dVD(e) =< ID(e), locName(e), Lex^-1(e), R_T(e), R_D(e) >$, if e is a concept ;
$dVD(e) =< ID(e), locName(e), Lex^-1(Dom(e)), Lex^-1(Range(e)), const(e) >$, if e is a property.
where $R_T(e), R_D(e)$ stand for the properties attached to e and their information. $Dom(e)$ and $Range(e)$ give respectively the list of domains and ranges of the property e and $const(e)$ gives the property constraints associated to e (e.g. *functional property*).
From the *direct virtual document*, we can define the notion of *extended virtual document*, which represents its virtual transitive closure.

Definition 6 (extended virtual document) : Given an ontology O and a concept $c \in O$, let's assume that $Sup_{Lex}(c) = \{t \in T \; \exists c_i \in C^o, c_i R_I c \; and \; t \in Lex^-1(c_i)\}$ denotes the terms (labels) associated with the ancestors of c and $Sup_{LocName}(c) = \{ln | \; \exists c_i \in C^o, c_i R_i c \; and \; ln \in locName(c_i)\}$. Formally,
$eVD(c) =< ID(e), locNamed(e), Lex^-1(e), R_T(e), R_D(e), Sup_{Lex}(c), SupLocName(c) >$

3.1.2 Detail of the Matching Approach

The matching process in ServOMap relies on Information Retrieval strategy thanks to the use of the ServO Ontology server. The latter enables ontological entities indexing and lookup by instantiating the metamodel it provides (Diallo, 2011) and of which an extract is depicted in Figure 3.2. The metamodel enables the representation of an OWL-based KOS. For the matching itself, three main steps are used : (*i*) the initialisation phase ; (*ii*) the candidate mappings retrieval ; and (*iii*)the post-processing phases as depicted in Figure 3.3. Theses phases are described in the following.

Initialisation phase

The initialisation phase comprises the following steps :

Loading of the ontologies to match The ontology loading step takes charge of processing the input ontologies. For each entity (concept, property), a *direct virtual document* from the set of annotations is generated for indexing purposes. Any ontology, regardless of its degree of formalism, is considered as a corpus of semantic documents to process following **Definition 1**. Each entity (concepts or properties including both object properties and data type properties) is therefore a document to process. For each input Ontology, ServO is used to dynamically generate a *direct virtual document* which corresponds to any retrieved entity and instantiate the ServO metamodel (Diallo, 2011). The objective is to gather the terminological description of each entity in order to build a vector of entity terms. Each *virtual document* has a set of fields for storing its different elements. The generation process takes into account the terminological units in different languages used to described each entity.

In order to optimise entities lookup, we have implemented during the loading step a bit-vector representation of the taxonomical structure of each input ontology, inspired by the approach in (Ait-Kaci et Amir, 2017). As depicted in Figure 3.1, in this bit-vector representation, a bit is set to 1 if : *i*) it has a rank equal to the index of the concept ; *ii*) the bit that has a rank equal to the index of a sub-concept of this concept. For instance, *professor* has three bits set to 1 which is the representation of the number 3. The concepts *full professor* and *associate professor* are represented by encoding the number 1 and 2 respectively.

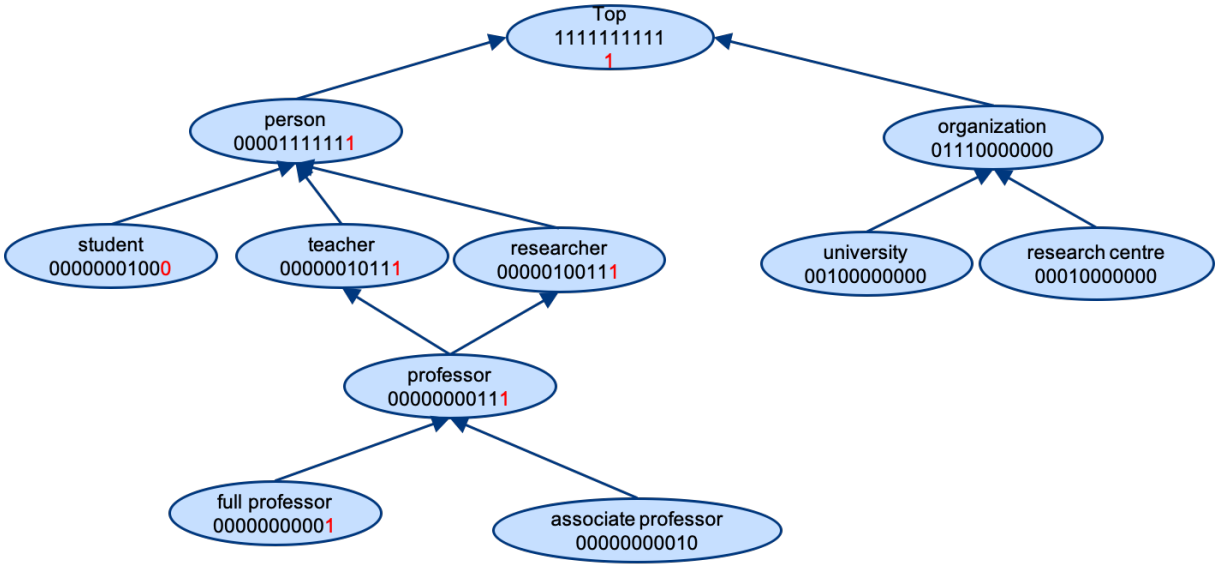

Figure 3.1 – Example of a bit-score encoding of the taxonomy of a KOS.

According to the evaluation we performed, the bit-vector representation enables in particular a fast look up for super-concepts and sub-concepts. For instance, for an Ontology with about 29,000 concepts, retrieving sub-concepts from a given concept takes 134 μs against 133,841 μs.

Metadata and metrics generation After loading the ontologies, a set of metrics are computed. They include the size of input ontologies in terms of concepts, properties and instances, as well as the list of languages denoting the annotations of entities (labels, comments), etc. Identifying the input size helps in later adapting the matching strategy. Two categories are distinguished according to the size of the considered ontologies : matching

two ontologies with number of concepts ≤ 500 and matching ontologies with a number of concepts ≥ 500. Further, we pre-identify whether the matching problem is a conceptual level or instances level matching. The purpose of detecting the set of languages is to allow the use of the latter as the appropriate list of specific stopwords and the use of stemming during pre-processing.

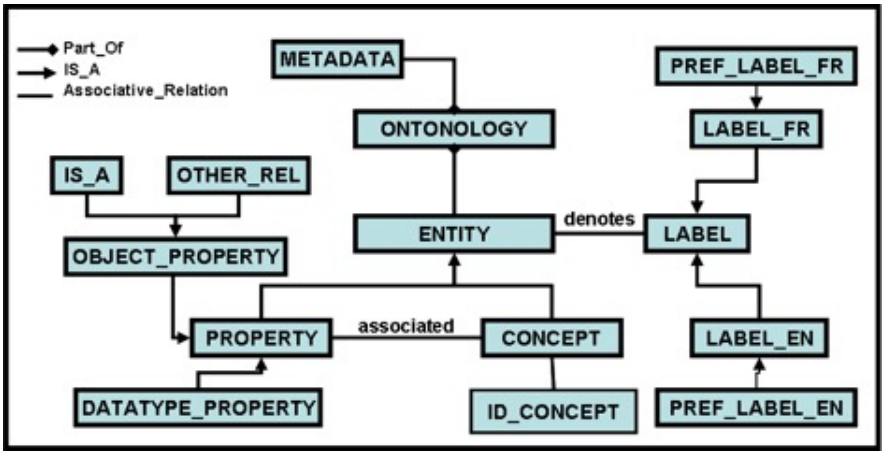

FIGURE 3.2 – Excerpt of the meta-model of ServO

Ontology indexing ServOMap has been the approach which introduced Ontology indexing as a strategy to reduce the search space for large ontologies alignment (DIALLO, 2011). As in the traditional IR domain, the purpose of the indexing step is to build an inverted index for each input ontology from the *virtual documents previously* generated. Each index contains a set of fields identified during the generation of virtual documents. Detailed examples of an index content could be found in (DIALLO, 2014).

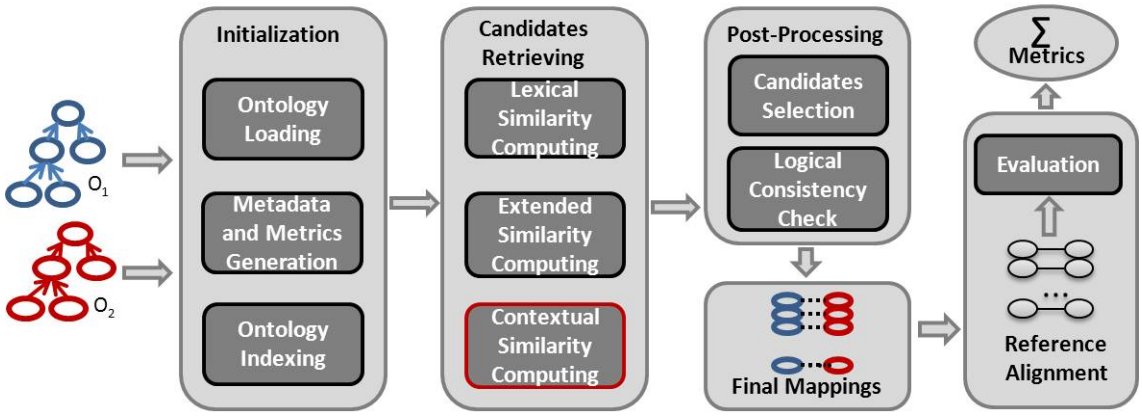

FIGURE 3.3 – Matching process with ServOMap

We proceed as follows to build the index. Each *dVD* is passed through a set of filters : stopwords removal, non-alphanumeric character removal (for harmonisation of each entity's terms description), lower-casing and label stemming and finally converting numbers to characters. Indeed, we use a VSM-like system from the IR field (BAEZA-YATES et RIBEIRO-NETO, 1999) to compare terms. Therefore, non-alphanumeric symbols are removed in order to harmonise the description of terms. The conversion of numbers to

characters contributes too in reducing errors during the matching process. For instance, for the two biomedical ontologies FMA and NCI, the Ninth_thoracic_vertebra of the FMA corresponds to the T9_Vertebra of the NCI thesaurus (knowing that the latter also has T8_Vertebra and so on).

In addition, to deal with the variation in naming concepts with natural language units, labels denoting concepts are enriched by their permutation before stemming. Permutation, in our context, is useful to acquire synonymous terms in the context of biomedical ontologies. It is common to come across concepts denoted by all possible permutations between words of the terms. For instance, the concept *Myocardial Infarction* with the UMLS concept with CUI C0027051 has among its labels : myocardial infarction, heart attack, infractions myocardial, attacks heart, etc. The permutation operation is applied to the first four words of each label. We limit it to the first four words for two reasons : *i*) most terms used to denote entities that we encounter are of less than five words ; *ii*) increasing this threshold affects the performance in terms of computation time without a significant gain in terms of F-measure. For instance, after enriching the term *thoracic vertebral foramen* we obtain, before stemming, the set *thoracic vertebral foramen, thoracic foramen vertebral, vertebral thoracic foramen, foramen vertebral thoracic, foramen thoracic vertebral, vertebral foramen thoracic*.

The permutation process slightly increases the size of the index (around 20%). For instance, the index of SNOMED-CT is 22.8 Megabyte with permutation against 18 Megabyte. We can notice that the increase of the size of index generated by the IR API that we use [2] is not linear according to the input data due to internal compression strategy. In addition, our experiments show that the impact of the permutation on the precision is negligible compared to the gain in terms of F-measure. It is of the order of 2%. This low impact could be explained by the combined use of the concatenation.

The indexing strategy involves the possibility for each concept's *dVD* to take into account its surrounding contest for the optimisation process : siblings, descendants and ancestors concepts.

Computing Candidate Mappings

The candidate mappings computing relies on three categories of similarity : an *element level similarity* which comprises a terminological and an extended terminological strategy and a *structural similarity* which is based on the contextual location of the entities within the ontologies being matched.

Given two inputs ontologies O_1 and O_2, their respective indexes, I_1 and I_2, and an optional background knowledge (BK), then the candidate mappings, denoted as $M_{candidate}$, is the union of the candidates generated using respectively terminological (denoted as M_{term} for concepts, M_{prop} for properties), extended (denoted as M_{ext}^{BK}) and contextual (denoted as $Mcontext$) similarity computing :

$$M_{candidate} = M_{term} \cup M_{prop} \cup M_{ext}^{BK} \cup M_{context}$$

Terminological similarity The terminological similarity is an element level similarity computed by comparing the vector representation of entities of the two ontologies which are being aligned (Diallo, 2014). Two successive operations are performed. First, each entity of the input ontologies, represented as a vector according to the *virtual document*

2. www.apache.lucene.org

representation described previously, are compared. This strategy is inspired by the VSM of IR (Baeza-Yates et Ribeiro-Neto, 1999) where for any two entities e_1 and e_2 to compare, e_1 plays the role of the query and e_2 the document to compare with. The resulting candidates are further filtered using a combination of various lexical similarities. Therefore, let's assume that :

— $ISub(s_1, s_2)$ is the ISub string similarity between two input strings, a measure adapted for ontology matching (Stoilos et al., 2005).

— $Q-Gram(s_1, s_2)$ is the n-gram similarity distance between two texts string, which is simply the number of common/distinct n-grams between two strings (Ukkonen, 1992).

— $Lev(s_1, s_2)$ is the Levenshtein distance between two strings, which is the minimum number of single-character edits (insertion, deletion, substitution) required to change one word into another (Levenshtein, 1966).

Given two entities (concepts or properties) e_1 and e_2 such that $e_1 \in O_1$ and $e_2 \in O_2$, the terminological similarity Sim_{term} is defined as :

$$\begin{aligned} Sim_{term} = (\alpha \cdot ISub(eVD(e_1), eVD(e_2))) \\ + (\beta \cdot QGram(eVD(e_1), eVD(e_2))) \\ +(\gamma \cdot Lev(eVD(e_1), eVD(e_2))) \end{aligned}$$

where $\alpha, \beta, \gamma \in [0, 1]$ are respectively the weight for the $ISub$, $Q-Gram$ and $Levenshtein$ distances $\alpha + \beta + \gamma = 1$).

Sim_{term} is computed for all the set of candidate couples retrieved using the IR API used to index the input ontologies and whose cosine similarity is greater than a given *MaxScore* chosen manually. Thereafter, the obtained pairs are filtered out in order to keep only those satisfying a terminological similarity condition. This condition is to keep all pairs $< e_1, e_2 >$ such that $Sim_lex(e_1, e_2) \geq \theta$. The threshold θ is chosen empirically.

Extended similarity The main purpose of the extended similarity is to deal with synonym issue in naming entities and is limited to the concepts denotation. We have implemented a general purpose background knowledge-based strategy. From the set of concepts not selected after the previous phase, we use the WordNet dictionary (Miller, 1995) for retrieval of alternative labels for concepts to be mapped. The idea is to check whether a concept in the first ontology is denoted by synonymous terms in the second one. All pairs in this case are retrieved as possible candidates.

Contextual Similarity The main idea is to compare the entities being matched according to their **structural representation**. That means the position of each entity's node within the corresponding Ontology hierarchy is taken into account. The algorithm starts with the candidate set generated by the terminological similarity. A Machine Learning strategy is then applied in order to learn new possible candidates from the one already generated. Let us assume that O_1 and O_2 are the two ontologies to match and that a_6 and b_6 are candidates mapping already identified in the previous steps, and supposed to be correct. Figure 3.4 depicts the strategy that is followed. The idea consists into transforming the matching into a ML problem. To do so, we start from the set of candidate couples identified previously, including a_6 and b_6, and which constitute the positive examples of the learning set *LearnSet*. Then, negative examples are included by looking up the surrounding of the positive candidate couples, in our example in Figure 3.4 for instance, a_2

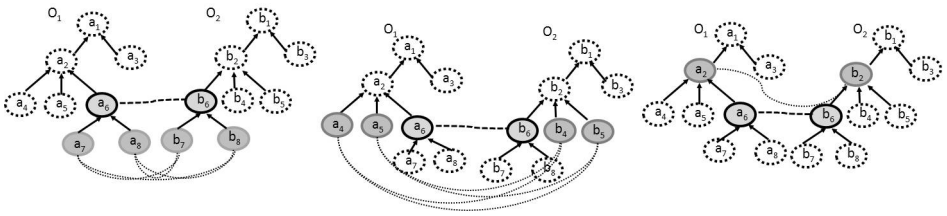

FIGURE 3.4 – Contextual similarity strategy

and b_7 or a_7 and b_2. These couples cannot be positive with our hypothesis that a_6 and b_6 is as well correct. The overall strategy is that for each couple $(c_1, c_2) \in M_{term}$ we obtain $\{(c_1, Sub^O_{c2}), (c_1, Sup^O_{c2}), (Sub^O_{c1}, c_2), (Sup^O_{c1}, c_2)\}$, etc.

For each pair in *LearnSet*, we compute a set features constituted of five similarity measures between the eVD (which does not include in this case the properties) : Q-Gram, Levenshtein, BlockDistance, Jaccard and Monge–Elkam.

— The block distance between two vectors a and b is $\sum_{i=1}^{n} |a_i - b_i|$.

— The Jaccard measure (JACCARD, 1912) computes the number of words two strings have in common divided by the total number of unique words.

— The Monge–Elkam measure (MONGE et ELKAN, 1996) is a simple but effective method of measuring similarity between two text strings containing several tokens, using an internal similarity function, $sim(a, b)$, able to measure the similarity between two individual tokens.

Post-processing Phase

In the post-processing step, the enrichment of the set of candidates mapping is performed. In addition, the selection of the best candidates by performing a logical consistency check is done, and finally, if a reference alignment is available, performing the system performance evaluation. The enrichment consists in incorporating those identified, not originally mapped pairs and mapping all of their sub-concepts after the similarity computing phase.

The selection of the final candidates from the set $Mcandidate$ is performed using a filtering algorithm, which implements the **Stable Marriage Problem principle** (MCVITIE et WILSON, 1971) and a **greedy selection strategy** (DIALLO, 2014) to select the best candidates based on their scores.

Performance of the Matching System

The ServOMap system and its different derivative versions over the years have been evaluated in the context of several OAEI campaigns between 2012 and 2015. It has been ranked regularly among the top three performing systems for large ontologies, the LargeBiomed dataset of the OAEI campaign introduced in 2012. LargeBiomed has been for several years the most challenging task in terms of scalability and complexity. It is dedicated to the evaluation of automated large scale (biomedical) matching systems. The ontologies in this dataset are semantically rich and contain tens of thousands of entities. The track consists of finding alignments between the FMA containing more than 78,000 concepts (ROSSE et MEJINO, 2003), the SNOMED-CT containing more than 306,000 concepts (LEE et al., 2014 ; SCHULZ, CORNET et al., 2011) and the NCI Thesaurus (NCI) containing more than 66,000 concepts (GOLBECK et al., 2003).

Task	#Mappings	Precision	Recall	F-score
FMA-NCI Small	2,725	0.943	0.85	0.894
FMA-NCI Large	3,163	0.711	0.836	0.793
FMA-SNOMED Small	6,978	0.955	0.74	0.834
FMA-SNOMED Large	7,940	0.833	0.735	0.781
NCI-SNOMED Small	12,716	0.933	0.642	0.761
NCI-SNOMED Large	14,312	0.822	0.637	0.718

TABLE 3.1 – Results of the ServOMap matching system

In order to measure the behaviour of the matching system according to the size of the input ontologies three matching problems are identified : the FMA-NCI matching problem, the FMA-SNOMED CT matching problem, and the SNOMED CT-NCI matching problem, with each problem divided into two subtasks : *small* and *large*. Six subtasks according to the size of the fragments of the input ontologies are considered. Therefore, for the FMA-NCI problem, the small task consists of matching 5% of the FMA (3,696 concepts) and 10% of the NCI (6,488 concepts) while the large task consists of matching the whole ontologies. For the FMA-SNOMED CT problem, the small task consists of matching 13% of the FMA (10,157 concepts) and 5% of SNOMED CT (13,412 concepts). The large task consists of the complete FMA and 40% of SNOMED CT (122,464 concepts). For the SNOMED CT-NCI problem, the small fragment consists of 17% of SNOMED CT (51,128 concepts) and 36% of the NCI Thesaurus (23,958 concepts) while the large task consists of the complete NCI Thesaurus and 40% of SNOMED CT.

UMLS Metathesaurus is used as the gold standard for the track reference alignments (BODENREIDER, 2004). The Metathesaurus is a very large, multi-purpose, and multilingual vocabulary database that contains information about biomedical and health related concepts, their various names and the relationships among them. It is worth noting that as UMLS may contain some incoherences and is not complete, the performance of an automated matching system could be affected when using a reference alignment from this resource. To overcome this limitation, organisers of OAEI provide more cleaner references alignment using either voting or repair (PESQUITA et al., 2013) strategies.

Table 3.1 summarises the best performance achieved by the ServOMap system on the LargeBio dataset over the different editions.

For the 2012 edition of OAEI ServOMap achieved in average the second best performing system in terms of F-score (0.780), behind the YAM++ system (0.782) (NGO et BELLAHSENE, 2012) and second computation time after a lightweight version of the LogMap system (JIMÉNEZ-RUIZ, GRAU et al., 2012) which only relies on string comparison between entities.

More detailed results could be found in (AGUIRRE et al., 2012 ; CUENCA GRAU et al., 2013 ; DIALLO, 2014)

Let us now report on a possible solution to overcome the limitation of fully automated Ontology alignment, as described previously.

3.1.3 User Involvement in Ontology Matching

The contributions described in this section have been published in the following papers :

M. Savadogo, B. M. J. Some, G. Diallo : Alignement interactif d'ontologies avec ServOMap. Technique et Science Informatiques 35(1) : 55-74 (2016)

.

N. Kheder, G. Diallo. ServOMBI at OAEI 2015. Proceedings of the 10th International Workshop on Ontology Matching collocated with the 14th International Semantic Web Conference (ISWC 2015), Bethlehem, PA, USA, October 12, 2015. CEUR Workshop Proceedings 1545. pp 200-207

.

M. Savadogo, BMJ. Some, G. Diallo. Alignement interactif d'ontologies avec ServOMap. Actes des 5èmes Journées Francophones sur les Ontologies (JFO), 14-16 Nov. 2014, Hammamet, Tunisie.

Interactive alignment has emerged after the observation that research works on Ontology alignment had been focused exclusively on proposing fully automated alignment methods and tools. A review of the results provided by the regularly OAEI participating alignment systems showed that there is a limit, in terms of alignment quality, that is difficult to overcome by solely automatic ontology alignment systems (Paulheim et al., 2013). Falconer and Noy (Falconer et Noy, 2011) identified the need to incorporate the user into the alignment process in order to review the proposed correspondence candidates and validate those that are correct and to propose additional correspondence that the system would not have found. The alignment process, thus, becomes semi-automatic because it combines automatic alignment algorithms and the knowledge (expertise) of experts in the studied field. The challenge is to provide a high quality alignment with minimal users interaction.

Brief Review of Related work

Different interaction mechanisms have been proposed during the alignment process. The interaction can be limited to the beginning of the process by only choosing the system configuration parameters. It may also concern the choice of the different metrics (Hamdi et al., 2010 ; Ritze et Paulheim, 2011). Other approaches recommend that the expert proposes an initial set of correspondences and non-correspondences (To et al., 2009).

In contrast to these strategies, interaction can involve correcting candidate matches (Severo et al., 2014), validating them (Shi et al., 2009) or evaluating the created alignment. Validating candidate correspondences can be done in an interactive way (Jimenez-Ruiz et al., 2012). This validation can also be done automatically when a partial reference alignment is available. The systems that participated in the task dedicated to the interactive alignment of the OAEI challenge in its 2013 edition mainly relied on this strategy (Faria, Pesquita et al., 2013).

The approach of (Lambrix et Kaliyaperumal, 2013) incorporates most of the strategies mentioned above. It is based on the use of user sessions to align large ontologies. The followed principle extends the approach implemented in the SAMBO system (B. Chen

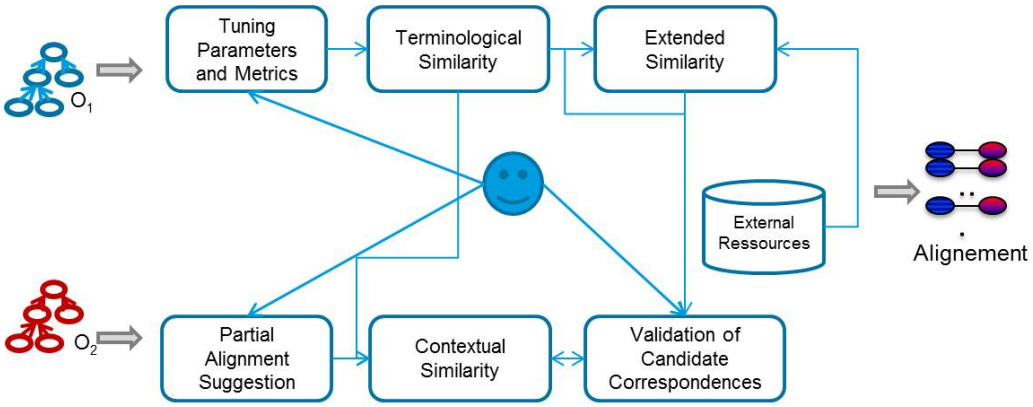

FIGURE 3.5 – Interactive Matching in ServOMap

et al., 2006) through a dedicated user interface, the user is integrated at several levels of the process. At the beginning of the process, the user starts with a new session, or by loading a previously stored session. A database is used to manage information about the sessions. In this first phase, the user provides the different parameters to be used for alignment. If partial alignment is available, the input ontologies are first partitioned based on this alignment.

Among these approaches, only that of (JIMÉNEZ-RUIZ et al., 2012) reported an evaluation in the context of large Ontology Alignment. We can therefore observe that as far as we know, none of the systems implements all the interaction mechanisms (choice of configuration parameters, choice of metrics, validation of candidate mappings) at the same time while being designed to align large ontologies.

Interactive Matching in ServOMap

The interactive matching strategy has been conducted respectively as part of the MSc internships of Mamadi Savadogo in 2014 and Nouha Kheder in 2015. The matching process in this context involves the user at several levels of the alignment process. The proposed approach is depicted in Figure 3.5. The user is involved in the result of each similarity calculation step and in the provision of intermediate results. There are three steps in similarity calculating according to the ServOMap approach, and described previously : terminological similarity, extended terminological similarity and contextual similarity. Before the contextual similarity, the user is given the possibility to propose a partial alignment that can play the role of the learning set (as described in Section **??**. Extensive terminological similarity uses external knowledge resources, in our case the WordNet lexical database (MILLER, 1995). Users involvement is done in the following steps.

1. **Tuning of the algorithm** : The user defines the settings to be used during the alignment process (section 3.6). As an indexing strategy is followed in ServOMap, users have the ability to select the kind of entities (concepts, properties, individuals) to index during the indexing step and the threshold to be used during the similarity computation.
2. **Choice of the metrics** : ServOMap includes various similarity measures for string comparison. The user has the ability to choose those that could be used for the current alignment process. As we have adopted the double threshold principle (B.

Chen et al., 2006), a user can choose the minimal and maximal threshold between 0 and 1.

3. **Partial alignment suggestion** : An initial set of available correspondences could be provided by the user to be reused during the alignment process. This is the case before the contextual similarity which needs a learning set for the generation of the candidate couples from the contextual similarity.
4. **Candidate correspondences validation** : The set of candidate correspondences that are computed are categorised as exact and dubious. The dubious one have their score within the double threshold range and they are processed by the user to validate those that are plausible. The validated one are re-injected into the alignment process in an iterative manner.

Performance of the interactive alignment

Two evaluation settings have been used in order to assess the performance achieved by the interactive alignment algorithm. They rely on the *Conference* track of the OAEI campaign (2013 and 2015), which has been developed in the context of the OntoFarm project [3]. It is constituted of 16 ontologies which describes the domain of Conference Organisation. The gold standard in this case is the result achieved by the classical alignment algorithm of ServOMap during the 2013 OAEI campaign.

The first setting is a traditional user based evaluation, where the user is asked to interact with the system. According to the double threshold principle the minimal score is set to 0.91 and the maximum to 1. The evaluation shows an increase of the performance by 8% in terms of F-score (0.7% Vs. 0.63%) between the classical and the interactive algorithm.

The second setting do not takes "real" users into account but use an Oracle instead which simulates the user behaviour. In the evaluation, we looked at how interacting with the user improves the matching results. The interactive matcher, evaluated in the context of the 2015 OAEI campaign, can present a correspondence to the Oracle, which then tells the system whether the correspondence is right or wrong. In this setting, our approach achieved an F-score of 0.78 (precision 0.65 and recall 1), which leads an increase of performance by 20%. Table 3.2 shows the achieved results according to different error rates introduced in the alignment references. We observe that, as expected, the highest F-score is obtained when no error is introduced in the reference alignment. Detailed results could be found in (Cheatham et al., 2015).

Error rate	0.0	0.1	0.2	0.3
Precision	1.0	1	1	1
Recall	0.617	0.587	0.553	0.519
F-Score	0.763	0.740	0.712	0.683

Table 3.2 – Performances of the interactive alignment Algorithm on the Conference dataset in an Oracle based setting.

3. https ://owl.vse.cz/ontofarm/

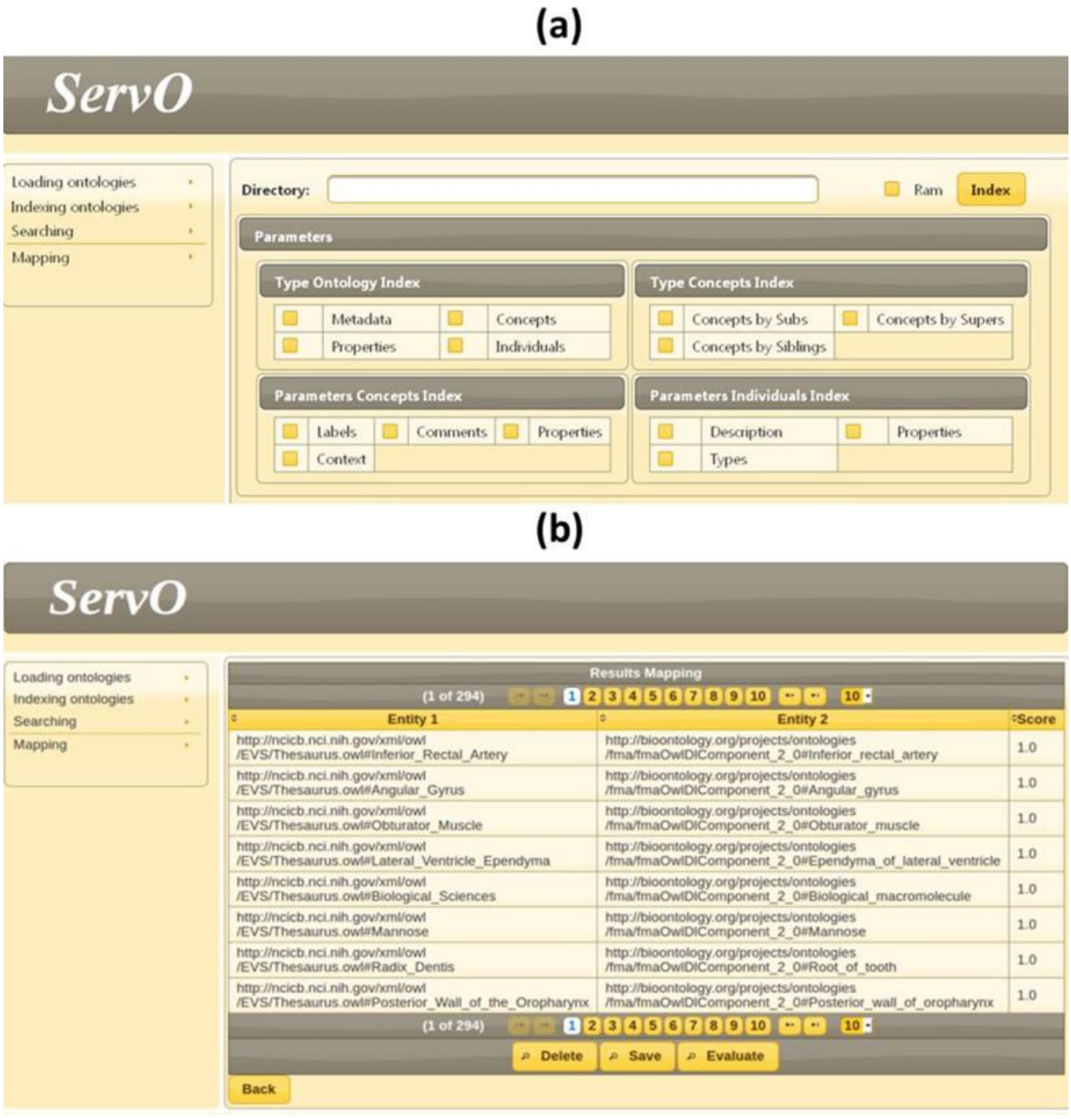

FIGURE 3.6 – Web Interface of the Interactive Matching with ServOMap

3.1.4 Multilingualism issue in Ontology Matching

More recently, we have invested some efforts on multilingual Ontology matching, a necessary task when it comes to deal with ontologies expressed in different natural languages (English, French, Italian, etc). It is done in collaboration with Marouen Kachroudi and Sadok Ben Yahia from LIPAH laboratory, University of Tunis El Manar (Tunisia). The following papers have contributed to this work.

M. Kachroudi, G. Diallo, S. Ben Yahia : Kepler at OAEI 2018. OM@ISWC 2018 : 173-178

E. Jimenez-Ruiz, T. Saveta, O. Zamazal, S. Hertling, M. Röder, I. Fundulaki, A.C. Ngonga Ngomo, M.A. Sherif, A. Annane, Z. Bellahsene, S. Ben Yahia, G. Diallo, D. Faria, M. Kachroudi, A. Khiat, P. Lambrix, H. Li, M. Mackeprang, M. Mohammadi, M. Rybinski, B.S. Balasubramani, C. Trojahn. Introducing the HOBBIT platform into the Ontology Alignment Evaluation Campaign. Proceedings of the 13th International Workshop on Ontology Matching co-located with the 13th International Semantic Web Conference (OM@ISWC) 2018 :49-60.

M. Kachroudi, G. Diallo, S. Ben Yahia : OAEI 2017 results of Kepler. OM@ISWC 2017 : 138-145

M. Kachroudi, G. Diallo, S. Ben Yahia. On the Composition of Large Biomedical Ontologies Alignment. ACM Proceedings of the 7th International Conference on Web Intelligence, Mining and Semantics (WIMS'17). pp :24 :1–24 :10, June 19-22th 2017, Amantea, Italy. Doi :10.1145/3102254.3102284

M. Kachroudi, G. Diallo, S. Ben Yahia. Initiating Cross-Lingual Ontology Alignment with Information Retrieval Techniques. Actes 6èmes Journées Francophones sur les Ontologies (JFO'2016), 13 - 14 October 2016, Bordeaux, France

Introduction

The multilingual Ontology matching approach has been implemented within the Kepler system, which is grounded from the previously described ServOMap system and the CLONA Ontology matching system from Kachroudi and Ben Yahia (Kachroudi, Diallo et Ben Yahia, 2016, 2018 ; Kachroudi, Diallo et Yahia, 2017).

Traditional (monolingual) alignment techniques often rely on lexical comparisons made between resource labels. Their scope is limited to ontologies in the same natural language or at least in comparable natural languages. However, there are many cases where monolingual alignment is not appropriate. In the FP7 EU-ADR project described in Section

1.4.2 for instance, the ontologies which encode the content of the different databases are expressed in different natural languages. Given the existing monolingual alignment tools limitations in such a context, there is a pressing need for the development of alignment techniques that can deal with ontologies in multiple natural languages.

A well known approach to achieve multi or cross-lingual Ontology alignment is to use translation techniques to convert a direct cross-lingual alignment problem into a monolingual one (Trojahn Dos Santos et al., 2010). Various techniques have been proposed in the literature. They include statistical machine translation and rule-based machine translations have been developed so as to improve the translation quality through word sense disambiguation (Navigli, 2009).

The Kepler approach follows a divide and conquer strategy and partition the input ontologies into small blocks to align. It comprises the following steps : Ontology partitioning, indexing step, translation and selection and filtering out candidates. The indexing step is brought from the classical ServOMap approach.

Ontology Partitioning

Ontologies to align are split into smaller parts to support the alignment task (Kachroudi, Zghal et al., 2013). Consequently, partitioning a set $B(C)$ is to find subsets B_1, B_2,...,B_n, encompassing semantically close elements bound by a relevant set of relationships, i.e., $O = \cup B_1, B_2, ..., B_n$ where B_i is an ontological block, and n is the resulting number of extracted blocks. Hence, an ontological portion can be defined as a reduced Ontology that could be extracted from another larger one by splitting up the latter according to its both constituents : structures and semantics. One way to obtain such a partitioning can be to maximise the relationships inside a block and minimise the relationship between the blocks themselves. The partitioning quality result can be evaluated using different criteria :

— The size of the generated blocks : that must have a reasonable size, i.e., a number of elements that can be handled by an alignment tool ;
— The number of the generated blocks : this number should be as small as possible to limit the number of blocks pairs to be aligned ;
— The compactness degree of a block : a block is said to be *substantially compact* whenever relations (lexical and structural ones) are stronger inside the block and low outside.

Translation to a pivot language

Two strategies are usually followed for systems which rely on translation : either translating the two ontologies to a pivot language or choosing one of the input ontologies language as pivot and translate the second one to it. The first approach is used in Kepler and English was chosen as the pivot language. The reason to this is the exploitation of an external semantic resource in our case, currently WordNet (Miller, 1995). The latter is a lexical database for the English language. To perform the translation, Microsoft Bing translator system [4] is used.

4. https ://www.bing.com/translator

Filtering out candidate mappings

All candidates are subject of a filtering to select the best candidates. An entity is considered as partially redundant if it belongs to two different candidate mappings. Given three ontological entities e_1 , e_2 and e_3. If e_1 is aligned to e_2, and then, e_1 is aligned to e_3, then this last alignment is qualified as doubtful as we only provide $1:1$ in the multilingual context. To overcome this issue, the topology of the two suspicious entities (e_3 neighbours with e_1 neighbours, e_2 neighbours with e_1 neighbours) are compared with respect to the redundant entity e_1, and retains the couple having the highest topological proximity value.

Performance of the multilingual approach

The multilingual component of Kepler has been evaluated in the context of the OAEI 2017 and 2018 editions respectively. The multilingual alignment task consists in aligning ontologies from the MultiFarm dataset (Meilicke et al., 2012). This dataset is composed of a subset of the OAEI Conference track dataset, translated in nine different languages (i.e., Chinese, Czech, Dutch, French, German, Portuguese, Russian, Spanish and Arabic). The main goal of the MultiFarm track is to evaluate the ability of the alignment systems to deal with multilingual ontologies. It serves the purpose of evaluating the strength and weakness of a given system across languages. English is used as pivot language for a greater consistency of obtained translations.

Two tasks were considered for the evaluation : aligning different ontologies and aligning same ontologies in different languages. For 2017, in the different ontologies case, the method is ranked 4 out of 8 systems with a F-score of 0.36. Whereas in the same ontologies case, the method occupies the first place with a F-score value of 0.52 (Achichi et al., 2017 ; Kachroudi, Diallo et Yahia, 2017). For 2018, in the different ontologies case, the method is ranked 3rd out of 4 systems which are able to provide non empty results, with a F-score of 0.27. Whereas in the same ontologies case, the method occupies the first place with a F-score value of 0.49 (Algergawy, Cheatham et al., 2018 ; Kachroudi, Diallo et Ben Yahia, 2018).

We observe that the use of a pivot language ensures greater consistency of obtained translations since it starts from the same text while an efficient exploitation of the topological structure of the ontologies leads to a good performance for same ontologies alignment.

3.2 Knowledge Warehousing for jointly handling KOS and their Alignments

The work described in this section is part of the **PhD thesis of Bruno Thiao Layel** started mid-2018 and in collaboration with the Synapse-Medicine start-up, a spin-off the BPH/ERIAS research group. This work has been partially valued in the following publications and is the subject of a software registration and contributed to a patent as well.

B. Thia-Layel, V. Jouhet, G. Diallo. Joint Handling of Semantic Knowledge Resources and their Alignments. Proceedings of the 13th International Workshop on Ontology Matching co-located with the 13th International Semantic Web Conference (OM@ISWC). 2018 : 226-227.

B. Thiao-Layel, V. Jouhet, G. Diallo. K-Ware : vers une gestion conjointe de ressources sémantiques et leurs alignements. 6èmes Jounrnées Francophons sur les Ontologies (JFO'2016), 13 - 14 Octobre 2016, Bordeaux, France

In Preparation

B. Thiao-Layel, V. Jouhet, G. Blin, G. Diallo. Heterogeneous Knowledge Integration with the K-Ware Framework. To be submitted to Journal on Data Semantics (JoDS).

Patent

L. Letenier, C. Goehrs, S. Cossin, B. Thiao-Layel, F. Thiessard, A. Pariente, G. Diallo, V. Jouhet. Dispositif et un procédé de génération d'une base de données relative aux médicaments. Num. demande France 1661257. 21 November 2016.

3.2.1 Context and need for jointly handling KOS

KOS, particularly terminology and ontology resources (TOR) are increasingly used as repositories and metadata for the annotation of various Information Systems data. This is particularly the case in the healthcare sector as well as the biomedical domain. Relying on such resources, we can enable semantic interoperability between different heterogeneous sources (Calvanese et al., 2018). In order to elucidate the need, let us introduce two use cases in two different areas where jointly handling heterogeneous KOS together with their alignments is both a necessity and an asset.

Motivating use cases

Drug related Data and Knowledge management at Synapse Medicine. Let us elaborate on the drugs related information and knowledge management at Synapse-Medicine. First of all, the Agence Nationale de Sécurité du Médicament et produits de santé (ANSM) in France provides a standardised information in the form of a Summary of Product Characteristics (SPCs) (please see 2.3), which are accessible in the online drugs database information [5]. These SPCs are available for pharmaceutical specialities that have obtained an Marketing Authorisation (MA) whether or not marketed. They are

5. https ://ansm.sante.fr/Services/Repertoire-des-medicaments

available within the online drugs database together with the patient leaflet. The latter is a document more specifically intended for the patient, and information related to a set of pharmaceutical and regulatory information, grouped in the form of a summary sheet including the name, the composition of active substances of the product. They are produced from information provided by the laboratories that have developed the drugs and are updated regularly. Whenever a drug is approved at European level, its SPC is available as a PDF document from the European Medicines Agency (EMA) [6].

In addition to that, official recommendations are produced by various public bodies (High Authority of Health - HAS, Health Insurance Agency, World Health Organisation - WHO, etc.) on the prescription and use of drugs. The WHO maintains the Anatomical Therapeutic Chemical Classification (ATC) System of active substances according to the body organ or system on which they act, with their therapeutic, pharmacological and chemical properties. It can be seen as knowledge base on substances. Finally, through the monitoring of drugs usage on the market and information provided by regional pharmacovigilance centres (CRPV) in France, the ANSM maintains and makes available in its Open Data portal [7] a repository of identified drug interactions in the form of a PDF document.

These data and knowledge about drugs are different in several respects : the source that produces them, the means by which they are available, their format (CSV files, HTML pages, relational databases, PDF files, etc.), their frequency of update, etc. Sometimes we are dealing with knowledge, even if it is not formalised in a formal and computable language, as it is the case with the HAS recommendations for example.

If one would like to fulfil the Synapse-Medicine need of efficiently identifying interactions between drugs, then there is a compelling need to integrate in a automated way :

— terminology resources in the field ;
— formal ontologies in the field ;
— structured representations of heterogeneous repositories ;
— links between terminologies, ontologies, repositories and the actual data to be analysed.

However, the number, rapid and asynchronous evolution and heterogeneity of semantic resources and repositories are a barrier to the implementation of this automated integration of drug related knowledge and data, beyond the management of drug interactions. From end user perspective, answering to this issue implies a solution taking into account the heterogeneity of the different resources, the links that may exist between them in one hand and with actual data in the other hand. The Observational Medical Outcomes Partnership (OMOP) Common Data Model is an example of initiative which attempts to provide a solution which allows systematic analysis of disparate observational databases (Hripcsak et al., 2015).

Example from the Cancer Domain. Cancerology is another example that fits the bill with the issue on handling and reusing data. Indeed, during a patient follow-up, a lot of information is produced by different actors and with different objectives. For administrative purposes, in France the coding of diagnoses for the French Programme de Médicalisation des Systèmes d'Information (PMSI) is based on the ICD10 terminology. In addition, pathologists use two different terminologies for the cyto-pathological description of tumour samples at different sites : the International Classification of Diseases for On-

6. http ://ema.europa.eu/
7. https ://www.data.gouv.fr/fr/datasets/base-de-donnees-publique-des-medicaments-base-officielle/

cology (ICD-O3) and the classification about lesions developed by the Association pour le Développement de l'informatique en Cytologie et en Anatomie Pathologique (ADICAP), Development of information technology in Cytology and Pathological Anatomy).

The sources concerned are therefore not semantically interoperable due to the use of different classifications. Different approaches have been discussed to exploit these data (Jouhet, Defossez et al., 2013 ; Tognazzo et al., 2009) and work has specifically tried to map the different diagnostic cancer terminology (Jouhet, Mougin et al., 2017). However, these codes can be used to characterise the same diagnosis but with a level of detail and a heterogeneous structure. There is therefore no real perfect match, but rather a close match between these codes, which requires complex matching relationships. It should also be noted that the set of diagnoses available for a given individual is a one-time description of a continuous process that takes place in the patient. Thus, the use of these data to monitor the evolution of an individual's Tumor Disease may require the implementation of formal models to describe the disease with regard to the diagnosis coded in the different terminologies. Thus, in order to allow the exploitation of this data, it is necessary to jointly implement formal terminology and models but also to manage the matches between them.

We have designed the K-Ware Knowledge Warehouse platform as a possible solution to answer to the above mentioned issues. It is intended to be used in the context of specific projects where KOS and their alignments must be integrated together. This integration with a detailed description of each resource using a robust and scalable model allows to add entities and new correspondences between them, to explore and validate the integrated resources but also allowing later querying the actual data from a very conceptual level thanks to the KOS that are managed. The objective is not to be a centralised repository for KOS like for instance BioPortal does (Noy et al., 2009) in the biomedical domain, but an embeddable component that can be integrated into the implementation of a heterogeneous resource integration project. In a more broad way, these KOS can be represented in different formalisms according to the usage and need, ranging from simple terminologies, like the MeSH (Dhammi et Kumar, 2014) and the ICD in its different versions over the years [8] to very formal ontologies based on representation languages such as Description Logics (Baader et Nutt, 2003).

Brief Overview of Related Work

In order to improve the management of heterogeneous data sources, it is first necessary to be able to manage the repositories and metadata that are used to annotate their content. A significant effort has been made for providing multi-knowledge resources repositories. This effort is particularly noticeable in the biomedical domain, where classifying existing objects is a secular tradition. Thus, the NCBO BioPortal (Noy et al., 2009) manages currently more than seven hundreds KOS together with available alignments between them. Ontology Lookup Service (OLS) (Côté et al., 2006) handles about two hundreds fifty KOS. Recently, a service for finding mappings or cross-references between terms from different KOS has been included. Aber-OWL is another Ontology repository that provides reasoning services for bio-ontologies (Hoehndorf et al., 2015). It enables ontology-based semantic access to biological data and literature thanks to the annotation provided by the handled ontologies. While BioPortal and OLS provide web services and interfaces to access the ontologies that they manage, they do not offer computational reasoning possibility

8. https ://www.who.int/classifications/icd/en/

to the semantic content of their ontologies. Moreover, before to be incorporated into its repository, BioPortal for instance, translates any hierarchical relation between entities of a given KOS (e.g. *skos :broader*) into the *rdfs :subClassOf* subsumption relation losing the possibility to keep the original semantics of the KOS. In contrast, Aber-OWL offers a reasoning infrastructure over its managed ontologies and enables semantic queries over ontologies specified by a user or contained in its repository. However, it does not allow explicit handling the alignments between the ontologies within its repository.

The Health Termino-Ontological Portal HeTOP is another project similar to BioPortal (Noy et al., 2009) but which addresses KOS expressed in the French language. It is a multi-terminological repository for biomedical resources where any KOS is integrated within an ad-hoc semantic model with the help of domain expert instead of solely relying on automatic imports on common standard. The Terminological, Linguistic and Ontological Knowledge Resources Management (TOK) is another approach that is distinguished by the fact that it differentiate KOS to be managed from their abstract model. It provides a metamodel with descriptors to represent relationships, hierarchical relations and resource types to be handled (Ghoula et al., 2010). However it is not sufficiently abstract for instance by defining independently from terminology or ontology the notion of hierarchical relationship.

To the best of our knowledge, there is no currently available framework which offers the possibility to handle both multiple KOS together with their respective alignments, while keeping their native semantics and offering a support for a transparent visualisation of these resources. In addition, with the development of the ontology matching domain (Otero-Cerdeira et al., 2015), as different systems could be used to generate alignments and sometimes relying on user input, either for mappings validation purpose or initial alignment providing (Savadogo et al., 2016), it is a crucial issue to keep track of users and involved alignment methods or tools.

3.2.2 The K-Ware knowledge warehouse framework

K-Ware more specifically addresses the issue of **managing the persistence** of different KOS by making them permanent while **taking into account their evolution** over time or according to need. Once properly stored, the use of hierarchical relationships derived, as the case may be, from ontological or terminological resources must be differentiated. Indeed, semantic resources have an *original semantics* that is necessary to preserve. For example, some semantic resources are structured around of the **traditional subsumption relationship** between concepts (the is-a relationship) as specified by the OWL language, while others are structured around generic pure **terminological relationships** (e.g. *broader* and *narrower*, as specified by Simple Knowledge Organisation System (SKOS) language (Miles et al., 2005). Therefore, it is necessary to preserve these differences when integrating heterogeneous KOS. Moreover, managing heterogeneous KOS involves **managing their interlinks called alignment** (Shvaiko, Euzenat et al., 2013) that improve the interoperability of sources annotated or described by these KOS. Their identification and handling (editing, validation, etc.) can be complex and challenging. Then, the **possible composition** and export at request of entire semantic resources or parts of them are services to be offered. The **visualisation** is one of the supports that allows an interactive exploration within the framework of large graphs that these KOS can constitute. Finally, the challenge is to provide services that allow to **keep the link with the data** of the sources annotated by these resources in order to be able to

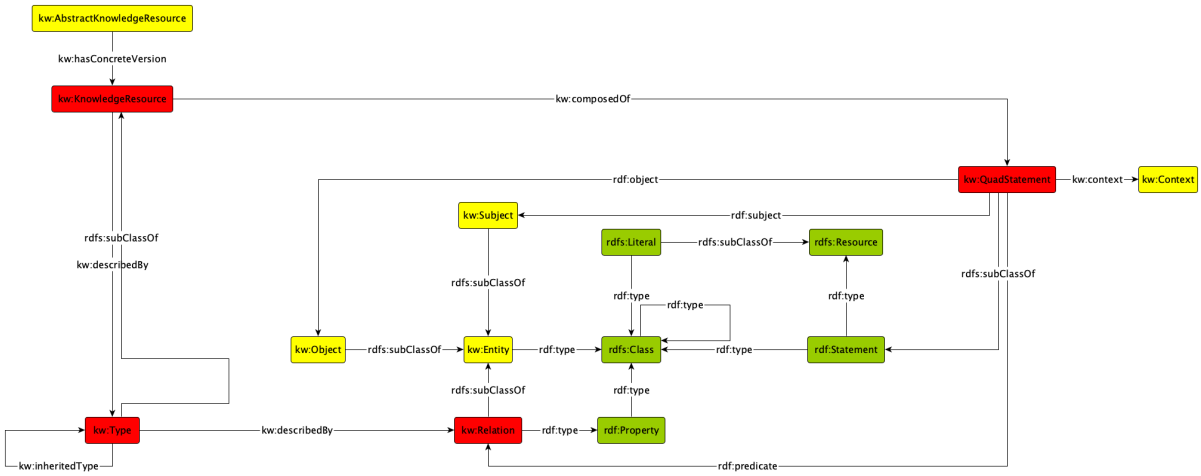

FIGURE 3.7 – K-Ware generic meta-model for handling heterogeneous KOS.

access them at semantic level. This linkage enables several perspectives on these sources by changing only each time the semantic resource used for querying.

Overview of the Approach

K-Ware Generic Metamodel. The K-Ware metamodel for handling any type of KOS is depicted in Figure 3.7. All green nodes are compatible with the RDFS Resource Description Framework model and the metamodel allows representing rdf-based knowledge resources (OWL, SKOS, RDF). Any handled KOS is represented by an *AbstractKnowledgeResource*, for instance the ICD, which can have a set of concrete versions (*KnowledgeResource*). Any *KnowledgeResource* inherits from a particular *Type*. Similarly, a *KnowledgeResource* is composed of a set of *QuadStatment*. A *QuadStatment* is a quadruplet constituted by a traditional RDF triplet (subject, predicat, object) and a given *Context* (the actual graph of the KOS being processed).

Structural hierarchical relations handling for KOS. As has been previously mentioned, KOS used to have a common structural organisation : the *is-a* relationship which is translated into *rdfs :subClassOf* or the *part-whole* relationships for formal ontologies and the *narrower/broader* relationships for taxonomies or structured vocabularies which are translated into *skos :narrower* and *skos :broader* respectively. We find similar notions when it comes to express inter relationships (mappings) between entities in different KOS. For instance, the *skos :narrowMatch* is used to state a hierarchical mapping link between two conceptual resources in different concept schemes.

The main information to be kept whenever one is interested in integrating many different semantic resources is their hierarchical aspect according to (HEPP et BRUIJN, 2007), which introduces a way to transform hierarchical classifications, thesauri and taxonomies into an OWL-based consistent Ontology by keeping their initial representation. As depicted in Figure 3.8, in order to allow navigating easily KOS represented in an RDF compliant type (like OWL or the SKOS language), we distinguish three types of relations (*Relation* : *HierarchicalRelation, LiteralDefinition* and *MappingRelation*). Further, a *HierarchicalRelation* is used to handle non taxonomic relations between a KOS entities. In a given KOS, an example of hierarchical relation could be *rdfs :subClassOf*, *skos :broder* or *part of*. A *literal Definition* allows rendering a human readable description of an entity

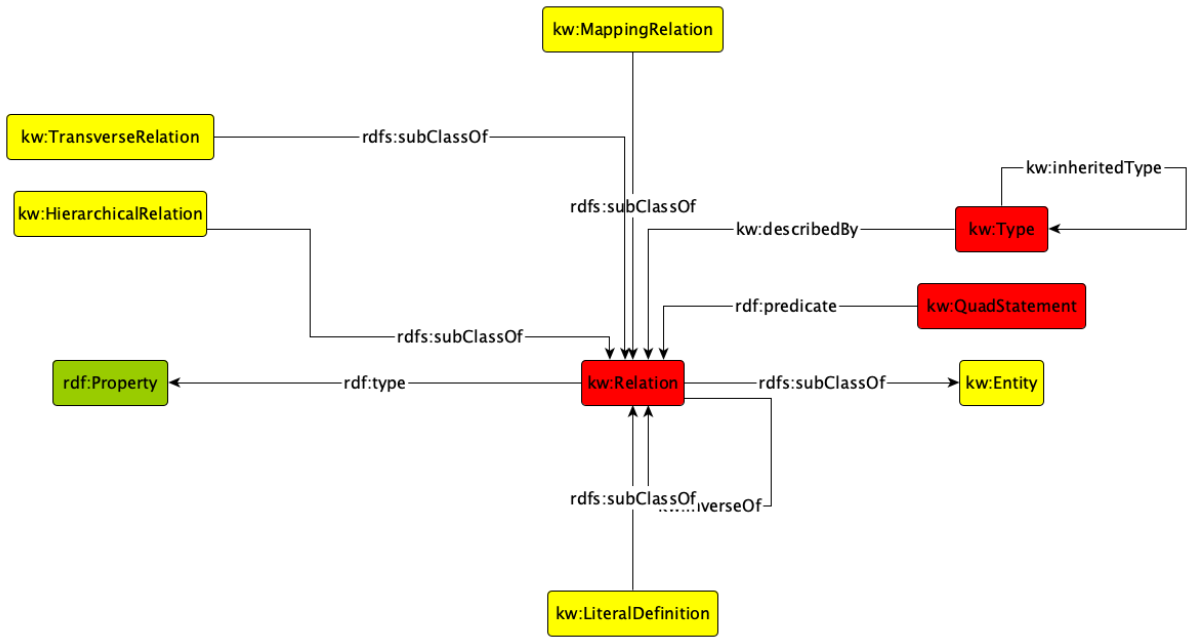

FIGURE 3.8 – Partial view of the metamodel for managing structural hierarchies within and between heterogeneous KOS.

(e.g., *rdfs :label*, *skos :prefLabel*, etc.). While a *mapping relation*, for instance *owl :equivalentTo* or *skos :exactMatch* handle correspondences (mappings) between entities. For mappings, it can be further specialised into *HierarchicalMappingRelation*, for instance *skos :broadMatch* or *skos :narrowMatch*. Further specialisations of the different type of relation that could be encountered within and between SKOS is available on the online K-Ware portal [9].

Handling of Alignments. Several alignments, generated by different tools (*ComputerisedProcess*, (for instance ServOMap 3.1 or LogMap (JIMÉNEZ-RUIZ, CUENCA GRAU et al., 2013)) and methods (*MappingMethod*, can exist between KOS. In order to take them properly into account, we have extended, with additional properties, the traditional definition of the notion of correspondence between entities introduced in (SHVAIKO et EUZENAT, 2013) and recalled in section 1.4.2. Therefore, and according to Figure 3.9, we identify at least the following components for a correspondence between KOS :

- the two entities to be mapped from different KOS linked with a *MappingRelation*, in an directed order ; $(e_1 - mappingRelation - e_2)$
- its confidence value ;
- the Mapping's author, the *User* who authored the correspondence, defined with a *foaf :Agent* entity from the FOAF Ontology [10] ;
- the Mapping's generation method, called **MappingMethod**, either it is an automated alignment method with an alignment **Tool** or a manual one ;
- finally, a flag which indicates whether the considered mapping is valid (and validated by a User) for an **Alignment** between two **KOS**.

9. http ://k-wa.re/models
10. http ://xmlns.com/foaf/spec/

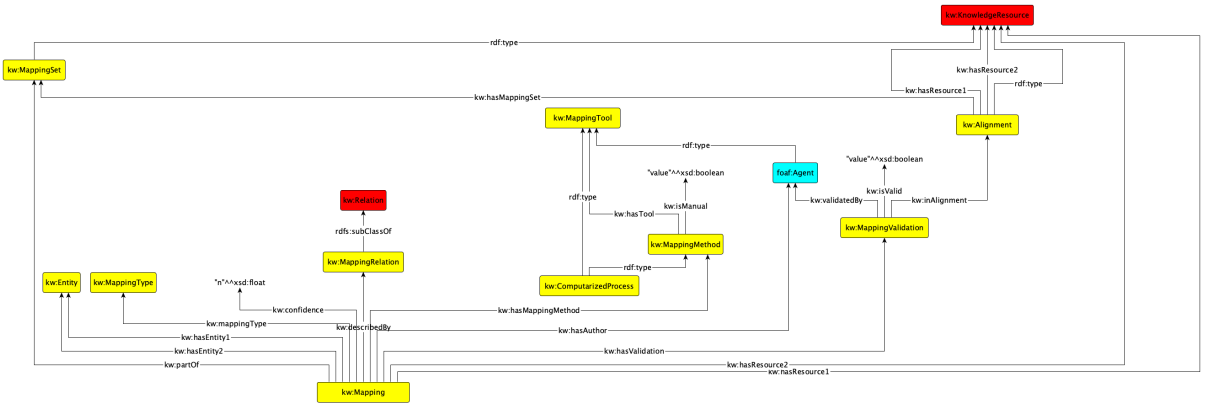

FIGURE 3.9 – Alignment management between different KOS (knowledge resources).

Implemented tool and API

We have implemented an API for interacting with the different models into K-Ware. It available online [11]. The full specification of the back-end API, designed with Swagger [12], is available online as well [13]. It allows, among others, navigating within a given KOS by exploiting transparently the hierarchical relationships. Further, it allows to navigate between different KOS thanks to the use of the defined mappings relations. Although the design of K-Ware is completely agnostic to any particular triplestore, currently, BlazeGraph [14] is used as back-end for storing the KOS and their different alignments.

The model and implemented algorithms are currently being adapted in the context of the Synapse-Medicine start-up.

3.3 Conclusion

In the first part of this chapter, we have described our contributions on handling the issues related to large scale Ontology matching and the subsequent user involvement and multilingualism. All the implemented algorithms and tools have been evaluated in the context of the international Ontology Alignment Evaluation Initiative competition, the main framework in the fields, using standard datasets.

Matching large ontologies requires some optimisation strategies. We have proposed *i*) a binary representation of the taxonomy of the input ontologies to fasten entities look-up and *ii*) an Ontology indexing strategy which reduces the search space during the similarity computing. A machine learning strategy is used to take into account the topological structure of input ontologies being matched.

Users involvement is of particular importance for improving the alignment quality. We designed and implemented an interactive Ontology approach through a web application to evolve the ServOMap large-scale alignment system. This web application provides a user friendly interface to facilitate user intervention during the process. The interactive system described in (JIMÉNEZ-RUIZ, GRAU et al., 2012) allows to simulate interactivity with the user through heuristics but does not offer a user interface. From this point of

11. http ://k-wa.re/app
12. https ://swagger.io/
13. https ://tinyurl.com/y5hwjguq
14. https ://www.blazegraph.com/

view, our system is close to the one described in (Lambrix et Kaliyaperumal, 2013) which is based on user sessions. Unlike the latter, our interactive system has been tested with large ontologies. However, it does not currently allow to reuse the results of one interactive session to another through the backup of user sessions.

The second part, which is a natural continuation of the work conducted on Ontology alignment, was dedicated to the K-Ware knowledge warehouse platform. Thanks to a generic meta-model, it allows handling jointly both heterogeneous knowledge organisation systems and their respective alignments. An API allows interacting with the designed model. It is being adapted to drugs specific knowledge and data management in the context of the Synapse Medicine start-up.

Chapitre 4

ICT as Digital Public Health Enabler in Countries with Limited Settings

Contents

The reported work in this chapter has been published in the following papers :

B. Kamsu-Foguem, G. Diallo, C. Foguem. Conceptual Graphs-based Knowledge Specification for Supporting Reasoning in African Traditional Medicine. Engineering Applications of Artificial Intelligence, Volume 26 Issue 4, April, 2013. Pages 1348-1365. DOI information : 10.1016/j.engappai.2012.12.004

E. Mutafungwa, F. Thiessard, MP. Diallo, R. Gore, V. Jouhet, F. Mougin, S. Ben Yahia, C. Karray, N. Kheder, R. Saddem, J. Hämäläinend, G. Diallo. Mobile Data as Public Health Decision Enabler : A Case Study of Cardiac and Neurological Emergencies. Proceedings of the Orange D4D Challenge, in conjunction with the 4th International conference on the scientific analysis of mobile phone datasets (NetMob'2015), 7-10 April 2015, MIT Media Lab, Cambridge, MA, USA.

BMJ. Some, G. Bordea, F. Thiessard, S. Schulz, G. Diallo : Design Considerations for a Knowledge Graph : the WATRIMed use case. Studies in Health Technology and Informatics, vol.259 :59-64. IOS Press. DOI :10.3233/978-1-61499-961-4-59

BMJ. Some, G. Bordea, F. Thiessard, G. Diallo : Enabling West African Herbal Traditional Medicine Digitalization : the WATRIMed Knowledge Graph. MEDINFO'2019, 26-30 Aug 2019, Lyon, France.

4.1 General Overview

With the aim of supporting the United Nations Sustainable Development Goals (SDG) [1], specifically "ensuring healthy lives and promoting well-being for all at all ages" (Goal 3) and "reducing inequality within and among countries" (Goal 10), the research work conducted under the Information and Communication Technology for Development (ICT4D) umbrella intends, from a public health perspective, to specifically target countries with low income and limited resources (according to United Nation Development Programme (UNDP) classification [2]), in particular the populations that are the most in needs. According to the World Health Organisation Commission on the social determinants of health, "the poor health of the poor, the social gradients in health and the marked health inequities between countries are caused by the unequal distribution of power, income, goods and services".

Therefore, we will briefly describe in this chapter a contribution related to two studies that leverage the potential of ICT technologies, from large heterogeneous data integration to semantic technologies with knowledge representation, in order to contribute tackling some Public Health issues in limited resources countries. The first one is related to the use of Big (meta)data that are passively collected by mobile phone operators from their network usage, and known as Call Data Records (CDR), to identify geographic areas in a country where people at risk of medical emergencies can have dramatic consequences due to a lack of appropriate health care infrastructures. This work is part of the Data for Development project (2014-2015) that I coordinated and which involved four research groups in the area of health informatics, datamining, mobile technology and big data simulation and modelling. The second study is the WATRIMed project [3] that I do coordinate and which is dedicated to bringing Semantic Web technologies, in particular formal knowledge representation, to modelling knowledge about herbal-based medicine. The ultimate aim is to contribute preserving a secular traditional knowledge while establishing bridge with knowledge from conventional medicine.

1. https ://www.un.org/sustainabledevelopment/sustainable-development-goals/
2. https ://www.undp.org/content/undp/en/home.html
3. http ://www.watrimed.org

4.2 Large Scale CDR Analysis and Integration for tackling Public Health Issues

4.2.1 Background

Big data, especially as applied to development and public policy issues, is at the early stage of research and implementation. In countries with very limited settings such as those from low and middle income one, it is a real challenge to obtain valuable data to be used for elaborating statistics. In this context, the emergence of new types of data or metadata, collected passively from the population and which can be anonymised to make them compliant for secondary use, presents a real opportunity. Indeed, they can complement self reported data from a posteriori questionnaires as used in Public Health studies for instance. In this context, CDRs are increasingly being used in research along with other forms of big data like those from sensors in the context of smart healthcare process (Cuzzocrea et al., 2018).

In the health sector, although CDRs are not health data per se, they have been used to improve health and health care, for example, via the generation of epidemiological models that can infer the spatial spread of infectious diseases from human mobility patterns (Jones et al., 2018). As of illustration, Malaria mapping in Kenya was possible by analysing mobile phone data to study the regional travel patterns of nearly 15 million mobile phone subscribers in Kenya over the course of a year (Wesolowski et al., 2012). They used spatially explicit mobile phone data and malaria prevalence information from Kenya and were able to identify importation routes that contribute to malaria epidemiology on regional spatial scales.

Tackling health related issues, thanks to the recent advancements in data analysis over huge amount of data sources, almost all the determinants of our health – from our individual genetic coding to our particular habits – is becoming knowable. In that context, Big Data may be the future for healthcare. Besides the possibility of achieving personalised medicine (Horgan et Lal, 2019), cross-linking and analysing various heterogeneous data sources can help early identification of factors that influence people health.

4.2.2 Big Call Data Records harnessing for risky zones identification

Some medical emergencies require rapid hospitalisation for their care. Myocardial Infraction (MA) and Stroke are two examples of diseases that fall within this framework and with a known maximum time limit for their treatment.

MA is an absolute cardiological emergency, the incidence remains high with 120,000 cases per year in France and about 610,000 people who die in the United States according to the Center for Disease and Control (CDC) [4]. According to WHO data, ischemic heart disease is the leading cause of death with 7.2 million of coronary heart disease death over the 50 million annual deaths worldwide. The prognosis remains serious since the MA is still responsible for 10 to 12% of total annual adult mortality. In case of MA, it is possible to perform a mechanical unblocking by the expansion of a balloon in a coronary to be carried out within 90 minutes after the first signs of the crisis.

Stroke is the leading cause in Western countries of acquired disability in adults, the

4. http ://www.cdc.gov

second cause of vascular dementia after Alzheimer's disease (Vijayan et Reddy, 2016), and the third leading cause of mortality. In Europe, the annual incidence of stroke is between 101 and 239 per 100,000 for men and between 63 and 159 per 100,000 for women. In developing country, such as Senegal, the burden of stroke and other non-communicable diseases has risen sharply. In Dakar the capital city for instance, stroke is the most frequent neurological disease with the highest mortality rate [5].

In this study we propose an approach that integrate huge amount of recorded and anonymized mobile data and crosslink them with various heterogeneous data and knowledge in order to identify and estimate population at risk of major Public Health issues, in particular stroke and MA. To that end, we rely on data provided in the context of the Data For Development (D4D) challenge launched in 2014 by Orange France Telecom and Sonatel mobile operator [6]. The provided data concerned the Senegal country.

4.2.3 Description of the Data

Senegal is a West African country bordered by 5 countries (Gambia, Guinea-Bissao, Guinea, Mauritania and Mali) and totalling 196,712 km2. The country is subdivided in 14 regions. It is further subdivided by 45 Départements, 123 Arrondissements and by a set of Collectivités Locales. The total population is estimated to 13,508,715 people according to the 2013 General Census of the Population. The capital city is Dakar. It concentrates the main business activity of the country. The main figures about the population in Senegal could be found in (ANSD, 2014). According to these figures, Dakar for instance has a global population of 3,137,196, an area of 547 km^2 and 5,735.3 of density. Whereas in the Kedougou region we observe a global population of 72,490, an area of 16,800 km^2 and 9.0 of density.

Sonatel Orange Dataset

The dataset provided by the mobile operator are based on fully anonymised CDRs of mobile phone calls and SMS between the company customers in Senegal between January 1st 2013 and December 31st 2014 (Montjoye et al., 2014). The collected CDR which initially comprises 9 million unique aliased mobile phone numbers have been reduced following two criteria (Montjoye et al., 2014) :

— users having more than 75% days with interactions per given period (biweekly for the second dataset, yearly for the third dataset)

— users having had an average of less than 1,000 interactions per week. The users with more than 1,000 interactions per week were presumed to be machines or shared phones.

Dataset 1 : it contains metadata about the traffic between each antenna for 2013. It includes both voice and text traffic. Table 4.1 gives an example for voice traffic between sites (antennas).

Dataset 2 : Fine-grained mobility data (site level) on a rolling 2-week basis for about 300,000 randomly sampled users meeting the two criteria mentioned before for each 2

5. according to the clinique neurologique du CHU de Fann http ://www.chnu-fann.com/index.php/neurologie

6. https ://sonatel.sn/

Timestamp	Outgoing site	Incoming site	Number of calls	Total call duration
2013-04-01 00	2	2	7	138
2013-04-01 00	2	3	4	136
2013-04-01 00	2	4	7	121
2013-04-01 00	2	5	13	272
2013-04-30 23	1651	1632	1	3601
2013-04-30 23	1653	575	1	20

TABLE 4.1 – Voice call traffic between antennas. Dataset 1 of Orange Sonatel for the Data for Development Challenge Senegal. Outgoing site and Incoming site are respectively the mobile antenna from which calls are issued and received.

user	timestamp	site
1	18/03/2013 21 :30	716
1	18/03/2013 21 :40	718
1	19/03/2013 20 :40	716
1	19/03/2013 20 :40	716
1	19/03/2013 20 :40	716
1	19/03/2013 20 :40	716
1	19/03/2013 21 :00	716
1	19/03/2013 21 :30	718
1	20/03/2013 09 :10	705
1	21/03/2013 13 :00	705

TABLE 4.2 – Example of fine-grained mobility data for randomly selected users.

weeks period 4.2.

Similarly, one year of coarse-grained (123 arrondissement level) mobility data about 150,000 randomly sampled users meeting the two criteria mentioned before for a year are also provided **Dataset 3)**. In addition to CDR related data, the administrative organisation of Senegal is provided as well as the antenna GPS coordinates (1,666 antennas) and the arrondissement to which they belong to. Shape files for Senegal are also provided.

The CDR data represents several hundred million tuples (50GO). Columns that might help re-identification have been 3-anonymised when binned to remove outliers (SWEENEY, 2002).

For the study we also retrieved maps based data from OpenStreetMap.org [7] as well as the list of hospitals compiled from the online Senegal Medical Directory [8] and the SenDoctor web site [9].

4.2.4 The Methodological Approach

The overall methodology depicted on Figure 4.1 is conducted in regards with the following assumptions :

— the average incidence rate of stroke in Senegal is estimated to 100 per 100,000 inhabitants. As there is no recent documented study which gives figures that could be used, we based our hypothesis on the fact that the incidence of stroke in Europe

7. http ://www.openstreetmap.org
8. http ://www.annuairemedical-senegal.com
9. http ://sendocteur.com/structureregion.php

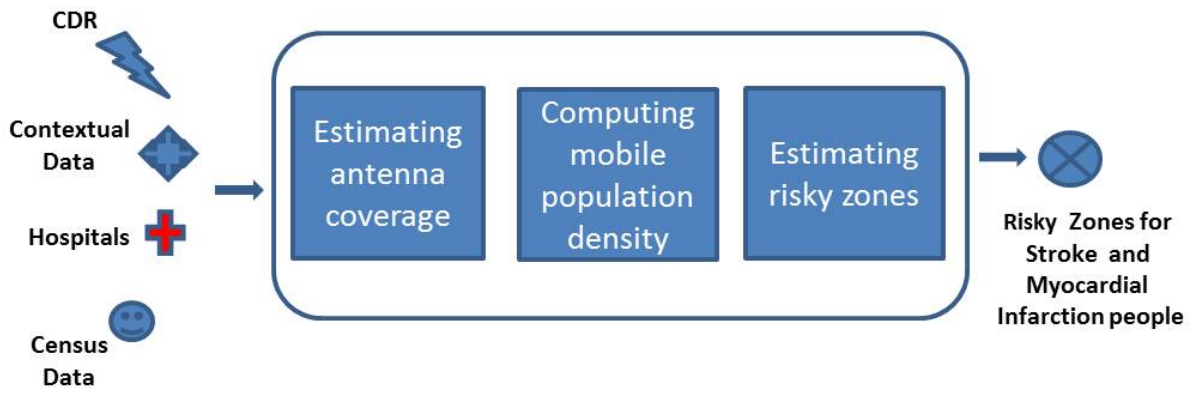

FIGURE 4.1 – Harnessing CDR data for risky zones identification for people with an emergency case of Stroke or Myocardial Infarction.

is between 63 and 159 for 100,000 women, and between 101 and 239 for 100,000 men at the time of the study. For Senegal, it is supposed to be less ;
— the average incidence rate of MA is estimated to 180 per 100,000 inhabitants in France. Based on that, we assumed that the incidence rate in Senegal is about 150 per 100,000 inhabitants ;
— we base our estimation distance from home to the nearest hospital in case of emergency of stroke on the recommendations from the French HAS. The recommendations estimate that for a severe stroke, the patient needs to be taken in charge in no more than 3 hours ;
— for MA, the maximum time for an efficient management is estimated to 1h30 (90 minutes).

Estimating Antenna Coverage

The geographical coverage areas of mobile (cellular) networks have often been described using equal-sized hexagonal coverage areas around each cell site. These hexagonal grid models have routinely been used for system performance studies for instance in Third Generation Partnership Project (3GPP) standardisation studies (HOLMA et TOSKALA, 2010). However, in reality the cellular layout is highly irregular due to constraints on the where the cell site could be located, spatio-temporal variations in mobile penetration and population density, the surrounding topography, presence of buildings, and so on (HOLMA et TOSKALA, 2010).

The use of *Voronoi diagrams* have been proposed as tesselation that overcomes the inaccurate hexagonal grid cellular representation when compared to real world cellular network layouts (BAERT et SEME, 2004). In the Voronoi diagram approach, cellular network for an area covered by N cells sites area is subdivided into convex polygonal regions around N points that correspond to the locations of the N sites. The irregular shape of the polygons allows providing a relatively better approximation of cell size by taking into account irregular site locations and proximity of neighbouring sites. The Voronoi tessalation generated for the provided 1,666 cell sites in Senegal is shown in Figure 4.2.

Computing Mobile Population Density In order to have an overall view of the distribution of mobile population at regional level, we used data from dataset 2 described

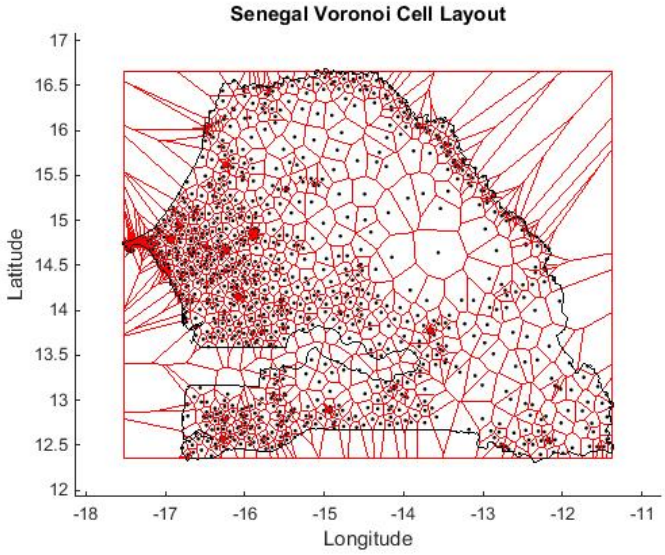

FIGURE 4.2 – Voronoi cell layout for Senegal based on provided 1,666 site locations.

previously. The idea is to aggregate a daily average mobile phones identified around a particular antenna. We make the hypothesis that census corresponds to the place where people are globally. We use two correction factors :

— α for adjusting identified unique users to the 300,000 two weeks based Orange Sonatel users ;

— β to adjust the number of people phoning to the expected number of people in the area according to the census provided by (ANSD, 2014) and the Orange/Sonatel market share in Senegal.

$\alpha = U_s.1.2.\frac{1}{Oms}$

$\beta = \frac{U_R}{O_R}$

where

— U_S is the number of unique users per day and per antenna computed from the dataset 2

— Oms is the estimated Orange Sonatel market share in 2013 according to GSMA survey

— the factor 1.2 is obtained by adjusting the total numbers of unique users to 300.000 as of dataset 2

The α and β adjustment coefficients are important to make the corrections of all counted number of people calling during a given day. This help having an idea of the real number of people at each place (antenna location) knowing that the total number of people should be 13,508,715 according to the 2013 Senegal population census (ANSD, 2014). For instance, the β correction factor for the Dakar region is 9.05 while it is 160.11 for the Sedhiou region and even 509.43 for the Kedougou region. Table 4.3 gives an example of estimated population by antenna site.

Estimating Risk Zones Many health geographers use distance as a simple measure of accessibility, risk, or disparity in terms of availability of health services in different locations (DUMMER, 2008). Time-critical medical emergencies like heart attacks and strokes considered in this study require that hospitals that provide emergency care are within a

Site	Dakar	Unique User-Day
1	Dakar	160.24
158	Tambacounda	170.05
1,405	Saint-Louis	181.04

TABLE 4.3 – Estimated population by antenna site.

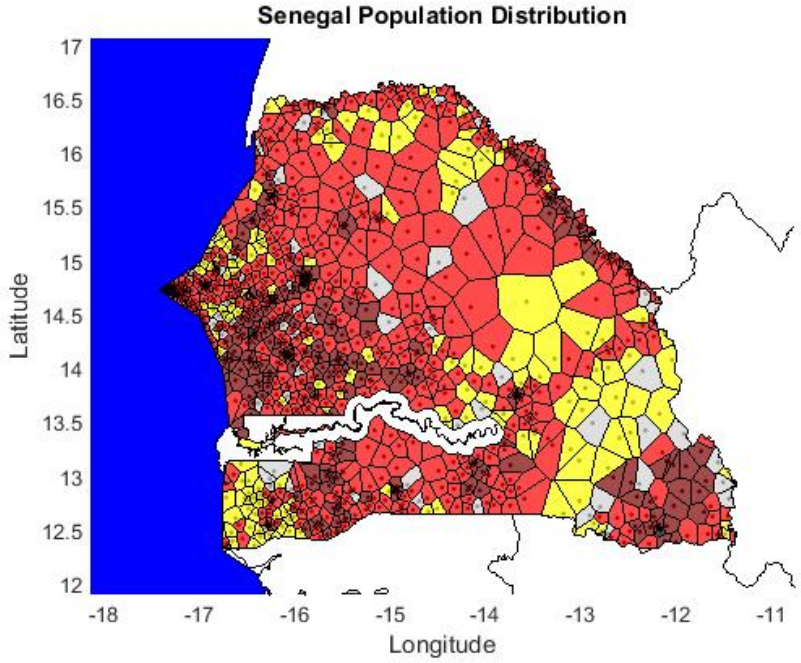

FIGURE 4.3 – Distribution of Senegal population according to the sites (antennas). Lets assume that N represents number of people in a given antenna coverage. We have used 5 different colours respectively grey for N < 100, yellow for 101<N<1000, red for 1,001<N<10,000, brown for 10001<N<100,000 and black for N >100,001.

certain distance of the victims requiring immediate care. For the case of medical emergencies due to heart attacks and strokes, locations from which victims are unable to reach the hospitals within a given time are considered to be high risk areas.

In this study we utilise the mobility dataset provided to evaluate the areas that are considered at high risk based on given time criteria (the distance to the nearest hospital is considered as a risk factor). This involves mapping the hospital and populated distribution on to the cellular network layout. A total of 40 hospitals located across the 14 regions of Senegal where considered (see Figure 4.4 with the capital city Dakar highlighted). The geographical location of the hospitals was approximated by representing them using site IDs of the nearest antenna site. Using this approximation it was noted that 85% of the 40 hospitals considered were within 2 km of their real geographical locations (see Figure 4.5). The impact of this error is minimal when evaluating the travel time to the hospitals.

The segmentation of locations according to their proximity to hospitals is also done based on cell areas. To that end, a common travel time is assumed to reach a hospital from a particular antenna for all people located in the same cell area. Furthermore, the antenna site location is assumed to the centroid of the cell area and the distance to the hospital for cell is the distance between the cell antenna site and the antenna site to which the hospital is associated.

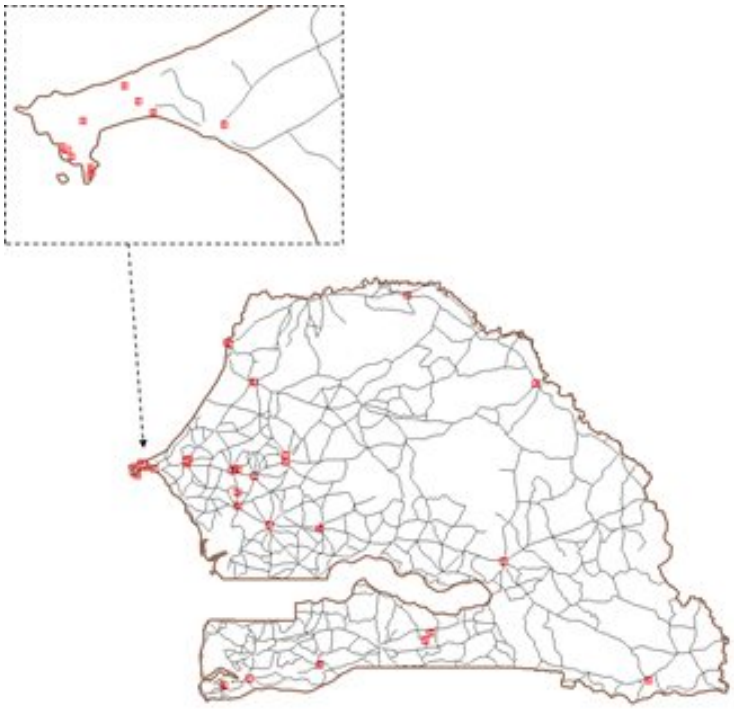

FIGURE 4.4 – Location of the 40 hospitals considered in the study. They are represented in red symbols while the road network is shown in black lines. The Dakar region is shown inset.

A significant number past health planning studies have used the straight-line ("as the crow flies") to measure the distance between two points on the map (BOSCOE et al., 2012). The approach uses either the spherical distance for geo coordinates (latitude and longitude) or Euclidean distance for projected coordinates. However, the real drive distances (and hence travel times) between the two points tend to be longer due to the fact that roads are built around natural obstacles, such as, mountains, boulders and so on. To that end, a correction factor known as *detour index* representing the ratio of the drive-distance to the straight-line distance has been introduced to obtain more accurate distance estimates with less computation or measurement effort (BOSCOE et al., 2012). The detour index approaches the lower bound of 1 the denser the road network. In developed economies detour indices in the range of 1.2 to 1.6 have been noted.

For the study we calculate the straight-line distance from each hospital to all cell sites and we then use a detour index of 2 to evaluate the drive distances between the points. The detour index assumption is rather conservative to take into account the relative low road network density and the less than ideal road conditions. Furthermore, for simplicity an average driving speed of 60 km/hr is assumed for all areas. An estimate of the travel time ranges from each cell area to the nearest hospital is illustrated in Figure 4.5. The maximum of 90 and 180 minutes are considered based to the treatment time windows for the cardiovascular conditions considered in this study. The cell areas beyond the 180 minutes catchment area are considered high risk areas for all conditions.

Main Results of the Study

From the expected 13,508 strokes each year in Senegal, only 462 (3.42%) will occur too far from a hospital to be able to have the thrombolysis treatment, because 96% of the

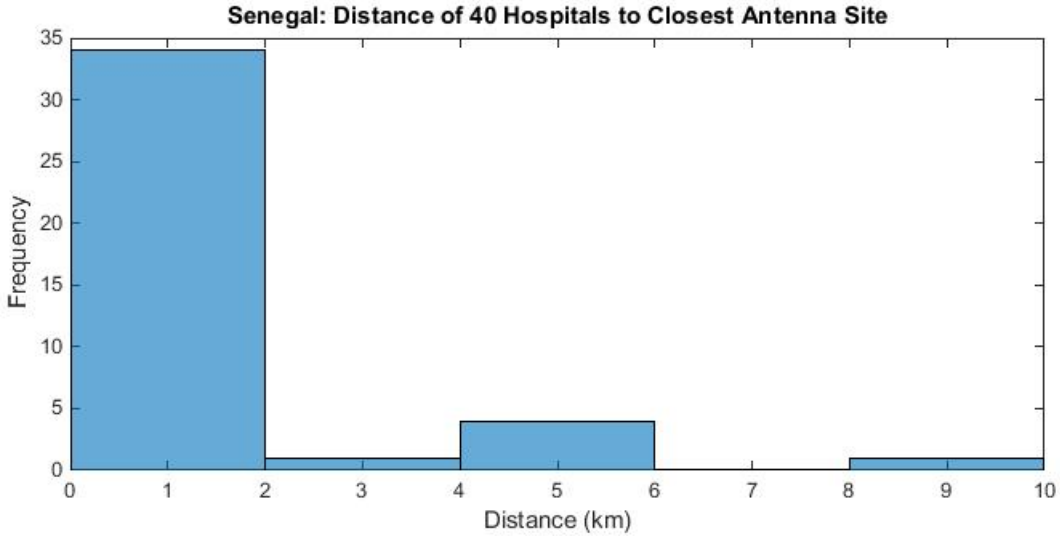

FIGURE 4.5 – Distance of hospitals to closest antenna sites.

population can reach a hospital in less than 3 hours (assuming that all the hospitals are able to do the treatment) according to Figure 4.6.

By cons, from the expected 24,315 MA, we observe that 4,241 (17.4%) will occur too far from a hospital to be able to have the balloon treatment because they cannot reach the hospital in less than 90 minutes (Figure4.7).

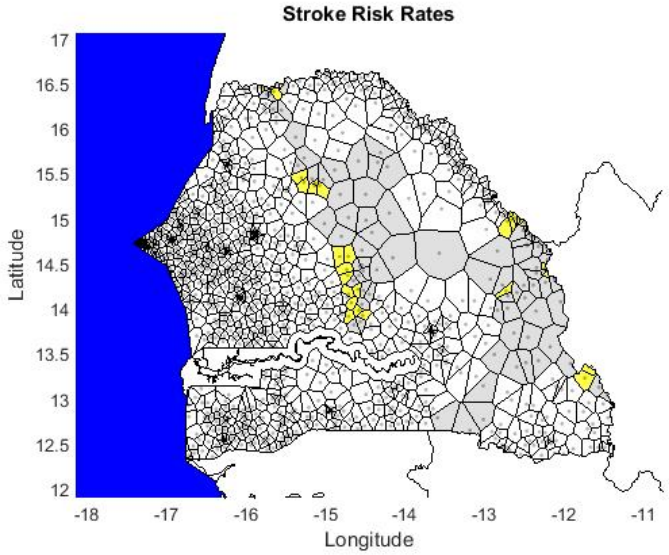

FIGURE 4.6 – Estimated incident cases of strokes from different areas too far from the nearest hospital to be treated by balloon (Grey $\prec$ 5 victims, Yellow 5 to 25).

Limitations of the study

The study presents some limitations due to bias introduced by the data used and some choices for designing the study. The first bias is related to the extrapolation of population at a given antenna coverage area as it is not possible to estimate precisely the Orange Sonatel share market in Senegal during the period covered by the provided data.

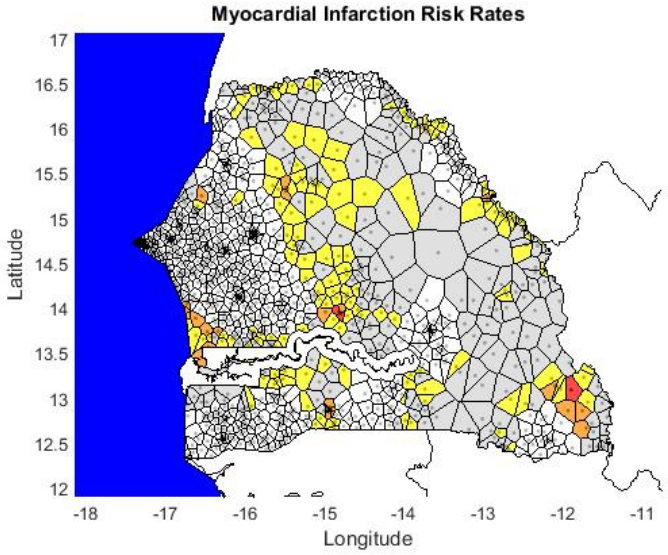

FIGURE 4.7 – Estimated incident cases of Myocardial Infarction from different areas too far from the nearest hospital to be treated by fibrinolysis (Grey $\prec$ 5 victims, Yellow 5 to 25, Orange 26 to 50, Red 51 to 100).

In addition, a filtering on data is performed by the mobile operator as indicated in section 4.2.3. Another bias which may affect the results is related to the computation of unique users per day for a given antenna site. We did not performed the estimation based solely on the users during night, which is likely to be more accurate. Eventually, we based the study on an estimated incidence rate of the considered medical emergency as there was no official figures available.

4.2.5 Conclusion

Stroke and Myocardial Infarction are two majors issues in Public Health. Several factors contribute to the increasing of the number of the cases annually. This is particularly true in Low and Middle income countries such as Senegal. We have shown that by using a considerable amount of CDR data (several hundreds million of tuples) integrated with census data and other contextual data, it is possible to perform location based estimation of the people in risk due to the lack of suitable infrastructures. The study confirms the potential of anonymised mobile phone CDRs for health research. This will help Public Health decision makers in their early stage decision making. The use of aggregated data is a safeguard but also presents a further limitation particularly for regulatory approvals (JONES et al., 2018). In this frame, Global Pulse [10], a flagship innovation initiative of the United Nations Secretary-General on Big Data, is pushing towards the access of useful data and tools to the discover of new uses of big data for development.

10. https ://www.unglobalpulse.org/

4.3 Bringing Semantic Technologies to Digitising Herbal-based Traditional Medicine

4.3.1 Introduction

Still in the ICT4D domain, and as the continuation of a research work initiated in collaboration with Bernard K Foguem from the ENIT engineering school of Tarbes (France) on a first work to contribute to the digitising of herbal-based African Traditional Medicine (ATM) (Kamsu-Foguem et al., 2013) using conceptual graphs (Sowa, 2008), we have initiated a more specific work on bringing semantic technologies for the formal representation of plant-based Traditional Medicine in West Africa, in collaboration with the West African Health Organisation (WAHO) together with domain experts from this region.

Traditional Medicine (TM) is the oldest form of healthcare delivery and is strongly established in many parts of the world, as a mainstay or a complement to conventional medicine for disease prevention and treatment. Its use is also rapidly spreading in high-income countries (HIC), where over 50% of the population have used complementary or alternative medicine at least once. The global market for herbal medicine currently stands at over US$ 60 billion annually and is progressing steadily [11]. Recently, the WHO TM Strategy 2014–2023(WHO, 2013) has recognised the growing use of TM and identifies need for harnessing the potential contribution of TM to health, wellness and people-centred healthcare. The strategy also seeks to promote the safe and effective use of TM by regulating, researching and integrating TM products, practitioners and practice into the formal health systems, where appropriate.

The high usage of TM is often driven by the inaccessibility, unaffordability or unavailability of conventional health care services and medicines in socioeconomic settings that are characterised by a high rate of poverty and a lack of suitable and affordable conventional medicine services and drugs. That underserved and mostly illiterate rural people account for the majority of the population, is an additional barrier that makes access to healthcare difficult. In response to the growing recognition of the potential of TM, the supra-national West African Health Organisation (WAHO) has given priority to TM in 2007, with the objective of supporting the institutionalisation of ATM in member countries' health systems, followed up by WAHO's 2016-2020 Strategic Plan. Within this plan, an important action item is the standardisation of descriptions of herbal and TM. Together with the lack of computable TM data, it is difficult to take benefit from them for primary and secondary use cases : patient follow-up and public health statistics, phytovigilance about herbal medications that are administered by traditional practitioners and used among the population, etc. An important step was the launch of the first edition of the West African pharmacopoeia in 2013, with inputs from ATM experts coming from different member states (WAHO, 2013).

In this study we aim at bringing Semantic Web technologies and formal knowledge representation in order to fulfil the following objectives :

— contribute to preserve the secular knowledge about TM in West Africa and bringing it to the digital world ;
— describe formally and non ambiguously TM entities and enabling their mapping to entities from conventional medicine ;
— acting as a support resources for training WAHO member states traditional practitioners and health professionals.

11. https ://www.marketresearchfuture.com/reports/herbal-medicine-market-3250

— enabling TM resources annotation and retrieval
— enabling support for knowledge gathering in order to perform later phytovigilance activity.

Initiatives to make health related information available to populations by using semantic web technologies have already been undertaken in other regions. One noticeable example is the HealthFinland case (Suominen et al., 2009).

4.3.2 African Traditional Medicine : some observations

Large Use and Potential of Herbal-based Traditional Medicine and lack of infrastructures

Africa is one region where TM usage has a significant foothold. The WHO estimates that as much as 80% of the population uses ATM in some form (Antwi-Bafour et al., 2014). In Mali, the first line of treatment for 60% of children with high fever resulting from malaria is the use of herbal medicines at home [12]. The high ATM usage is mainly driven by the inaccessibility, unaffordability or unavailability of conventional healthcare services and medicines. As an example, the cost of Abacavir®, a drug used in antiretroviral therapy, is about $ 602.66 per month [13] while in 2015, the Gross Domestic Product (GDP) reached $ 723 per capita in Mali and $ 531.3 per capita in Guinea according to figures from World Bank [14].

Beside, even when available at a reasonable price, the distribution of products is limited by the low density of pharmaceutical personnel and/or medical infrastructures. As an illustration, WHO estimates that the density of pharmacists was 21 per 1,000,000 people in Côte d'Ivoire in 2008 and only 9 per 1,000,000 people in Mali in 2010 [15]. Globally, the WHO considers that still over 40% of the population in Sub-Saharan Africa lack access to essential medicines, even in private health facilities [16]. By contrast, ATM being based on plants, plant compounds, minerals, or animals that are sourced locally, it benefits from an easier accessibility to remedies, a larger availability and a better affordability. In addition, the strong ties to local traditions, indigenous knowledge, and close proximity to tradi-practitioners increase the confidence of patients and facilitate the follow up. ATM can therefore be considered as a key enabler in the formidable quest of extending healthcare services in Africa and provide some relief to overstretched health systems. The role played by the West African Health Organisation (WAHO) in enhancing ATM in the Economic Community of West African States (ECOWAS)(please see the members on Figure 4.8) makes the region particularly relevant for applying innovative methods stimulating the access and use of ATM by the local population.

Lack of Formalisation and Regulation which Raises Safety Issues

Conventional medicine has increasingly benefited from advances in knowledge formalisation, organisation and translation, as well as from the rising adoption of semantic and terminological standards. Examples are coding systems for diseases, the International Classification of Disease (WHO-ICD) and drugs (WHO-ATC), ontologies and controlled

12. http ://www.who.int/mediacentre/factsheets/2003/fs134/en/
13. https ://tinyurl.com/qczpp96
14. http ://data.worldbank.org/indicator/NY.GDP.PCAP.CD ?locations=ML
15. http ://apps.who.int/gho/data/node.main.HSHRH ?lang=en
16. http ://www.who.int/medicines/mdg/MDG08ChapterEMedsEn.pdf

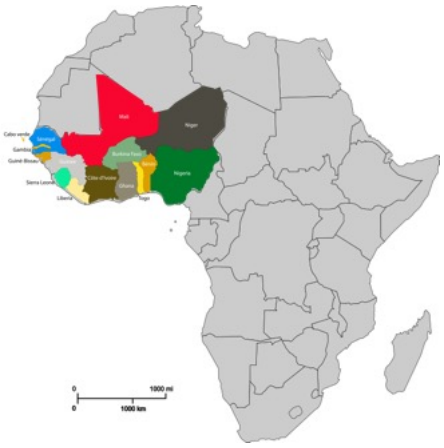

FIGURE 4.8 – Economic Community of West African Member States.

vocabularies (SNOMED-CT), information models (HL-7, openEHR), as well as clinical practice guidelines. In contrast, the practice of ATM (and TM in general) happens outside of any standardised knowledge management framework and remains largely unaffected by regulatory conditions that guarantee safety, quality and efficacy, as highlighted by the WHO. Indeed, ATM knowledge is mostly tacit and therefore largely shared in informal manner and disseminated orally with the risk to lose or alter crucial medical information. Moreover, although the raw material is usually based on plants and minerals, ATM remedies remain biologically active agents and can induce negative or dangerous effects if used inappropriately (TCHACONDO et al., 2011). Yet the lack of clinical follow-up on their use prevents monitoring of possible Adverse Drug Reactions (ADR) and raises critical safety issues associated to the practice. Furthermore, environmental pollution, varying medicinal properties (e.g., of same plants in different locations), misidentifying and adulteration of remedies constitute reasons for concern.

Lack of cooperation between conventional and traditional medicine

Despite the widespread usage of ATM in Low- and Middle-Income Countries (LMIC), the great potential offered by conventional medicine as regards to the understanding of pathologies, symptoms and mechanisms and the evident complementarity between the two practices, both systems co-exist in parallel and confine their interactions to the bare minimum so far. The integration of ATM into formal primary healthcare systems and national public health strategy is therefore still a gap to fill. The two segments would yet benefit from a collaboration that would result in a fruitful mutual learning and transfer of knowledge on the mechanism of action of molecules from specific plants (about 30% of modern medicines are made from plants that were first used traditionally (STROHL, 2000)).

To contribute overcoming these issues in the context of West Africa, we have launched the design and development of the West African herbal-based (WATRIMed) Knowledge Graph (KG) by reusing exiting materials provided by authoritative (WAHO) agencies and domain experts. A KG mainly describes real world entities and their interrelations, organised in a graph, dines possible classes and relations of entities in a schema and covers various topical domains (PAULHEIM, 2016).

An initiative dedicated to the digitisation of TM using Semantic Web technologies has already been taken in Asia. The Institute of Information on Traditional Chinese Medicine

(TCM) from the China Academy of Chinese Medical Science has been maintaining the Traditional Chinese Medicine Database System since 1984 [17]. It currently involves 40 sources related to TCM, including the TCM Literature Analysis and Retrieval Database. As well, a semantic e-Science infrastructure for TCM has been developed in order to support large-scale database integration and service coordination in a virtual organisation (H. Chen et al., 2007). It exploits domain ontologies to integrate TCM database resources and services within a semantic framework and delivers a semantics-based functionality including browsing, searching, querying and knowledge discovering to users. More recently, a knowledge graph for TCM Health Preservation, which integrates terms, databases and other resource has been constructed in order to facilitate TCM knowledge sharing (Yu et al., 2017).

4.3.3 The WATRIMed Knowledge Graph

The digital representation of health information is a critical aspect, which allows different health systems and organisations to interoperate. Unfortunately, the fact that ATM knowledge is mostly shared orally and the lack of documentation in ATM practice makes it difficult to classify and map ATM-related knowledge.

The WATRIMed effort aims at bringing West African TM to the digital world similarly to previous attempts of digitising the previously mentioned Traditional Chinese Medicine(Z. Wu et al., 2008) and more general African TM (Kamsu-Foguem et al., 2013 ; Lô et al., 2017), using a state-of-the art, flexible and shareable knowledge representation approach.

There is an increasing consensus that medical KR should use shareable standards for enabling computational support of data management. The converging of tools and methods is opposed to the richness of domains, concepts, and especially domain terms in a multitude of languages. A broad account for health knowledge representation should therefore be able to formalise knowledge using the following KR assets as basic building blocks (cf. Section 1.4.1 : a) concepts ; b) individuals, i.e. tangible, non-repeatable entities ; c) terminologies, i.e. units of human language, denoting entities from a) or b).

— a) Concepts (aka types, repeatables), i.e. language-independent entities of meaning that normally extend to classes of individual things ;
— b) Individuals, i.e. tangible, non-repeatable entities ;
— c) Terminologies, i.e. units of human language, denoting entities from a) or b).

Glued together by standardised languages (e.g. RDF, SKOS, OWL) and principles (e.g., linked data), the resulting propositions ideally result in shareable, interoperable, and computable KR assets. In addition, in order to facilitate this share and interpretation, the resulting propositions can be linked to an upper level Ontology (Mascardi et al., 2007). This is the choice made with WATRIMed.

4.3.4 Materials and Resources

The West African Herbal Pharmacopoeia gathers information on medicinal plants used in West Africa, building on a first African Pharmacopoeia including 105 plants created in 1985, followed by a book on medicinal plant analysis in 1986. It describes every plant by the following features : a summary description of the plant, its ethno-medicinal usage,

17. http ://cowork.cintcm.com/engine/windex1.jsp

related clinical information and safety, its chemical constitution, contraindications, the regions where the plant grows, a photograph, information on biological and pharmacological activity, and possible dosages and mode of administration.

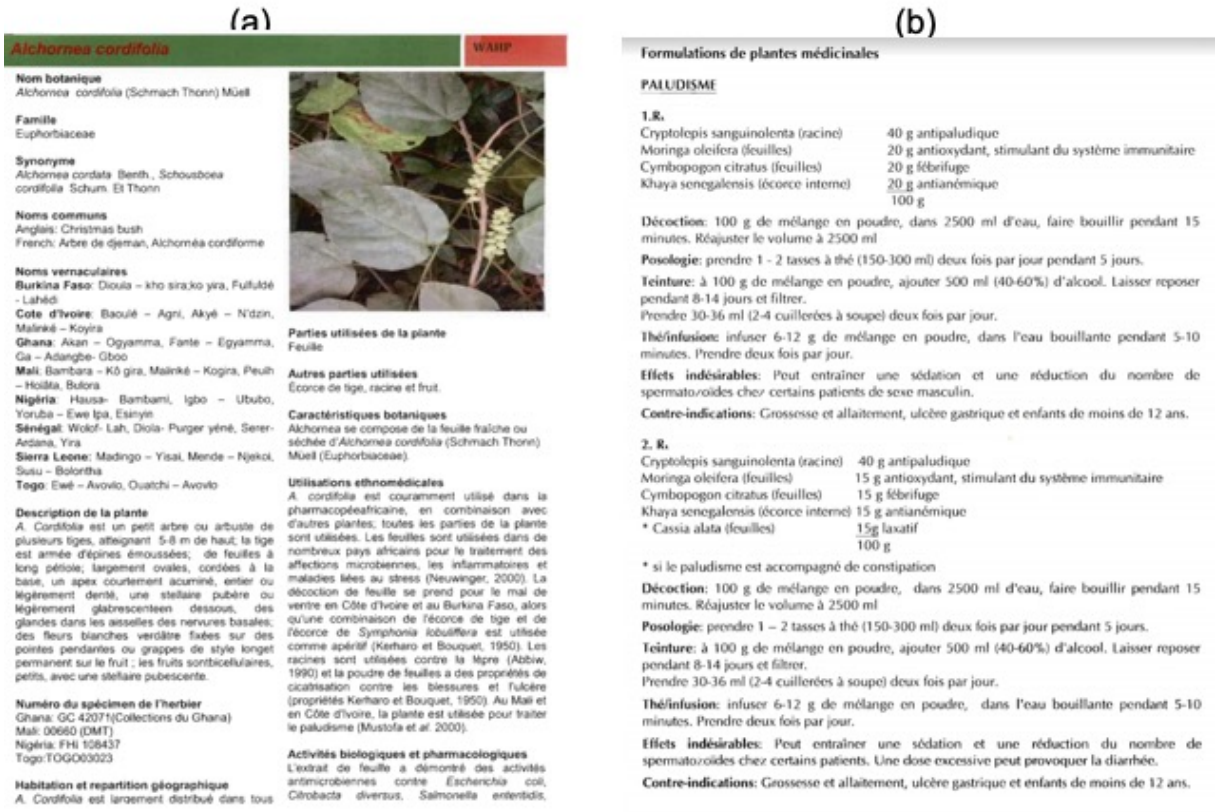

FIGURE 4.9 – Extract of the WAHO pharmacopoeia content (a), and an extract of the Inventory of Proven Effective Medicinal Plant Formulations (b).

WAHO also provides the "Inventory of formulations of medicinal plants with proven efficacy" resource (Figure 4.9 (OOAS, 2013). It includes 103 formulations (recipes) which are used to treat 53 diseases encountered in the West Africa region.

In addition, with the goal of building a KG by linking the WAHO herbal pharmacopoeia with TM knowledge, we identified the following set of publicly available Knowledge Bases (KBs), which allow to enrich the core information and to widen its scope while opening the perspective of wide-scale integration :

- DBpedia for Plants and Diseases, a crowd-sourced community effort to extract content from Wikipedia and by deploying this information on the Web though Linked Data (AUER et al., 2007).
- STITCH and PubChem for chemical compounds information. STITCH integrates information about interactions from metabolic pathways, crystal structures, binding experiments and drug-target relationships (KUHN et al., 2007). It enables understanding interactions between proteins and small molecules. PubChem is an open chemistry database aggregating information on chemical structures, chemical and physical properties, etc. (KIM et al., 2016).
- IPNI for plants names and bibliographic references. The International Plant Names Index (IPNI) is a database of the names and associated basic bibliographical details of seed plants, ferns and lycophytes (PENEV et al., 2016).

— GeoNames for information about countries and regions. It covers all countries with over 11 million place names.
— Wikidata and Yago for local dialects and vernacular names of plants and recipes. Wikidata is a free and open knowledge base which stores structured data from Wikipedia, Wikivoyage, Wikisource, etc... Yago is a knowledge base derived from Wikipedia, WordNet and GeoNames (Suchanek et al., 2007) with more than 10 million entities (like persons, organisations, cities, etc.) and contains more than 120 million facts about them.

4.3.5 Process of Building the WATRIMed Knowledge Graph

The workflow to build the WATRIMed Knowledge Graph comprises four main components (Figure 4.10) :

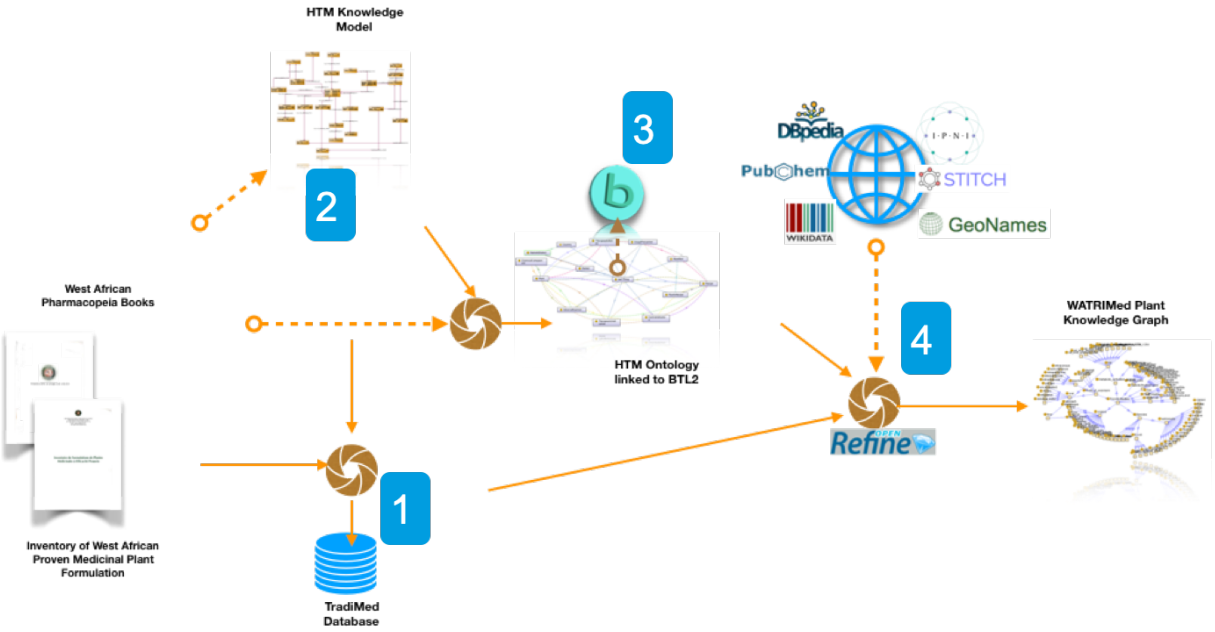

Figure 4.10 – Workflow of the WATRIMed Knowledge Graph building

1. Designing and Feeding the TradiMed Database (part 1 of Figure 4.10). The WAHO Herbal Pharmacopoeia and the WAHO Inventory of WA Proven Medicinal Plant formulations are manually looked up (as they were not available in an electronic processable format) for extracting the relevant information. In addition, the input sources include the WAHO training manual for traditional healers on the six (06) priority diseases in West Africa : HIV/AIDS, tuberculosis, sickle-cell anaemia, malaria, diabetes and hypertension. These extracts are fed into a relational database (TradiMed).
2. Conceptual Designing of the HTM Ontology, with the identification of the main TM entities to handle (part 2 of Figure 4.10). Starting from the resources made available by WAHO and inputs with local experts in TM, a model formally described in the Description Logics formalism (Baader et Nutt, 2003) and represented in the OWL language is designed as the core of WATRIMed and then implemented

using the Protégé editor [18]. It formally describes medicinal plants, their components (chemical compounds and plant parts) and recipes. It could be seen as manual design of an Ontology from existing resources.

3. Mapping the core entities of the HTM Ontology to a foundational Ontology (part 3 of Figure 4.10). In comparison with a domain Ontology, like the HTM Ontology, a foundational (or upper level) Ontology consists in very general categories that are common for a large set of domains. As mentioned earlier, it contains entities that can be used to bridge the knowledge represented by domain-specific ontologies. A number of foundational ontologies have been proposed (Mascardi et al., 2007). We have chosen to align the HTM Ontology with the BioTopLite2 Upper level Ontology (BTL2) (Schulz et Boeker, 2013). BTL2 is a light version of the BioTop Upper level Ontology for the Life Sciences (Beisswanger et al., 2008). BioTop has been launched in 2006 and described in OWL DL. For classes, BioTop inherits the top-level distinction of the Basic Formal Ontology (BFO) (Arp, Smith et Spear, 2015) between the classes *Continuant* and *Occurrent* and between *Independent Continuant* and *Dependent Continuant*.
4. Linking HTM entities with external relevant entities (part 4 of Figure 4.10) from the identified KBs and described in the Section 4.3.4. The aim is the acquisition of additional information for the WATRIMed entities and to establish bridges with entities from conventional medicine as described in publicly available KBs. This process relies on the OpenRefine tool [19]. It offers easy access to external web services and data.

Each WATRIMed Ontology unit is assigned a human language names. Further, the HTM model enables keeping different vernacular (local) names for denoting plants in different regions.

4.3.6 Current WATRIMed Results

The WATRIMed model and Knowledge Graph, together with a SPARQL endpoint are freely accessible online [20].

The Herbal based Traditional Medicine Knowledge Model and the WATRIMed Knowledge Graph

The Herbal-based Traditional Medicine knowledge model, referred as the HTM Ontology comprises 1,129 Concepts and 75 Properties (57 Object properties and 18 Data properties). The main component is the *MedicinalPlant* concept. It is linked with the *ChemicalCompound* entity by the object property *hasChemicalComponent*. A *MedicinalPlant* has a set of naming in different vernacular names. For instance, the *Moyatabél* vernacular name in Burkina Faso's *Fulfulde* of the plant *Alstonia boonei* is described with the following complex expression (assuming that *wat* is the prefix of WATRIMed and *btl2* the BioTopLite2 one) :

A *Vernacular Name* is part of *Vocabulary* which represents the dialects in which the vernacular names will be given. It should be noted that a given dialect can be spoken in several countries.

18. https ://protege.stanford.edu/
19. http ://openrefine.org/
20. http ://www.watrimed.org

wat :Moyatabél **type of** wat :'vernacular name'
and (btl2 :'is part of' **some** (wat :Vocabulary
and (btl2 :represents **value** wat :Fulfulde)
and (btl2 :'is participant in' **some**
(wat :Usage and btl2 :'is included in' **value** wat :'Burkina Faso')
and (btl2 :represents **only** wat :Alstonia-boonei)

The Mapping to the BioTopLite Upper Ontology

The mapping of the classes provided the following results, with respectively *btl2 :* and *wat :* the namespace prefixes of BioTopLite2 and WATRIMed :

— **Simple subclass mappings** was done for the following WATRIMed concepts : *wat :'Adverse Reaction'* under btl2 :process, *wat :'Chemical Component'* under *btl2 :compound*, *wat :Vocabulary* under *btl2 :information object* as well as the *wat :'Vernacular name'*. The *wat :MedicinalPlant* concept is under *btl2 :organism*, *wat :'Plant Part'* under *btl2 :organism part*, and *wat :Recipe* is subclass of *wat :Therapeutic Mixture* which is under *btl2 :compound of collective material* entities.

— **Complex subclass mappings** are used for *wat :Recipe* and *wat :'Chemical Compound'*. Thus, for the *wat :'Acacia nilotica'* concept for instance, we have the following expression :

wat :Acacia-nilotica **SubClassOf**
(btl2 :'has part' **some** wat :'Arabic acid')
and (btl2 :'has part' **some** wat :'Chlorogenic acid')
and (btl2 :'has part' **some** wat :'Gallic acid')
and (btl2 :'has part' **some** wat :Leucoanthocyanidin)
and (btl2 :'has part' **some**
wat :3-beta-acetoxy-17-beta-hydroxyandrost-5-ene)

— **No mappings** for *wat :UsagePrecaution* and *wat :ContraIndication*.

The mapping of relations does not consider the *dispositional* relations Arp, Smith et Spear, 2015. The WATRIMed relations are mapped as follows : the *wat :'has chemical component'* property is sub property of *btl2 :'has part'* ; *wat :'derives from'* is mapped as a sub property of chaining properties (*btl2 :'has part of' btl2 :'at some time' o btl2 :'is part of'*).

Linking WATRIMed and External Publicly Available Knowledge Bases

The following entities of the HTM Ontology have been linked to external resources identified among the external publicly available KBs : *MedicinalPlant, TheurapeuticIndication, ContraIndication, ChemicalComponent* and *Vocabulary*. Currently there are 115 MedicinalPlant respectively linked to 100 DBpedia entities and 100 IPNI resources. Setting up these external links enabled to enrich the description of the plants, because the information provided by the two KBs is complementary. For *TherapeuticIndication*, about 40% of them are linked to DBpedia entities (42 out of 110). However, only 6 out of 110 could be linked to some Yago entity. Eighteen *ContraIndication* entities have been linked to Yago entities (12%). All the *ChemicalComponent* entities have been linked to external resources by fetching URLs from STITCH and PubChem. We have identified 13 out of 122 links for Vocabulary with Yago entities and 46% (56 out of 122) links with Wikidata.

4.3.7 Discussion

For many people in Africa, Traditional Medicine is either the first line of treatment or is used as a last resort when all the available possibilities in the conventional medicine are exploited. Despite its affordability, it comes with various issues, in particular due to the oral transmission of knowledge and lack of digitised resources that could contribute to improve the sustainability of experiences gathered. We have designed and made available the first release of the WATRIMed Knowledge Graph. Some ideas to go further and issues to be further addressed and investigated on WATRIMed :

From Linking to External Knowledge Bases perspective The automated linking of the HTM instances (through the TradiMed Database) and entities from external, publicly available KBs relies on column names with a terminological similarity look-up. The matching to establish between two given entities depends on their lexical similarity, which could not be sufficient in case of synonymy for instance. This requires an in-depth human validation process. So far, we have manually checked the established correspondences. To illustrate the difficulty of the automated matching process, only 12% of the *ContraIndication* have been linked to external entities. There is a difference between the KB strategies in matching entities with OpenRefine : for instance, while for Yago it is quite strict (exact match), correspondences identification is more relaxed with DBpedia.

From Linking to an Upper Level Ontology perspective There are several rationales for rooting a domain ontology in a foundational ontology. First, precisely defined classes and relations reduces the ambiguity of domain terms. In our case, *Recipe* is placed under *btl2 :'material object'*, which precludes its interpretation as an information entity. Second, it precludes modelling errors : if *Recipe* for instance were put under *btl2 :'information object'* and linked with its material ingredients via *btl2 :has part*, it would contradict an upper level axiom. Third, ontologies that share a common upper level, are more suited to be reused in other contexts. This is in line with the increasing characterisation of ontologies as standards (with SNOMED CT as example) and addresses the FAIR criteria for scientific data stewardship (Wilkinson et al., 2016). Forth, building new domain Ontology as an extension of a foundational ontology speeds up Ontology building and maintenance. Without an explicit foundational Ontology, the authors will follow their own implicit upper level models, which heavily depend on the use case and are not shared with those who have to use and maintain it. On the downside is the tendency towards more complexity, especially regarding nested axioms, which at least partly can be compensated by simplifications, e.g. by using new object properties as relation chains. Further simplification steps might be necessary when the ontologies are used in large knowledge graphs, the performance of which might be affected by overly complex OWL models.

4.4 Conclusion

WATRIMed initiative is the first large-scale attempt to provide a formalised knowledge about ATM in the context of West Africa. It benefits from decades of experience gathered by the West African Health Organisation, which promotes and contributes to regulate TM usage among its member states. The aim is to provide a fully integrated digitised and semantically explicit resource to the Linked Open Data cloud. Currently the provided

graph contains about 30,000 facts. The current content has recently been validated by an expert in the field by comparing the content with findings reported in scientific literature for each plant studied.

From the current resource, it could be observed that the most used Medicinal Plants is *Moringa oleifera* followed by *Alchornea Cordifolia* and *Khaya Senegalensis*. Malaria, cough and diarrhoea are the most treated diseases and symptoms by these three Medicinal Plants. Haussa, Ewe and Bambara are the most common local dialects encountered in naming medicinal plants. Leaves, Stem Barks and Roots are the most common plant parts used in traditional recipes.

It is envisioned to progressively harvest and integrate medicinal plant related data from additional input resources made available by public bodies (European Medicinal Agency in Europe, Food Drug Administration (FDA) in the US, etc.) and other publicly available knowledge bases (Super Natural II (Banerjee et al., 2015), ChEBI [21], OMIM [22], etc.).

21. https ://www.ebi.ac.uk/chebi/
22. https ://www.omim.org/

Chapitre 5

General Conclusion and Research Statement

Contents

5.1 General Conclusion

In this manuscript, I have described the contributions of my work in recent years on the different aspects related to healthcare knowledge representation, large scale biomedical textmining and knowledge integration. The data that have been reported in this work are rather mainly those from sources other than those from EHRs (scientific literature, social networks - web forum posts, Call Data Records). An underlying marker that has remained in filigree throughout these described works is the semantic aspect that allows to formally encode the manipulated information.

For the processing of unstructured textual documents, the approaches implemented are based on the use of semantics encoded through different KOS, mainly structured vocabularies or terminologies. Semantic indexing and annotation is vital for biomedicine and healthcare related documents classification and retrieval. The biomedical literature mining reported here relies on statistical model which exploits information related to word frequency. Recent advances in AI and artificial neural network, which allows representing more semantic information, deserve to be investigated in this context.

The work describing the contributions related to the large-scale alignment of ontologies has highlighted the value of using techniques to optimise the complexity inherent to large semantic resources in the biomedical field in order to scale up. The K-Ware framework which enables the integration of heterogeneous KOS complements existing biomedical ontologies repository-based solutions.

The analysis and integration of CDR data in a context of limited resources (low and middle income countries) has also shown that it is possible to use data collected in a passive way, and which are not de facto health data, as alternative data to be used to answer public health questions. However, CDR data raise also some ethical issues. Indeed,

the level of detail within the data can lead to the possibility of de-anonymisation, which can bring up particularly high concerns in the context of the General Data Protection Regulation.

The different algorithms and tools that have been developed throughout this research have been tested in the context of international challenges on standard benchmarks with promising results compared to related work.

It can be observed that with the development of ICTs and the Web of data, the opportunity is offered to have data from alternative and complementary sources that were hitherto marginal, and public knowledge bases that can be integrated with traditional data (such as those from Electronic Health Records) to identify the patient in his/her complexity.

The current enthusiasm for artificial intelligence, particularly in the service of health, and the creation of the data catalogue through the health data hub initiative in France, constitute a favourable ecosystem for interdisciplinary work combining health and IT. This suggests that collaborations can be expected to produce algorithms and tools to better process the complex data that will become increasingly available and allow health professionals to have more time to devote to care and public health decision-makers to make decisions based on better quality indicators.

There is still a long way to go. From my personal perspective, I will now elaborate more on the work that is being initiated or is envisioned as of perspective to my research.

5.2 Research Statement

In the following sections, I outline the different research directions as a continuation of my current research work.

5.2.1 On jointly handling heterogeneous knowledge resources and their alignments

We have described in Chapter 3.1, the work we have pursued on generating alignments between different related KOS, particularly for large biomedical semantic resources (Ba et Diallo, 2013) and (Diallo, 2014). This work contribute to facilitating heterogeneous data integration at the semantic level.

As part of the **PhD thesis of Bruno Thiao-Layel**, and as a natural continuation of the former work, we have also described the research work conducted in the context of the K-Ware framework and the generic metamodel that it provides for defining at a high abstract level the notion of KOS (Thiao-Layel et al., 2016, 2018). The metamodel is compliant with the RDF recommendation of the W3C. This allows handling semantic resources described in SKOS, RDFS or OWL. Each KOS is described as an *rdfs :Resource* where any entity is an instance of the *rdfs :Class* predicate. This metamodel provides, in particular, the advantage of preserving the original semantics of the resources to be managed, compared to other existing solution based on biomedical semantic resources repositories (Grosjean et al., 2014 ; Noy et al., 2009).

The first avenue to be explored will be on the Ontology merging or integration (Pinto et J. P. Martins, 2001) domain in the context of K-Ware to provide so called Ontology modules. Indeed, as the framework is able to handle heterogeneous KOS, the idea is to be able to combine either entire managed KOS, whatever their type (formal Ontology

described in OWL, structured vocabulary in SKOS), or a portion of them thanks to the use of the mappings relationships defined at the abstract level in K-Ware (Thiao-Layel et al., 2016).

The second direction to explore is to address the challenge of being able to access, at a semantic level, the concrete data available in an information source. In a more precise way, referring the abstract notion encoded in KOS to access to the actual data as represented in different sources. For instance, a medical report can reports that a patient has a high fever while in a database entry about body temperature can be reported as 39°. Semantically they refer to the notion, the fact to have a fever.

The work to be conducted in this frame is to explore the enrichment of the actual K-Ware metamodel on order to be able to take into account the link between an actual data in an information source with the semantic entity from a particular KOS which describe it. This issue will be addressed toward investigating into the exploitation of the possibilities offered by the domain of Master Data Management, particularly metadata registries or upper level ontologies like the Information Artifact Ontology (IAO) (Ceusters, 2012).

A metadata registry according to the Dublin Core Metadata Initiative, like the ISO/-CEI 11179 standard, is a formal system that provides authority information on the semantics and structure of each element in an information source. For each element, the register shall define it, the qualifiers associated with it and correspondences with equivalents in other languages or other schemes. Similarly, The Information Artifact Ontology (IAO) was created to serve as a domain-neutral resource for the representation of types of information. An information artifact can be typically a database entry, a medical report or scientific publication.

This area of work has direct practical application in the context of the management of heterogeneous drugs information at Synapse Medicine, on the one hand, and in the context of formulating complex requests, at the end user level, on the clinical data warehouse deployed at the Bordeaux University Hospital in the other hand. A similar idea in a limited context, in order to identify anticoagulant indicators (PREPS PACHA project) has already been set up as part of the internships of Hadrien Sentenac (I co-supervised with Vianney Jouhet) in 2018 on this clinical warehouse.

5.2.2 Dictionary and Deep Learning based approach for unstructured information extraction from clinical DataWarehouse

As part of the continuation of **Sébastien Cossin's PhD thesis** work jointly supervised with Vianney Jouhet, the idea is to capitalise on the work carried out in the context of large biomedical textmining and the experience gained in identifying in particular medical concepts on unstructured textual documents, like drugs mention identification with the Romedi approach (Cossin, Lebrun et al., 2019), in order to examine outcome prediction from clinical texts. Indeed, clinical notes associated with a patient encounter contain rich information about the entirety of the admission (Shickel et al., 2018). And a clinical DataWarehouse is being implemented within the Bordeaux University Hospital. This clinical DataWarehouse comprises data from more than 1.5 millions patients. These data are harvested from the traditional Hospital Information System data sources. A significant part of the data consists of unstructured data in textual form (clinical notes, discharge summaries, etc...) in French. It is common that the same type of note can appear very differently depending on its author, due to various shorthand abbreviations, ordering pre-

ferences, and writing style (Shickel et al., 2018). Handling them properly in order to extract valuable information is an important issue, particularly when they are expressed in the French language due to the lack of resources for performing efficient NLP tasks (Névéol et al., 2018).

The idea is therefore to take benefit from the availability of huge amounts of clinical data within the DataWarehouse as well as health expertise from the users from whom learning strategies could be drawn. Thus, The approach to be pursued will be based on a combination of dictionary-based approaches, such as the one implemented in the context of drugs mention identification with Romedi, with deep learning approaches for computation phenotyping but also for the automated filling of case report forms (eCRF). The latter case will facilitate collecting data from the DataWarehouse for data reusing in the context of clinical research.

5.2.3 From Raw Materials to Knowledge Graph Enrichment and Exploitation

The continuation work on WATRIMed and knowledge graph acquisition will be pursued towards the directions that are detailed in the following sections.

Enriching WATRIMed with iconic languages for illiterate people

Tradi-practitioners and illiterate people could benefit from the content of WATRIMed. Enhancing support to these layman and illiterate could also mean offering the possibility to express complex concepts and notions with a simple visual-based language. I would like therefore investigate into an iconic visualisation language as a communication channel for tradi-practitioners and plant-based medication patients. A similar approach has already been used in the context of the conventional medical domain with the Visualisation of Concepts in Medicine (VCM) language (Griffon et al., 2014; Lamy et al., 2008). It is constituted about thousands of easily understandable icons, representing abstract concepts in medicine including signs, symptoms, physiological states, surgical procedures and drugs. For example, an icon can represent the status of the patient, and the colour of the icon indicates the temporality of the state. We envision to adapt and extend such a language to the plant modelling domain in order to enable iconic representation of plant related entities (bark, plant, leaf, symptoms, etc.) and enrich the WATRIMed knowledge graph.

Knowledge Graph-based Computational Drug Repositioning

Drug discovery is a time-consuming, high-investment, and high-risk process. As of consequence, drug repositioning or repurposing, also known as old drugs for new uses, has become a popular strategy in recent years. On average about 1-2 years is needed to identify new drug targets and 8 years to develop a repositioned drug against 10-15 years to develop a new drug (Sertkaya et al., 2014). Particularly, drug discovery strategies based on natural products and traditional medicines are re-emerging as attractive options. Indeed, natural products have often proved to be a critical starting point in drug design (Banerjee et al., 2015). It is estimated that approximately one-third of the top-selling drugs are derived from medicinal herbs (Strohl, 2000). Among the well-known examples of valuable natural products used widely in today's medical and animal health industries include lovastatin (anticholesterolemic agent) and paclitaxel and doxorubicin (antitumor

agents). From the traditional medicine perspective, the Hymenocardia acida medicinal plant for instance, which is widely used for its healing properties, has been the subject of extensive scientific investigations for the treatment of hypertension in Guinea. This lead to the Guinex HTA drug to treat high blood pressure.

Alongside biological experimental approaches, several computational methods for drug repurposing exist (e.g., (Alshahrani et Hoehndorf, 2018)). To address the drug repositioning issue, the idea is to rely on neuro-symbolic approach and combine multi-modal information (information mined from biomedical scientific literature, structured information from knowledge graph, public databases about target proteins and diseases, etc.) to predict drug targets and indications for known drugs and medicinal plants. To do so, the WATRIMed Knowledge graph will be enriched with drug-targets interaction from the SuperTarget database (Hecker et al., 2012). The latter integrates drug-related information associated with medical indications, adverse drug effects, drug metabolism, pathways and Gene Ontology (GO) terms for target proteins. Feature representations of entities will be learnt using Machine Learning approaches on relational knowledge graphs similarly to (Nickel et al., 2016).

Detecting Herb Drug Interactions using Artificial Neural Networks

In the context of the Horizon 2020 MSCA Marie-Curie Postdoc of Georgeta Bordea, the kANNa project (April 2019 - April 2021) that I supervise, we are investigating into the identification of interactions between herbs and drugs. Indeed, detecting early signals of possible adverse reactions and providing up-to-date support for decision making will be made feasible by automatically extracting facts about herb-drug interactions from a large number of scientific articles using recent development in Deep Learning, combining and linking them with rich information already available in general-purpose and domain-specific knowledge bases. The following topics will be investigated.

1. Integrating information extraction features into the process of monitoring herb-drug interactions from medical literature : kANNa aims to extract accurate, complete and up to date knowledge about herb-drug interactions from scientific publications. This work will investigate, implement, apply and evaluate a state of the art information extraction pipeline based on Deep Networks and ontologies to detect, classify, integrate and share herb-drug interactions.
2. Enhancing additional knowledge acquisition from sparse, incomplete, and unreliable evidence. Indeed, in scientific literature herb-drug interactions are not clearly identified but rather often implicit and inferred from additional knowledge. For example, in the case of co-administration of warfarin and St John's Wort the proposed mechanism is induction of CYP450 isoenzymes by constituents of St. John's Wort, but there is only limited clinical data available. Therefore we will investigate methods for enriching the extracted knowledge graph by learning missing relations.

The work will exploit public databases like DrugBank (detailed information about drugs and their targets) [1], and the WATRIMed knowledge graph, which will be enriched with information about drugs target as previously described, to perform knowledge graph completion (Paulheim, 2016) to predict missing relations and facts. Knowledge graph completion, which emerged in the context of constructing general-purpose knowledge bases such as DBpedia and YAGO, will allow to predict relations between entities by using

1. http ://drugbank.ca/

known relations in the graph for training, making use of relational learning algorithms that work well with the relatively smaller training datasets that are available in specialised domains.

5.2.4 Integrating CDR and Public Databases to address Public Health issues

Together with colleagues from ETH Zürich – Switzerland (Center for Development and Cooperation), and the Aalto University we envision to build on the previous research work we have done on the Call Data Records with two proposed objectives : the first goal is to improve the previous work by relaxing some of the assumptions we had to make and by using higher resolution input data. Our second goal is to extend the research by mapping additional risk factors and looking at further health conditions.

In our past study, we generated maps of people at risk for not receiving timely acute care during critical medical emergency conditions (myocardial infarction and stroke). To that end, we estimated travel times of the population to their nearest hospital as a health risk factor. The risk factor maps and corresponding numerical estimates for people at risk can be used to inform early stage decision making on the location of health centres, their equipment for handling different medical emergency cases and preventative strategies to address the risk factors for the condition in the population.

The proposed improvements include the following measures :

— time to care estimates can be refined by using fast travel times between antenna sites from the users fine grained mobility (Dataset 2) of the CDR data. Here insights from the approach described in (Kujala et al., 2016) can be utilised ;

— population density estimates can be improved using high resolution remote sensing data from the WorldPop project [2] .

— the location based estimates of incidences can be improved by combining the approach with a incidence model for stroke and myocardial infarction that utilises maps of risk factors. To this aim, data for risk factors like smoking, age, gender and diet taken from the Demographic and Health Surveys (DHS) will be used as input to a machine learning modelling procedure. DHS are conducted in several developing countries, usually in collaboration with the national statistical agency and other organisations [3].

These measures substantially improve the maps of the number and location of people at risk following stroke and myocardial infarction due to the risk factor of high travel times.

Furthermore we aim to apply the mapping approach towards further risk factors and additional conditions and diseases. For maternal mortality for example, risk factors include home birth (30% of mothers still deliver their newborns at home) and high travel times to a hospital. This results in increased mortality due to birth complications requiring time critical professional intervention. Risk factors for neonatal mortality include the vaccination and treatment seeking behaviour, with 25% of children not receiving all required basic vaccinations within the first 2 years of life and only 25% of newborns receiving treatment at health facilities for fever and/or diarrhoea.

We hypothesise that there is potential for using knowledge representation/machine learning methods and mobile data to estimate maps of these and other health condition/-

2. http ://www.worldpop.org.uk/
3. https ://dhsprogram.com/

disease risk factors using DHS surveys as ground truth. The idea has been demonstrated to be successful for some variables (Bruckschen et al., 2015) and work on mapping poverty using mobile phone and satellite data (Steele et al., 2017). The extension to other variables however still remains to be explored, constituting a gap in current research.

Health condition and disease risk factors could be modelled using an Ontology-based formal knowledge representation model. The advantage to this is to be able to reusing the mapping model in different contexts.

The generated maps can contribute to addressing a multitude of public health concerns. A high resolution map of immunization (DHS) can improve the logistics and priority areas for vaccination campaigns. A map for the probability to seek treatment for fever (DHS) can identify regions where health outcomes could potentially be significantly improved by an information campaign. Areas with long travel times to health facilities could be considered for the placement of new health facilities. Disease incidence models can be improved by higher granularity input on risk factors.

Bibliographie

Achichi, M. et al. (2017), « Results of the Ontology Alignment Evaluation Initiative 2017 », in : *OM : Ontology Matching*, Wien, Austria, p. 61–113 (cf. p. 68).

Agrawal, R. et E. L. Wimmers (2000), « A framework for expressing and combining preferences », in : *ACM SIGMOD Record* 29.*2*, p. 297–306 (cf. p. 9).

Aguirre, J.-L. et al. (2012), « Results of the Ontology Alignment Evaluation Initiative 2012 », in : *International Workshop on Ontology Matching (OM)* (cf. p. 61).

Ait-Kaci, H. et S. Amir (2017), « Classifying and querying very large taxonomies with bit-vector encoding », in : *J. Intell. Inf. Syst.* 48.*1*, p. 1–25 (cf. p. 18, 56).

Algergawy, A., M. Cheatham et al. (2018), « Results of the Ontology Alignment Evaluation Initiative 2018 », in : *Proceedings of the 13th International Workshop on Ontology Matching co-located with the 17th International Semantic Web Conference, OM@ISWC 2018, Monterey, CA, USA, October 8, 2018.* P. 76–116 (cf. p. 68).

Algergawy, A., S. Massmann et al. (2011), « A Clustering-Based Approach for Large-Scale Ontology Matching », in : *ADBIS*, p. 415–428 (cf. p. 54).

Allan, J. et al. (2003), « Challenges in information retrieval and language modeling : report of a workshop held at the center for intelligent information retrieval, University of Massachusetts Amherst, September 2002 », in : *ACM SIGIR Forum* 37.*1*, p. 31–47 (cf. p. 47).

Alshahrani, M. et R. Hoehndorf (2018), « Drug repurposing through joint learning on knowledge graphs and literature », in : *BioRxiv* (cf. p. 103).

ANSD (2014), *Rapport définitf du RGPHAE 2013*, rapp. tech. SEN-ANSD-RGPHAE-2013-V1.1, Dakar, Senegal (cf. p. 80, 83).

Antwi-Bafour, S. et al. (2014), « The Place of Traditional Medicine in the African Society : The Science, Acceptance and Support », in : *American Journal of Health Research* 2.*2*, p. 49 (cf. p. 89).

Aronson, A. R. (2001), « Effective mapping of biomedical text to the UMLS Metathesaurus : the MetaMap program », in : *Proc AMIA Symp*, p. 17–21 (cf. p. 30).

Aronson, A. R. et F.-M. Lang (2010), « An overview of MetaMap : historical perspective and recent advances », in : *Journal of the American Medical Informatics Association* 17.*3*, p. 229–236 (cf. p. 38).

Aronson, A. R., J. G. Mork et al. (2004), « The NLM Indexing Initiative's Medical Text Indexer », in : *Stud Health Technol Inform* 107.*Pt 1*, p. 268–272 (cf. p. 30).

Arp, R. et B. Smith (2008), « Function, Role, and Disposition in Basic Formal Ontology », in : *Nature Precedings* 1941 (cf. p. 11).

Arp, R., B. Smith et A. D. Spear (2015), *Building ontologies with basic formal ontology*, Mit Press (cf. p. 11, 94, 95).

Audeh, B. et al. (2017), « Vigi4Med Scraper : A Framework for Web Forum Structured Data Extraction and Semantic Representation », in : *PLOS ONE* 12.*1*, e0169658 (cf. p. 41, 42).

Auer, S. et al. (2007), « DBpedia : A Nucleus for a Web of Open Data », in : *Proceedings of the 6th International The Semantic Web and 2Nd Asian Conference on Asian Semantic Web Conference*, ISWC'07/ASWC'07, Berlin, Heidelberg : Springer-Verlag, p. 722–735 (cf. p. 92).

Aumueller, D. et al. (2005), « Schema and ontology matching with COMA++ », in : *SIGMOD Conference*, p. 906–908 (cf. p. 54).

Avillach, P., J.-C. Dufour et al. (2012), « Design and validation of an automated method to detect known adverse drug reactions in MEDLINE : a contribution from the EUâ€"ADR project », in : *Journal of the American Medical Informatics Association*, amiajnl–2012 (cf. p. 15).

Avillach, P., F. Mougin et al. (2009), « A semantic approach for the homogeneous identification of events in eight patient databases : a contribution to the European eu-ADR project », in : *Studies in health technology and informatics* 150, p. 190–194 (cf. p. 15).

Ba, M. et G. Diallo (2013), « Large-scale biomedical ontology matching with ServOMap », in : *{IRBM}* 34.*1*, p. 56–59 (cf. p. 17, 18, 100).

Ba, M. et G. Diallo (2012), « Knowledge Repository as Entity Similarity Computing Enabler », in : *Signal Image Technology and Internet Based Systems (SITIS), 2012 Eighth International Conference on*, IEEE, p. 975–981 (cf. p. 17).

Baader, F. et W. Nutt (2003), « The Description Logic Handbook », in : sous la dir. de F. Baader et al., New York, NY, USA : Cambridge University Press, p. 43–95 (cf. p. 19, 71, 93).

Baert, A.-E. et D. Seme (2004), « Voronoi Mobile Cellular Networks : Topological Properties », in : *Third International Symposium on Parallel and Distributed Computing/-Third International Workshop on Algorithms, Models and Tools for Parallel Computing on Heterogeneous Networks*, Cork, Ireland : IEEE, p. 29–35 (cf. p. 24, 82).

Baeza-Yates, R. A. et B. Ribeiro-Neto (1999), *Modern Information Retrieval*, Boston, MA, USA : Addison-Wesley Longman Publishing Co., Inc. (cf. p. 57, 59).

Balikas, G. et al. (2015), « BioASQ : A Challenge on Large-Scale Biomedical Semantic Indexing and Question Answering », in : *Multimodal Retrieval in the Medical Domain*, sous la dir. de H. Müller et al., t. 9059, Cham : Springer International Publishing, p. 26–39 (cf. p. 35).

Banerjee, P. et al. (2015), « Super Natural II—a database of natural products », in : *Nucleic Acids Research* 43.*D1*, p. D935–D939 (cf. p. 97, 102).

Bates, D. W. et al. (1999), « Patient risk factors for adverse drug events in hospitalized patients. ADE Prevention Study Group », in : *Arch. Intern. Med.* 159.*21*, p. 2553–2560 (cf. p. 21).

Bechhofer, S. et al. (2008), « Using Ontologies and Vocabularies for Dynamic Linking », in : *IEEE Internet Computing* 12.*3*, p. 32–39 (cf. p. 47).

Begaud, B. et al. (2017), *Les données de vie réelle, un enjeu majeur pour la qualité des soins et la régulation du système de santé : L'exemple du médicament.* Rapp. tech., Ministère de la Santé et des Solidarités. Paris. FRA (cf. p. 39, 40).

Beisswanger, E. et al. (2008), « BioTop : An Upper Domain Ontology for the Life Sciences : A Description of Its Current Structure, Contents and Interfaces to OBO Ontologies », in : *Appl. Ontol.* 3.*4*, p. 205–212 (cf. p. 94).

Berners-Lee, S. T. (2011), « Designing the Web for an Open Society », in : *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, New York, NY, USA : ACM, p. 3–4 (cf. p. 5, 10, 53).

Berners-Lee, T. et al. (2001), « The semantic Web », in : *Scientific American* (cf. p. 5).

Bi, W. et J. T. Kwok (2013), « Efficient Multi-label Classification with Many Labels », in : *Proceedings of the 30th International Conference on International Conference on*

Machine Learning - Volume 28, ICML'13, Atlanta, GA, USA, JMLR.org, p. III–405–III–413 (cf. p. 31).

Biebow, B. et S. Szulman (1999), « TERMINAE : A Linguistic-Based Tool for the Building of a Domain Ontology », in : *Proceedings of the 11th European Workshop on Knowledge Acquisition, Modeling and Management*, EKAW '99, London, UK, UK : Springer-Verlag, p. 49–66 (cf. p. 10).

Bigeard, E. et al. (2018), « Detection and Analysis of Drug Misuses. A Study Based on Social Media Messages », in : *Frontiers in Pharmacology* 9 (cf. p. 21, 41).

Blei, D. M. et al. (2003), « Latent Dirichlet Allocation », in : *J. Mach. Learn. Res.* 3, p. 993–1022 (cf. p. 31).

Bodenreider, O. (2004), « The Unified Medical Language System (UMLS) : integrating biomedical terminology », in : *Nucleic Acids Research* 32.*Database-Issue*, p. 267–270 (cf. p. 10, 17, 54, 61).

Boscoe, F. P. et al. (2012), « A Nationwide Comparison of Driving Distance Versus Straight-Line Distance to Hospitals », in : *The Professional Geographer* 64.*2*, p. 188–196 (cf. p. 85).

Breiman, L. (2001), « Random Forests », in : *Machine Learning* 45.*1*, p. 5–32 (cf. p. 34).

Breslin, J. G. et al. (2006), « SIOC an Approach to Connect Web-Based Communities », in : *Int. J. Web Based Communities* 2.*2*, p. 133–142 (cf. p. 43).

Bruckschen, F. et al. (2015), « Cookbook for a socio-demographic basket », in : *D4D. Challenge Senegal Sessions Scientific Papers*, MIT Media Lab, Cambridge, USA (cf. p. 105).

Calì, A. et al. (2002), « On the Expressive Power of Data Integration Systems », in : *Proceedings of the 21st International Conference on Conceptual Modeling*, ER '02, Berlin, Heidelberg : Springer-Verlag, p. 338–350 (cf. p. 6).

Calvanese, D. et al. (2018), « Ontology-Based Data Access and Integration », in : *Encyclopedia of Database Systems*, sous la dir. de L. Liu et M. T. Özsu, New York, NY : Springer New York, p. 2590–2596 (cf. p. 15, 69).

Cameron, D. et al. (2013), « PREDOSE : A semantic web platform for drug abuse epidemiology using social media », in : *Journal of Biomedical Informatics* 46.*6*, p. 985–997 (cf. p. 41).

Ceusters, W. (2012), « An information artifact ontology perspective on data collections and associated representational artifacts », in : *Studies in Health Technology and Informatics* 180, p. 68–72 (cf. p. 101).

Cheatham, M. et al. (2015), « Results of the Ontology Alignment Evaluation Initiative 2015 », in : *Proceedings of the 10th International Workshop on Ontology Matching collocated with the 14th International Semantic Web Conference (ISWC 2015), Bethlehem, PA, USA, October 12, 2015.* P. 60–115 (cf. p. 64).

Chen, B. et al. (2006), « Structure-Based Filtering for Ontology Alignment », in : *Enabling Technologies : Infrastructure for Collaborative Enterprises, 2006. WETICE'06. 15th IEEE International Workshops on*, p. 364–369 (cf. p. 55, 62, 63).

Chen, H. et al. (2007), « Towards Semantic e-Science for Traditional Chinese Medicine », in : *BMC Bioinformatics* 8.*S3*, S6 (cf. p. 91).

Chomicki, J. (2003), « Preference formulas in relational queries », in : *ACM Transactions on Database Systems* 28.*4*, p. 427–466 (cf. p. 9).

Claro, D. B. et al. (2006), « Web Services Composition », in : *Semantic Web Services, Processes and Applications*, sous la dir. de J. Cardoso et A. P. Sheth, t. 3, Boston, MA : Springer US, p. 195–225 (cf. p. 15).

Corby, O. et al. (2004), « Querying the Semantic Web with Corese Search Engine », in : *Proceedings of the 16th European Conference on Artificial Intelligence*, ECAI'04, event-place : Valencia, Spain, Amsterdam, The Netherlands, The Netherlands : IOS Press, p. 705–709 (cf. p. 23, 47, 48).

Cossin, S., V. Jouhet et al. (2018), « IAM at CLEF eHealth 2018 : Concept Annotation and Coding in French Death Certificates », in : *CoRR* abs/1807.03674 (cf. p. 43).

Cossin, S., L. Lebrun et al. (2019), « Romedi : an open data source about French drugs on the semantic web », in : *The 17th World Congress of Medical and Health Informatics*, Lyon, France (cf. p. 22, 43, 44, 101).

Côté, R. G. et al. (2006), « The Ontology Lookup Service, a lightweight cross-platform tool for controlled vocabulary queries », in : *BMC bioinformatics* 7, p. 97 (cf. p. 71).

Crespo Azcárate, M. et al. (2013), « Improving image retrieval effectiveness via query expansion using MeSH hierarchical structure », in : *Journal of the American Medical Informatics Association* 20.*6*, p. 1014–1020 (cf. p. 30).

Cuenca Grau, B. et al. (2013), « Results of the Ontology Alignment Evaluation Initiative 2013 », Anglais, in : *Proc. 8th ISWC workshop on ontology matching (OM)*, Sydney, Australie, p. 61–100 (cf. p. 61).

Cunningham, H. et al. (2001), « GATE : an architecture for development of robust HLT applications », in : *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*, Philadelphia, Pennsylvania : Association for Computational Linguistics, p. 168 (cf. p. 48).

Cuzzocrea, A. et al. (2018), « Intelligent Sensor Data Fusion for Supporting Advanced Smart Health Processes », in : *Complex, Intelligent, and Software Intensive Systems*, sous la dir. de L. Barolli et O. Terzo, t. 611, Cham : Springer International Publishing, p. 361–370 (cf. p. 79).

d'Aquin, M. et E. Motta (2011), « Watson, More Than a Semantic Web Search Engine », in : *Semant. web* 2.*1*, p. 55–63 (cf. p. 17).

Deerwester, S. et al. (1990), « Indexing By Latent Semantic Analysis », in : *Journal of the American Society for Information Science* 41, p. 391–407 (cf. p. 31).

Dehainsala, H. et al. (2007), « OntoDB : An Ontology-Based Database for Data Intensive Applications », in : *Advances in Databases : Concepts, Systems and Applications*, sous la dir. de R. Kotagiri et al., t. 4443, Berlin, Heidelberg : Springer Berlin Heidelberg, p. 497–508 (cf. p. 9).

Dhammi, I. et S. Kumar (2014), « Medical subject headings (MeSH) terms », in : *Indian Journal of Orthopaedics* 48.*5*, p. 443 (cf. p. 71).

Diallo, G. (2006), « Une Architecture à base d'Ontologies pour la Gestion Unifiées des Données Structurées et non Structuréees », PhD Thesis, Université Joseph-Fourier-Grenoble I (cf. p. 6).

— (2011), « Efficient Building of Local Repository of Distributed Ontologies », in : IEEE, p. 159–166 (cf. p. 17, 55–57).

— (2014), « An effective method of large scale ontology matching », in : *Journal of Biomedical Semantics* 5.*1*, p. 44 (cf. p. 16–18, 57, 58, 60, 61, 100).

Diallo, G., N. Grabar et al. (2012), « Towards complex queries on data from complex patients », in : AMIA'2012 (cf. p. 15).

Diallo, G., K. Khelif et al. (2008), « Semantic Browsing of a Domain Specific Resources : The Corese-NeLI Framework », in : *Web Intelligence/IAT Workshops*, p. 50–54 (cf. p. 22, 45).

Diallo, G., P. Kostkova et al. (2008), « Process of building a vocabulary for the infection domain », in : *Computer-Based Medical Systems, 2008. CBMS'08. 21st IEEE International Symposium on*, IEEE, p. 308–313 (cf. p. 8, 11, 47, 48).

Diallo, G., M. Simonet et al. (2006a), « An approach to automatic ontology-based annotation of biomedical texts », in : *Advances in Applied Artificial Intelligence*, Springer Berlin Heidelberg, p. 1024–1033 (cf. p. 6).

— (2006b), « Bringing together structured and unstructured sources : the OUMSUIS approach », in : *On the Move to Meaningful Internet Systems 2006 : OTM 2006 Workshops*, Springer Berlin Heidelberg, p. 699–709 (cf. p. 6).

Díaz-Galiano, M. et al. (2009), « Query expansion with a medical ontology to improve a multimodal information retrieval system », in : *Computers in Biology and Medicine* 39.*4*, p. 396–403 (cf. p. 30).

Djeddi, W. E. et al. (2017), « XMap results for OAEI 2017 », in : *Proceedings of the 12th International Workshop on Ontology Matching co-located with the 16th International Semantic Web Conference (ISWC 2017), Vienna, Austria, October 21, 2017.* P. 196–200 (cf. p. 54).

Dragisic, Z. et al. (2016), « User Validation in Ontology Alignment », in : *The Semantic Web – ISWC 2016*, sous la dir. de P. Groth et al., t. 9981, Cham : Springer International Publishing, p. 200–217 (cf. p. 17).

Dramé, K., G. Diallo et al. (2014), « Reuse of termino-ontological resources and text corpora for building a multilingual domain ontology : An application to Alzheimer's disease », in : *Journal of biomedical informatics* 48, p. 171–182 (cf. p. 10, 13, 20, 29, 43).

Dramé, K., F. Mougin et al. (2016), « Large scale biomedical texts classification : a kNN and an ESA-based approaches », in : *Journal of Biomedical Semantics* 7.*1* (cf. p. 20, 38).

Drame, K. et al. (2012), « Towards a bilingual Alzheimer's disease terminology acquisition using a parallel corpus. », in : *MIE*, p. 179–183 (cf. p. 13).

Dummer, T. J. (2008), « Health geography : supporting public health policy and planning », in : *Canadian Medical Association Journal* 178.*9*, p. 1177–1180 (cf. p. 83).

Falconer, S. M. et N. F. Noy (2011), « Interactive Techniques to Support Ontology Matching », in : *Schema Matching and Mapping*, p. 29–51 (cf. p. 16, 62).

Faria, D., C. Martins et al. (2015), « AML results for OAEI 2015 », in : *Proceedings of the 10th International Workshop on Ontology Matching (OM-2015) Colocated with the 14th International Semantic Web Conference (ISWC-2015)*, t. 1545, CEUR-WS, Bethlehem (PA US), p. 116–123 (cf. p. 54).

Faria, D., C. Pesquita et al. (2013), « AgreementMakerLight results for OAEI 2013 », in : *CEUR Workshop Proc. OM2013*, t. 1111, p. 101–108 (cf. p. 54, 62).

Finger, F. et al. (2016), « Mobile phone data highlights the role of mass gatherings in the spreading of cholera outbreaks », in : *Proceedings of the National Academy of Sciences* 113.*23*, p. 6421–6426 (cf. p. 4).

Finin, T. et al. (2004), « Swoogle : A search and metadata engine for the semantic web », in : *In Proceedings of the Thirteenth ACM Conference on Information and Knowledge Management*, ACM Press, p. 652–659 (cf. p. 17).

Gabrilovich, E. et S. Markovitch (2009), « Wikipedia-based Semantic Interpretation for Natural Language Processing », in : *J. Artif. Int. Res.* 34.*1*, p. 443–498 (cf. p. 20, 38).

Gedzelman, S. et al. (2005), « Building an ontology of cardio-vascular diseases for concept-based information retrieval », in : *Computers in Cardiology, 2005*, IEEE, p. 255–258 (cf. p. 10).

Ghoula, N. et al. (2010), « TOK : A Meta-model and Ontology for Heterogeneous Terminological, Linguistic and Ontological Knowledge Resources », in : *2010 IEEE/-WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology* (cf. p. 72).

Gini, R., P. B. Ryan et al. (2013), « Comparison Among EU-ADR, OMOP, Mini-Sentinel And MATRICE Strategies For Data Extraction And Management », in : *PHARMACOEPIDEMIOLOGY AND DRUG SAFETY*, t. 22, WILEY-BLACKWELL 111 RIVER ST, HOBOKEN 07030-5774, NJ USA, p. 189–190 (cf. p. 14).

Gini, R., M. Schuemie et al. (2016), « Data Extraction And Management In Networks Of Observational Health Care Databases For Scientific Research : A Comparison Among EU-ADR, OMOP, Mini-Sentinel And MATRICE Strategies », in : *eGEMs (Generating Evidence & Methods to improve patient outcomes)* 4.*1*, p. 2 (cf. p. 14).

Goble, C. et al. (2001), « Conceptual Open Hypermedia = the Semantic Web ? », in : *Proceedings of the Second International Conference on Semantic Web - Volume 40*, SemWeb'01, Hongkong, China : CEUR-WS.org, p. 44–50 (cf. p. 8, 47).

Golbeck, J. et al. (2003), « The National Cancer Institute's Thesaurus and Ontology », in : *Web Semantics : Science, Services and Agents on the World Wide Web* 1.*1* (cf. p. 17, 60).

Griffon, N. et al. (2014), « Design and usability study of an iconic user interface to ease information retrieval of medical guidelines », in : *Journal of the American Medical Informatics Association* 21.*e2*, e270–e277 (cf. p. 102).

Grosjean, J. et al. (2014), « Comparing BioPortal and HeTOP : towards a unique biomedical ontology portal ? », in : *IWBBIO'14 : 2nd International Work-Conference on Bioinformatics and Biomedical Engineering* (cf. p. 100).

Gruber, T. R. (1995), « Toward principles for the design of ontologies used for knowledge sharing ? », in : *International Journal of Human-Computer Studies* 43.*5-6*, p. 907–928 (cf. p. 9, 10).

Guarino, N. (1998), *Formal Ontology in Information Systems : Proceedings of the 1st International Conference June 6-8, 1998, Trento, Italy*, 1st, Amsterdam, The Netherlands, The Netherlands : IOS Press (cf. p. 10).

Gurwitz, J. H. et al. (2003), « Incidence and preventability of adverse drug events among older persons in the ambulatory setting », in : *JAMA* 289.*9*, p. 1107–1116 (cf. p. 21).

Hahn, U. et S. Schulz (2004), « Building a Very Large Ontology from Medical Thesauri », in : *Handbook on Ontologies*, sous la dir. de S. Staab et R. Studer, Berlin, Heidelberg : Springer Berlin Heidelberg, p. 133–150 (cf. p. 10).

Hall, M. et al. (2009), « The WEKA Data Mining Software : An Update », in : *SIGKDD Explor. Newsl.* 11.*1*, p. 10–18 (cf. p. 35).

Hamdi, F. et al. (2010), « TaxoMap alignment and refinement modules : results for OAEI 2010 », in : *5th International Workshop on Ontology Matching* (cf. p. 54, 62).

Hao, Y. et Y. Zhang (2007), « Web Services Discovery Based on Schema Matching », in : *Proceedings of the Thirtieth Australasian Conference on Computer Science - Volume 62*, ACSC '07, Darlinghurst, Australia, Australia : Australian Computer Society, Inc., p. 107–113 (cf. p. 15).

Hazell, L. et S. A. W. Shakir (2006), « Under-Reporting of Adverse Drug Reactions : A Systematic Review », in : *Drug Safety* 29.*5*, p. 385–396 (cf. p. 39).

Hecker, N. et al. (2012), « SuperTarget goes quantitative : update on drug-target interactions », in : *Nucleic Acids Research* 40.*D1*, p. D1113–D1117 (cf. p. 103).

Heeks, R. (2017), *Information and Communication Technology for Development (ICT4D)*, 1re éd., Routledge (cf. p. 23).

Hepp, M. et J. de Bruijn (2007), « GenTax : A Generic Methodology for Deriving OWL and RDF-S Ontologies from Hierarchical Classifications, Thesauri, and Inconsistent Taxonomies », in : *The Semantic Web : Research and Applications*, sous la dir. de D. Hutchison et al., t. 4519, Berlin, Heidelberg : Springer Berlin Heidelberg, p. 129–144 (cf. p. 73).

Hoehndorf, R. et al. (2015), « Aber-OWL : a framework for ontology-based data access in biology », in : *BMC Bioinformatics* 16, p. 26 (cf. p. 9, 71).

Hofmann, T. (1999), « Probabilistic latent semantic indexing », in : *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '99*, Berkeley, California, United States : ACM Press, p. 50–57 (cf. p. 31).

Holma, H. et A. Toskala (2010), *WCDMA for UMTS : HSPA Evolution and LTE, 5th Edition*, Wiley (cf. p. 82).

Holub, P. et al. (2018), « Enhancing *Reuse* of Data and Biological Material in Medical Research : From FAIR to FAIR-Health », in : *Biopreservation and Biobanking* 16.*2*, p. 97–105 (cf. p. 4).

Horgan, D. et J. A. Lal (2019), « Making the Most of Innovation in Personalised Medicine : An EU Strategy for a Faster Bench to Bedside and Beyond Process », in : *Public Health Genomics*, p. 1–20 (cf. p. 79).

Hripcsak, G. et al. (2015), « Observational Health Data Sciences and Informatics (OHDSI) : Opportunities for Observational Researchers », in : *Studies in Health Technology and Informatics* 216, p. 574–578 (cf. p. 70).

Hu, W. et al. (2006), « Partition-Based Block Matching of Large Class Hierarchies », in : *Proceedings of the First Asian Conference on The Semantic Web*, ASWC'06, Berlin, Heidelberg : Springer-Verlag, p. 72–83 (cf. p. 54).

Huang, M. et al. (2011), « Recommending MeSH terms for annotating biomedical articles », in : *Journal of the American Medical Informatics Association* 18.*5*, p. 660–667 (cf. p. 30, 38).

Jaccard, P. (1912), « The Distribution of the Flora in the Alpine Zone », in : *New Phytologist* 11.*2*, p. 37–50 (cf. p. 60).

Jean, S. et al. (2015), « OntoQL : An Alternative to Semantic Web Query Languages », in : *International Journal of Semantic Computing* 09.*01*, p. 105–137 (cf. p. 9).

Jiang, J. et al. (2013), « FoCUS : Learning to Crawl Web Forums », in : *IEEE Transactions on Knowledge and Data Engineering* 25.*6*, p. 1293–1306 (cf. p. 41).

Jiménez-Ruiz, E., B. Cuenca Grau et al. (2013), « LogMap and LogMapLt results for OAEI 2013 », in : *CEUR Workshop Proc. OM2013*, p. 131–138 (cf. p. 74).

Jiménez-Ruiz, E., B. C. Grau et al. (2012), « Large-scale Interactive Ontology Matching : Algorithms and Implementation », in : *ECAI 2012 - 20th European Conference on Artificial Intelligence. Including Prestigious Applications of Artificial Intelligence (PAIS-2012) System Demonstrations Track, Montpellier, France, August 27-31 , 2012*, p. 444–449 (cf. p. 61, 75).

Jimenez-Ruiz, E. et al. (2012), « Exploiting the UMLS Metathesaurus in the Ontology Alignment Evaluation Initiative », in : *2nd International Workshop on Exploiting Large Knowledge Repositories (E-LKR)*, CEUR-WS.org (cf. p. 62).

Jiménez-Ruiz, E. et al. (2012), « Large-scale Interactive Ontology Matching : Algorithms and Implementation », in : *ECAI 2012 - 20th European Conference on Artificial Intelligence. Including Prestigious Applications of Artificial Intelligence (PAIS-2012) System Demonstrations Track, Montpellier, France, August 27-31 , 2012*, p. 444–449 (cf. p. 63).

Jiminez-Ruiz, E. et al. (2011), « Logic-based assessment of the compatibility of UMLS ontology sources », in : *Journal of biomedical semantics* 2 Suppl 1, S2 (cf. p. 54).

Jones, K. H. et al. (2018), « Challenges and Potential Opportunities of Mobile Phone Call Detail Records in Health Research : Review », in : *JMIR mHealth and uHealth* 6.7, e161 (cf. p. 4, 23, 79, 87).

Jouhet, V., G. Defossez et al. (2013), « Automated Selection of Relevant Information for Notification of Incident Cancer Cases within a Multisource Cancer Registry », in : *Methods of Information in Medicine* 52.05, p. 411–421 (cf. p. 71).

Jouhet, V., F. Mougin et al. (2017), « Building a model for disease classification integration in oncology, an approach based on the national cancer institute thesaurus », in : *Journal of Biomedical Semantics* 8.1, p. 6 (cf. p. 71).

Kachroudi, M., G. Diallo et S. Ben Yahia (2016), « Initiating Cross-Lingual Ontology Alignment with Information Retrieval Techniques », in : *6émes Journées Francophones sur les Ontologies*, Actes des 6émes Journées Francophones sur les Ontologies, Bordeaux, France (cf. p. 18, 66).

— (2018), « KEPLER at OAEI 2018 », in : *Proceedings of the 13th International Workshop on Ontology Matching co-located with the 17th International Semantic Web Conference, OM@ISWC 2018, Monterey, CA, USA, October 8, 2018.* P. 173–178 (cf. p. 19, 66, 68).

Kachroudi, M., G. Diallo et S. B. Yahia (2017), « OAEI 2017 results of KEPLER », in : *Proceedings of the 12th International Workshop on Ontology Matching co-located with the 16th International Semantic Web Conference (ISWC 2017), Vienna, Austria, October 21, 2017.* P. 138–145 (cf. p. 19, 66, 68).

Kachroudi, M., S. Zghal et al. (2013), « OntoPart : At the cross-roads of ontology partitioning and scalable ontology alignment systems », in : *International Journal of Metadata, Semantics and Ontologies* 8, p. 215–225 (cf. p. 18, 67).

Kalfoglou, Y. et M. Schorlemmer (2003), « Ontology mapping : the state of the art », in : *The Knowledge Engineering Review* 18.1, p. 1–31 (cf. p. 15, 53).

Kalyanam, J. et al. (2017), « Exploring trends of nonmedical use of prescription drugs and polydrug abuse in the Twittersphere using unsupervised machine learning », in : *Addictive Behaviors* 65, p. 289–295 (cf. p. 41).

Kammoun, A. et G. Diallo (2013), « ServOMap Results for OAEI 2013 », in : *CEUR Workshop Proc. OM2013*, t. 1111, p. 169–176 (cf. p. 18).

Kamsu-Foguem, B. et al. (2013), « Conceptual graph-based knowledge representation for supporting reasoning in African traditional medicine », in : *Engineering Applications of Artificial Intelligence* 26.4, p. 1348–1365 (cf. p. 13, 88, 91).

Katsahian, S. et al. (2015), « Evaluation of Internet Social Networks using Net scoring Tool : A Case Study in Adverse Drug Reaction Mining », in : *Stud Health Technol Inform* 210, p. 526–530 (cf. p. 41).

Khelif, K. et al. (2007), « An Ontology-based Approach to Support Text Mining and Information Retrieval in the Biological Domain », in : *JUCS - Journal of Universal Computer Science 12* (cf. p. 48).

Kibbe, W. A. et al. (2015), « Disease Ontology 2015 update : an expanded and updated database of human diseases for linking biomedical knowledge through disease data », in : *Nucleic Acids Research* 43.*Database issue*, p. D1071–1078 (cf. p. 54).

Kiessling, W. (2002), « Foundations of Preferences in Database Systems », in : *Proceedings of the 28th International Conference on Very Large Data Bases*, VLDB '02, event-place : Hong Kong, China, VLDB Endowment, p. 311–322 (cf. p. 9).

Kim, S. et al. (2016), « PubChem Substance and Compound databases », in : *Nucleic Acids Research* 44.*D1*, p. D1202–D1213 (cf. p. 92).

Kocev, D. et al. (2007), « Ensembles of Multi-Objective Decision Trees », in : *Machine Learning : ECML 2007*, sous la dir. de J. N. Kok et al., t. 4701, Berlin, Heidelberg : Springer Berlin Heidelberg, p. 624–631 (cf. p. 20, 31).

Kong, H.-J. (2019), « Managing Unstructured Big Data in Healthcare System », in : *Healthcare Informatics Research* 25.*1*, p. 1 (cf. p. 28).

Kostkova, P. et al. (2008), « User profiling for semantic browsing in medical digital libraries », in : *The Semantic Web : Research and Applications*, Springer Berlin Heidelberg, p. 827–831 (cf. p. 8, 15).

Krumholz, H. M. (2014), « Big Data And New Knowledge In Medicine : The Thinking, Training, And Tools Needed For A Learning Health System », in : *Health Affairs* 33.*7*, p. 1163–1170 (cf. p. 19).

Kuhn, M. et al. (2007), « STITCH : interaction networks of chemicals and proteins », in : *Nucleic Acids Research* 36.*Database*, p. D684–D688 (cf. p. 92).

Kujala, R. et al. (2016), « Estimation and monitoring of city-to-city travel times using call detail records », in : *EPJ Data Science* 5.*1* (cf. p. 104).

Lai, S. et al. (2015), « Recurrent Convolutional Neural Networks for Text Classification », in : *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, event-place : Austin, Texas, AAAI Press, p. 2267–2273 (cf. p. 31).

Lambrix, P. et R. Kaliyaperumal (2013), « A Session-Based Approach for Aligning Large Ontologies », in : *The Semantic Web : Semantics and Big Data*, sous la dir. de P. Cimiano et al., t. 7882, Lecture Notes in Computer Science, Springer Berlin Heidelberg, p. 46–60 (cf. p. 62, 76).

Lambrix, P. et Q. Liu (2009), « Using Partial Reference Alignments to Align Ontologies », in : *The Semantic Web : Research and Applications*, sous la dir. de L. Aroyo et al., t. 5554, Berlin, Heidelberg : Springer Berlin Heidelberg, p. 188–202 (cf. p. 54).

Lamy, J.-B. et al. (2008), « An iconic language for the graphical representation of medical concepts », in : *BMC Med Inform Decis Mak* 8, p. 16 (cf. p. 102).

Langley, P. et al. (1992), « An Analysis of Bayesian Classifiers », in : *AAAI* (cf. p. 20, 34).

Lassila, O. et al. (1998), *Resource Description Framework (RDF) Model and Syntax Specification* (cf. p. 8, 11, 54).

Lee, D. et al. (2014), « Literature review of SNOMED CT use », in : *Journal of the American Medical Informatics Association* 21.*e1*, e11–e19 (cf. p. 14, 17, 60).

Levenshtein (1966), « Binary codes capable of correcting deletions, insertions, and reversals », in : (cf. p. 59).

Li, B. et al. (2018), « YaTCM : Yet another Traditional Chinese Medicine Database for Drug Discovery », in : *Computational and Structural Biotechnology Journal* 16, p. 600–610 (cf. p. 13, 14).

Lin, J. et W. J. Wilbur (2007), « PubMed related articles : a probabilistic topic-based model for content similarity », in : *BMC Bioinformatics* 8, p. 423 (cf. p. 30, 32).

Liu, J. et al. (2017), « Deep Learning for Extreme Multi-label Text Classification », in : *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR '17*, Shinjuku, Tokyo, Japan : ACM Press, p. 115–124 (cf. p. 31).

Liu, T.-Y. (2007), « Learning to Rank for Information Retrieval », in : *Foundations and Trends® in Information Retrieval* 3.*3*, p. 225–331 (cf. p. 30).

Lô, G. et al. (2017), « Linking African Traditional Medicine Knowledge », in : *Semantic Web Applications and Tools for Healthcare and Life Sciences (SWAT4LS)*, Rome, Italy (cf. p. 13, 91).

Lowe, H. J. et G. O. Barnett (1994), « Understanding and using the medical subject headings (MeSH) vocabulary to perform literature searches », in : *JAMA* 271.*14*, p. 1103–1108 (cf. p. 12, 20, 29, 54).

Madjarov, G. et al. (2012), « An extensive experimental comparison of methods for multi-label learning », in : *Pattern Recognition* 45.*9*, p. 3084–3104 (cf. p. 31).

Mambone, Z. et al. (2015), « DServO : A Peer-to-Peer-based Approach to Biomedical Ontology Repositories », in : *Studies in Health Technology and Informatics* 216, p. 1103 (cf. p. 17).

Mao, Y. et Z. Lu (2013), « NCBI at the 2013 BioASQ challenge task : Learning to rank for automatic MeSH indexing », in : *BioASQ challenge 2013*, Springer-Verlag Berlin Heidelberg (cf. p. 35, 36, 38).

Màrquez, L. et H. Rodríguez (1998), « Part-of-speech tagging using decision trees », in : *Machine Learning : ECML-98*, sous la dir. de J. G. Carbonell et al., t. 1398, Berlin, Heidelberg : Springer Berlin Heidelberg, p. 25–36 (cf. p. 48).

Mascardi, V. et al. (2007), « A Comparison of Upper Ontologies », in : *WOA 2007 : Dagli Oggetti agli Agenti. 8th AI*IA/TABOO Joint Workshop "From Objects to Agents" : Agents and Industry : Technological Applications of Software Agents, 24-25 September 2007, Genova, Italy*, p. 55–64 (cf. p. 91, 94).

Mazzocchi, F. (2018), « Knowledge Organization System (KOS) : An Introductory Critical Account », in : *KNOWLEDGE ORGANIZATION* 45.*1*, p. 54–78 (cf. p. 5, 22).

McGuinness, D. L., F. Van Harmelen et al. (2004), « OWL web ontology language overview », in : *W3C recommendation* 10.*2004-03*, p. 10 (cf. p. 6, 10, 11).

McVitie, D. G. et L. B. Wilson (1971), « The stable marriage problem », in : *Commun. ACM* 14.*7*, 486â€“490 (cf. p. 18, 60).

Meilicke, C. et al. (2012), « MultiFarm : A benchmark for multilingual ontology matching », in : *Web Semantics : Science, Services and Agents on the World Wide Web* 15, p. 62–68 (cf. p. 17, 19, 68).

Miles, A. et al. (2005), « SKOS core : simple knowledge organisation for the web », in : *International Conference on Dublin Core and Metadata Applications*, pp–3 (cf. p. 10, 11, 72).

Miller, G. A. (1995), « WordNet : A Lexical Database for English », in : *COMMUNICATIONS OF THE ACM* 38, p. 39–41 (cf. p. 59, 63, 67).

Monge, A. E. et C. P. Elkan (1996), « The Field Matching Problem : Algorithms and Applications », in : *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)*, p. 267–270 (cf. p. 60).

Montjoye, Y.-A. d. et al. (2014), « D4D-Senegal : The Second Mobile Phone Data for Development Challenge », in : *CoRR* abs/1407.4885 (cf. p. 80).

Mork, J. G. et al. (2013), *The nlm medical text indexer system for indexing biomedical literature*, Proceedings of the first Workshop on Bio-Medical Semantic Indexing

and Question Answering, a Post-Conference Workshop of Conference and Labs of the Evaluation Forum 2013 (CLEF 2013), Valencia (cf. p. 30, 38).

Mrabet, Y. et al. (2007), « Recognising Professional-Activity Groups and Web Usage Mining for Web Browsing Personalisation », in : *IEEE/WIC/ACM International Conference on Web Intelligence (WI'07)*, Fremont, CA, USA : IEEE, p. 719–722 (cf. p. 23, 47).

Mungall, C. J. et al. (2012), « Uberon, an integrative multi-species anatomy ontology », in : *Genome Biology* 13.*1*, R5 (cf. p. 54).

Mutafungwa, E. et al. (2015), « Mobile Data as Public Health Decision Enabler : A Case Study of Cardiac and Neurological Emergencies », in : *Data For Developement Challenge, NetMob'15* (cf. p. 24).

Navigli, R. (2009), « Word sense disambiguation : A survey », in : *ACM Computing Surveys* 41.*2*, p. 1–69 (cf. p. 67).

Névéol, A. et al. (2018), « Clinical Natural Language Processing in languages other than English : opportunities and challenges », in : *Journal of Biomedical Semantics* 9.*1*, p. 12 (cf. p. 28, 102).

Ngo, D. H. et Z. Bellahsene (2012), « YAM++ : (not) Yet Another Matcher for Ontology Matching Task », Anglais, in : *Bases de DonnÃ©es AvancÃ©es*, sous la dir. de N. Anciaux, France, p. 5 (cf. p. 61).

Nickel, M. et al. (2016), « Holographic Embeddings of Knowledge Graphs », in : *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, Phoenix, Arizona : AAAI Press, p. 1955–1961 (cf. p. 103).

Noy, N. F. et al. (2009), « BioPortal : ontologies and integrated data resources at the click of a mouse », in : *Nucleic Acids Research* 37.*Web Server*, W170–W173 (cf. p. 13, 71, 72, 100).

Oliveira, J. L. et al. (2013), « The EU-ADR Web Platform : delivering advanced pharmacovigilance tools », in : *Pharmacoepidemiology and Drug Safety* 22.*5*, p. 459–467 (cf. p. 15).

Oliver, H. et al. (2009), « A user-centred evaluation framework for the Sealife semantic web browsers », in : *BMC bioinformatics* 10.*Suppl 10*, S14 (cf. p. 8, 12, 23, 45, 49, 52).

OOAS (2013), *Inventaire de formulations des plantes médicinales à éfficacité prouvé*, Organisation Ouest Africaine de la Santé (cf. p. 92).

Otero-Cerdeira, L. et al. (2015), « Ontology matching : A literature review », in : *Expert Systems with Applications* 42.*2*, p. 949–971 (cf. p. 15, 53, 72).

Paulheim, H. (2016), « Knowledge graph refinement : A survey of approaches and evaluation methods », in : *Semantic Web* 8.*3*, sous la dir. de P. Cimiano, p. 489–508 (cf. p. 6, 17, 90, 103).

Paulheim, H. et al. (2013), « Towards Evaluating Interactive Ontology Matching Tools », English, in : *The Semantic Web : Semantics and Big Data*, sous la dir. de P. Cimiano et al., t. 7882, Lecture Notes in Computer Science, Springer Berlin Heidelberg, p. 31–45 (cf. p. 16, 62).

Penev, L. et al. (2016), « A common registration-to-publication automated pipeline for nomenclatural acts for higher plants (International Plant Names Index, IPNI), fungi (Index Fungorum, MycoBank) and animals (ZooBank) », in : *ZooKeys* 550, p. 233–246 (cf. p. 92).

Pesquita, C. et al. (2013), « To repair or not to repair : reconciling correctness and coherence in ontology reference alignments », in : *OM*, p. 13–24 (cf. p. 61).

Pinto, H. S. et J. P. Martins (2001), « A methodology for ontology integration », in : *Proceedings of the international conference on Knowledge capture - K-CAP 2001*, Victoria, British Columbia, Canada : ACM Press, p. 131 (cf. p. 15, 100).

Ponte, J. M. et W. B. Croft (1998), « A Language Modeling Approach to Information Retrieval », in : *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, event-place : Melbourne, Australia, New York, NY, USA : ACM, p. 275–281 (cf. p. 30).

Porter, M. F. (1997), « Readings in Information Retrieval », in : sous la dir. de K. Sparck Jones et P. Willett, San Francisco, CA, USA : Morgan Kaufmann Publishers Inc., p. 313–316 (cf. p. 32).

Quinlan, J. R. (1993), *C4.5 : Programs for Machine Learning*, San Francisco, CA, USA : Morgan Kaufmann Publishers Inc. (cf. p. 34).

Read, J. et al. (2011), « Classifier chains for multi-label classification », in : *Machine Learning* 85.*3*, p. 333–359 (cf. p. 31).

Ribadas, F. et al. (2013), « Two hierarchical text categorization approaches for BioASQ semantic indexing challenge », in : *CEUR Workshop Proceedings* 1094 (cf. p. 38).

Ritze, D. et H. Paulheim (2011), « Towards an automatic parameterization of ontology matching tools based on example mappings », in : t. 814, CEUR Workshop Proceedings (CEUR-WS.org) (cf. p. 62).

Rosse, C. et J. L. V. Mejino (2008), « The Foundational Model of Anatomy Ontology », in : *Anatomy Ontologies for Bioinformatics*, sous la dir. d'A. Burger et al., t. 6, London : Springer London, p. 59–117 (cf. p. 17).

Rosse, C. et J. L. V. Mejino Jr (2003), « A reference ontology for biomedical informatics : the Foundational Model of Anatomy », in : *Journal of biomedical informatics* 36.*6*, p. 478–500 (cf. p. 60).

Roussel, V. et al. (2013), « Estimating the excess of inappropriate prescriptions of anti-dopaminergic anti-emetics during acute gastroenteritis epidemics in France : Inappropriate Anti-Dopaminergic Anti-Emetics Use For Gastroenteritis », in : *Pharmacoepidemiology and Drug Safety*, n/a–n/a (cf. p. 21).

Ruch, P. (2006), « Automatic assignment of biomedical categories : toward a generic approach », in : *Bioinformatics* 22.*6*, p. 658–664 (cf. p. 30).

Safran, C. et al. (2007), « Toward a National Framework for the Secondary Use of Health Data : An American Medical Informatics Association White Paper », in : *Journal of the American Medical Informatics Association* 14.*1*, p. 1–9 (cf. p. 4).

Salton, G. et al. (1975), « A Vector Space Model for Automatic Indexing », in : *Commun. ACM* 18.*11*, p. 613–620 (cf. p. 30, 32, 43).

Savadogo, M. et al. (2016), « Alignement interactif d'ontologies avec ServOMap », in : *Technique et Science Informatiques* 35.*1*, p. 55–74 (cf. p. 18, 72).

Schroeder, M. et al. (2006), « Sealife : a semantic grid browser for the life sciences applied to the study of infectious diseases », in : *Stud Health Technol Inform* 120, p. 167–178 (cf. p. 8, 11, 46).

Schulz, S. et M. Boeker (2013), « BioTopLite : An Upper Level Ontology for the Life Sciences. Evolution, Design and Application », in : p. 1889–1899 (cf. p. 14, 94).

Schulz, S., R. Cornet et al. (2011), « Consolidating SNOMED CT's Ontological Commitment », in : *Appl. Ontol.* 6.*1*, p. 1–11 (cf. p. 60).

Sertkaya, A. et al. (2014), *Examination of clinical trial costs and barriers for drug development*, rapp. tech. 1, p. 1–92 (cf. p. 102).

Severo, B. et al. (2014), « VOAR : A Visual and Integrated Ontology Alignment Environment », anglais, in : *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, sous la dir. de N. C. (Chair) et al., Reykjavik, Iceland : European Language Resources Association (ELRA) (cf. p. 62).

Shi, F. et al. (2009), « Actively Learning Ontology Matching via User Interaction », in : *The Semantic Web - ISWC 2009*, sous la dir. d'A. Bernstein et al., t. 5823, Lecture Notes in Computer Science, Springer Berlin Heidelberg, p. 585–600 (cf. p. 62).

Shickel, B. et al. (2018), « Deep EHR : A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis », in : *IEEE journal of biomedical and health informatics* 22.*5*, p. 1589–1604 (cf. p. 101, 102).

Shvaiko, P. et J. Euzenat (2013), « Ontology Matching : State of the Art and Future Challenges », in : *IEEE Trans. Knowl. Data Eng.* 25.*1*, p. 158–176 (cf. p. 14–16, 18, 53, 74).

Shvaiko, P., J. Euzenat et al., éds. (2013), *Proceedings of the 8th International Workshop on Ontology Matching co-located with the 12th International Semantic Web Conference (ISWC 2013), Sydney, Australia, October 21, 2013*, t. 1111, CEUR Workshop Proceedings, CEUR-WS.org (cf. p. 72).

Some, B. M. J. et al. (2019), « Design Considerations for a Knowledge Graph : The WATRIMed Use Case », in : *Studies in Health Technology and Informatics*, p. 59–64 (cf. p. 11, 13).

Sowa, J. F. (2008), « Chapter 5 Conceptual Graphs », in : *Foundations of Artificial Intelligence*, t. 3, Elsevier, p. 213–237 (cf. p. 13, 88).

Spyromitros, E. et al. (2008), « An Empirical Study of Lazy Multilabel Classification Algorithms », in : *Artificial Intelligence : Theories, Models and Applications*, sous la dir. de J. Darzentas et al., t. 5138, Berlin, Heidelberg : Springer Berlin Heidelberg, p. 401–406 (cf. p. 31, 33, 35).

Steele, J. E. et al. (2017), « Mapping poverty using mobile phone and satellite data », in : *Journal of The Royal Society Interface* 14.*127*, p. 20160690 (cf. p. 105).

Stoilos, G. et al. (2005), « A String Metric for Ontology Alignment », in : *Proceedings of the International Semantic Web Conference (ISWC 05)*, sous la dir. d'Y. Gil, t. 3729, LNCS, Springer-Verlag, p. 624–637 (cf. p. 59).

Strohl, W. R. (2000), « The role of natural products in a modern drug discovery program », in : *Drug Discovery Today* 5.*2*, p. 39–41 (cf. p. 90, 102).

Suchanek, F. M. et al. (2007), « Yago : A Core of Semantic Knowledge », in : *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, New York, NY, USA : ACM, p. 697–706 (cf. p. 93).

Suominen, O. et al. (2009), « HealthFinland–a National Semantic Publishing Network and Portal for Health Information », in : *Journal of Web Semantics* 7.*4*, p. 287–297 (cf. p. 89).

Sweeney, L. (2002), « Achieving k-Anonymity Privacy Protection Using Generalisation And Suppression », in : *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10.*05*, p. 571–588 (cf. p. 81).

Tapucu, D., G. Diallo, Y. Ait-Aimeur et al. (2009), « Ontology-Based Database Approach for Handling Preferences », in : *Data Warehousing Design and Advanced Engineering Applications : Methods for Complex Construction*, Ladjel Bellatreche, IGI Global (cf. p. 9).

Tapucu, D., G. Diallo, S. Jean et al. (2009), « Définition et Exploitation des Préférences au Niveau Sémantique », in : *ACM proceedings of 2ème Journées Francophones sur les Ontologies*, p. 29–36 (cf. p. 9).

Tchacondo, T. et al. (2011), « Herbal remedies and their adverse effects in Tem tribe traditional medicine in Togo », in : *Afr J Tradit Complement Altern Med* 8.*1*, p. 45–60 (cf. p. 90).

Thiao-Layel, B. et al. (2016), « K-Ware : vers une gestion conjointe de ressources sémantiques et leurs alignements », in : *6ièmes Journées Francophone sur les Ontologies*, Actes des 6ièmes Journées Francophone sur les Ontologies, Bordeaux, France : Gayo Diallo, Okba Kazar, Fleur Mougin (cf. p. 19, 100, 101).

— (2018), « Joint handling of semantic knowledge resources and their alignments », in : *Proceedings of the 13th International Workshop on Ontology Matching co-located with the 17th International Semantic Web Conference, OM@ISWC 2018, Monterey, CA, USA, October 8, 2018*. P. 226–227 (cf. p. 19, 100).

To, H.-V. et al. (2009), « An adaptive machine learning framework with user interaction for ontology matching », in : Proc. Workshop on Information Integration on the Web, p. 35–40 (cf. p. 62).

Tognazzo, S. et al. (2009), « Probabilistic classifiers and automated cancer registration : An exploratory application », in : *Journal of Biomedical Informatics* 42.*1*, p. 1–10 (cf. p. 71).

Trieschnigg, D. et al. (2009), « MeSH Up : effective MeSH text classification for improved document retrieval », in : *Bioinformatics* 25.*11*, p. 1412–1418 (cf. p. 30, 34, 38).

Trojahn Dos Santos, C. et al. (2010), « An API for multilingual ontology matching », in : *Proc. 7th conference on Language Resources and Evaluation Conference (LREC)*, Valletta, Malta : No commercial editor., p. 3830–3835 (cf. p. 67).

Tsoumakas, G. et al. (2009), « Mining Multi-label Data », in : *Data Mining and Knowledge Discovery Handbook*, sous la dir. d'O. Maimon et L. Rokach, Boston, MA : Springer US, p. 667–685 (cf. p. 30, 36).

Ukkonen, E. (1992), « Approximate String-matching with Q-grams and Maximal Matches », in : *Theor. Comput. Sci.* 92.*1*, p. 191–211 (cf. p. 59).

Uschold, M. et M. Gruninger (2004), « Ontologies and semantics for seamless connectivity », in : *ACM SIGMOD Record* 33.*4*, p. 58 (cf. p. 10).

Vijayan, M. et P. H. Reddy (2016), « Stroke, Vascular Dementia, and Alzheimer's Disease : Molecular Links », in : *Journal of Alzheimer's Disease* 54.*2*, p. 427–443 (cf. p. 80).

WAHO (2013), *West African Herbal Pharmacopoeia*, First edition (cf. p. 88).

Wang, P. et al. (2011), « Matching Large Ontologies Based on Reduction Anchors », in : *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Volume Three*, IJCAI'11, Barcelona, Catalonia, Spain : AAAI Press, p. 2343–2348 (cf. p. 54).

Wang, Y. et al. (2008), « Exploring Traversal Strategy for Web Forum Crawling », in : *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '08, event-place : Singapore, Singapore, New York, NY, USA : ACM, p. 459–466 (cf. p. 40).

Weber, G. M. et al. (2014), « Finding the Missing Link for Big Biomedical Data », in : *JAMA* (cf. p. 28).

Wesolowski, A. et al. (2012), « Quantifying the impact of human mobility on malaria », in : *Science* 338.*6104*, p. 267–270 (cf. p. 79).

WHO (2013), *WHO traditional medicine strategy : 2014-2023* (cf. p. 88).

Wilkinson, M. D. et al. (2016), « The FAIR Guiding Principles for scientific data management and stewardship », in : *Scientific Data* 3, p. 160018 (cf. p. 3, 96).

Wu, H. et al. (2014), « Deep semantic embedding », in : *CEUR Workshop Proceedings* 1204, p. 46–52 (cf. p. 31).

Wu, Z. et al. (2008), « Semantic Web Development for Traditional Chinese Medicine », in : *Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence, AAAI 2008, Chicago, Illinois, USA, July 13-17, 2008*, p. 1757–1762 (cf. p. 14, 91).

Yan, X. et al. (2013), « A biterm topic model for short texts », in : *Proceedings of the 22nd international conference on World Wide Web - WWW '13*, Rio de Janeiro, Brazil : ACM Press, p. 1445–1456 (cf. p. 41).

Yan, Y. et al. (2018), « Biomedical literature classification with a CNNs-based hybrid learning network », in : *PLOS ONE* 13.*7*, sous la dir. de C. C. Reyes-Aldasoro, e0197933 (cf. p. 31).

Yang, J.-M. et al. (2009), « Incorporating site-level knowledge to extract structured data from web forums », in : *Proceedings of the 18th international conference on World wide web - WWW '09*, Madrid, Spain : ACM Press, p. 181 (cf. p. 40).

Yu, T. et al. (2017), « Knowledge graph for TCM health preservation : Design, construction, and applications », in : *Artificial Intelligence in Medicine* 77, p. 48–52 (cf. p. 91).

Zayani, C. A. et al. (2017), « Semantic-based Reconstruction of User's Interests in Distributed Systems », in : *Computación y Sistemas* 21.*3* (cf. p. 9).

Zhang, M.-L. et Z.-H. Zhou (2007), « ML-KNN : A lazy learning approach to multi-label learning », in : *Pattern Recognition* 40.*7*, p. 2038–2048 (cf. p. 31).

Zhu, D. et al. (2013), « An incremental approach to MEDLINE MeSH indexing », in : *CEUR Workshop Proceedings* 1094 (cf. p. 38).

Deuxième partie

Curriculum Vitae

Chapitre 5

CV Gayo DIALLO - April 2019

Highlights

Affiliation

Associate Prof. in Computer Sciences — Univ. of Bordeaux
HDR, Hors Classe — Bordeaux Population Health Research Center
Section 27 — INSERM 1219 & LaBRI UMR5800
Mail : Gayo.Diallo@u-bordeaux.fr — Web : http ://www.gayodiallo.org

Main Research Topics

My research activity is dedicated to investigate into methods and tools for facilitating the access to health-care related data and knowledge at a semantic level. To do so, I rely on theoretical approaches from the Artificial Intelligence domain, specifically Knowledge Engineering (Ontology Based Health Care related Knowledge Representation and Modelling, Large Scale Ontology Matching), Semantic Information Retrieval and Biomedical TextMining and Information and Communication Technology for Development.

Th main application domain of this research activity is the Health Care domain, in particular the Public Health one.

Keywords : Heterogeneous Data and Knowledge Semantic Integration, Ontologies & Medical Knowledge Engineering, Semantic Information Retrieval, ICT for Development (ICT4D)

Collective Duties and Professional Membership

- Since 11/2017 Director of Team ERIAS, emerging research group of Bordeaux BPH Research Center INSERM 1219 : team dedicated to provide AI based methods and tools for facilitating secondary use of healthcare related data (15 people including MSc, 5 permanent staff)
- Member of the Research Council of the University of Bordeaux 2014- present
- Member of the academic council of the University of Bordeaux
- Elected on the board of the BPH Center INSERM 1219 06/2016- present
- Member of the Council of the Institute of Public Health (ISPED) (2018-present)
- Member of the French Association AFIA : National Ass. for Artificial Intelligence
- Member of the French Health Informatics Association

Awards and Distinction

-

— Awarded of the French Ministry of Higher Education and Research's Scientific Excellency Bonus (2013-current)
— Winner of the Data For Development Big Data International Challenge Practical Application Prize (2015) : a challenge involving 150 scientific teams and 60 projects in various topics related to Calls Data Records cross-linking and analysis. I coordinated an international consortium of 4 labs (Team ERIAS Bordeaux INSERM 1219, VMASC Lab of Old Dominion University (USA), ComNet Aalto University (Finland), LIPPA Univ Tunis El Manar (Tunisia)
— Awarded of the SAMPO Program (French-Finland) mobility fund in 2015 to visit Labs in Finland and setup collaborations
— The ServOMap large scale ontology matching system I have designed ranked among the top 3 systems for largebio track during the Ontology Alignment Evaluation Initiative (OAEI) challenge, 2012, 2013 and 2015 edition. The Kepler system which implements ServOMap features ranked 1st for MultiLingual matching task of OAEI 2017 and 2018.

Cursus

Professional Background

— Habilitation to Supervise Research (HDR), Univ. Bordeaux **June 2019**
— Authorised to Direct PhD Thesis (ADT), Univ. Bordeaux **Since Oct. 2016**
— Assoc. Prof., Univ. Bordeaux **Sep. 2009 - current**
— Postdoc Researcher, LISI Laboratory/ENSMA Poitiers **Oct. 2008 - Aug. 2009**
— Research Assistant, CeRC, City University London **Mars 2007 - Sep. 2008**
— Lecturer in Computer Sciences, Univ. Grenoble 2 **Oct. 2004 - Aug. 2006**
— Junior Lecturer (Moniteur), Univ. Grenoble 2 **Oct. 2001 - Sep. 2004**
— Research Assistant PhD Student, Univ. Grenoble 1 **Oct. 2001 - Sep. 2004**
— Teaching Assistant, IUT 2 Grenoble **Feb. 2001 - May. 2001**

Educational Background

— Habilitation to Supervise Research (HDR), Univ. Bordeaux **June 2019**
— PhD in Computer Sciences, Univ. Grenoble 1 **Dec. 2006**
Topic **:** An Ontology based architecture for the management of structured and unstructured data **Jury :**
  — M. Yves Chiaramella, Professeur Université Grenoble 1, Président
  — M. Djamel Zighed, Professeur Université de Lyon 2, Rapporteur
  — M. Christophe Roche, Professeur Université de Savoie, Rapporteur
  — Mme. Sylvie Calabretto, MCF HDR INSA de Lyon, Examinatrice
  — M. Michel Simonet, Chargé de Recherche au CNRS, Directeur de Thèse
  — Mme. Ana Simonet, MCF Univ. Grenoble 2
— Degree in Business Activities Creation **May 2003**
Univ. Grenoble 1, 2 and 3
— European Master in Management and Technology of Information Systems **2001**

Univ. Geneve (Switzerland), Federal Polytechnic School (EPFL) Lausanne, Univ. Grenoble 1, & National Polytechnic Institute (INPG) Grenoble

— MSc. in Information Systems **June 2001**
Univ. Grenoble 1, Joseph Fourier, France

— Engineering Degree in Computer Science **Sep. 1999**
National Computer Sciences School - INI, Algiers, Algeria

Teaching activities

— Active participation into the launch of the ISPED/Bordeaux University MSc. in Information Systems and IT Technology for Healthcare (SITIS)
— Coordinator of the Web Based Data course of the new Master in Digital Health Data Science (Ecole Universitaire de Recherche - Digital Public Health Graduate School)
— Internship Coordinator of the Master Speciality SITIS
— Coordinator of 2 courses (Biomedical TextMining and Information Retrieval, Databases management and Object Oriented Programming) MSc. SITIS ISPED and Databases Management of the MSc. Mathematics Stochastic and Statistics, Univ. of Bordeaux. Co-lead of the Web Semantic course of the Cognitive National School of Bordeaux INP.
— 200 hours in average of teaching every year, including online courses as well as teaching abroad

Research Activities

Supervision of Junior Researchers : PostDoc, PhD and Master Students

PostDoc

— **Goergeta Bordea** ; (04/2019-04/2021), Marie-Curie Horizon 2020 PostDoc, KaNNA : Herbs-Drugs interactions detection using artificial neural networks

PhD Students

— **Bruno Thiao Layel**,(2018-2021). CIFRE-ANRT PhD student with the Start-Up Synapse-Medicine : Services for joint handling of data and knowledge for drug related information management. Co-supervision with Guillaume Blin (LaBRI) and Vianney Jouhet (ERIAS & Bordeaux Hospital)
— **Sébastien Cossin**, (2018-2021). Information Extraction and Visualization from Clinical DataWarehouse : case of the Bordeaux Universitary Hospital. Co-supervision with Vianney Jouhet (ERIAS & Bordeaux Hospital)
— **Aimé Patrice Koumamba**, (2019-2022). Implementation of an information system for statistics and health monitoring in Gabon, Joint PhD Thesis between University of Bordeaux and Doctoral School of Franceville (Gabon). Co-supervision with Edgard Ngoungou.

— **Khadim Dramé** : (2011-2014, defended) PhD in Health Informatics. Univ. of Bordeaux. Co-supervised with Fleur Mougin. **Period** : 11/2011-12/2014, **topic** : Contribution to the construction of ontologies and information retrieval : application to the medical field, **rate** : 50%, **Funding** : Fondation Plan Alzheimer - Projet SemBip). After his PhD thesis defence, Dramé was a postdoc INRIA Bretagne then Assistant Prof. at University of Ziguinchor (Senegal) since Sep. 2015.

I am also participating in the supervision of the CIFRE-ANRT PhD thesis of Myriam Lopez. **Period** : 2018-2021. Localisation and prediction of failures for the factory of the future (Lectra company) under the direction of Marie Beurton-Aimar and Sofiane Maabout(LaBRI).

Master and Bachelor students

— **Thomas Cheung-Toi-Cheung**, 2019 (rate 50%), MSc Information System and IT Technologies for Health, University of Bordeaux. **Topic** : Medicinal plants Knowledge graph validation and enrichment. Co-supervised with Georgeta Bordea.

— **Raphael Bazenet**, 2019 (rate 50%), Master 1 Public Health, ISPED - University of Bordeaux. **Topic** : Visualising the content of a medicinal plants Knowledge Graph. Co-supervised with Rabia Azzi.

— **Adrien Sentenac**, 2018 (rate 50%), MSc Information System and IT Technologies for Health, University of Bordeaux. **Topic** : Automated anticoagulant indicators computing from the Bordeaux Hospital Information Systems : case of the Pacha project. Co-supervised with Vianney Jouhet.

— **Louis Biliet**, 2018 (taux 33%), MSc Information System and IT Technologies for Health, University of Bordeaux. **Topic** : Metrics for Drugs Misuse detection from social networks. Supervised with Frantz Thiessard and Sébastien Cossin.

— **Pierre Lemaire**, 2017 (2sd year Ecole Nationale Supérieur de Cognitique, Co-supervision 50%). **Topic** : an ontological model for Drug-Food interactions detection.

— **Corentin Bienassis**, 2017 (2sd year Ecole Nationale Supérieur de Cognitique (ENSC), Co-supervision 50%). **Topic** : Drugs misuse visualising from social networks data.

— **David Myers** (50%, 2016) : MSc Information System and IT Technologies for Health, University of Bordeaux. **Topic** : Comparison and a solution for immunology data integration.

— **Bruno Thiao Layel** (100%, 2015) : MSc Information System and IT Technologies for Health, University of Bordeaux. **Topic** : Persistence and visualising of heterogeneous data and knowledge : pharmacovigilance use cases.

— **Nouha Kheder** (100%, 2015) : IT Engineering school - Univ Tunis El Manar. **Topic** : Re-engineering of an alignment system and binary-based indexing optimisation.

— **Sébastien Cossin** (33%, 2015) : MSc Information System and IT Technologies for Health, University of Bordeaux. **Thème** : Multi-label classification of scientific articles using the MeSH thesaurus.

— **Zakaria Mambone** (50%, 2014) : IT design engineering school (CICI)- Univ Bobo Dioulasso. **Topic** : Decentralised and cooperative Ontology Server.

— **Mahamadi Savadogo** (50%, 2014) : IT design engineering school (CICI)- Univ

Bobo Dioulasso. **Topic** : User involvement in large scale Ontology matching.
— **Amal Kammoun** (100%, 2013) : IT Engineering school - Univ Tunis El Manar. **Topic** : Design and realisation of a large scale Ontology matching system.
— **Maher Harrouchi** (33%, 2013) : IT Engineering school - Univ Tunis El Manar. **Topic** : Semantic Portal for Alzheimer resources related annotation and retrieval. Co-supervision with Khadim Dramé and Fleur Mougin.
— **Mouhamadou Ba** (100%, 2012) : MSc in IT Univ. Gaston Berger St. Louis, Senegal. **Topic** : Dynamic Knowledge Server and Semantic Similarity measures between entities.
— **Khadim Dramé** (50%, 2011) : MSc in IT Univ. Gaston Berger St. Louis, Senegal. **Topic** : Design and Development of an Ontology for Alzheimer Disease. Co-supervion with Fleur Mougin
— **Nacima Belaidi** (80%, 2009) : MSc IT and Telecommunication - BDTW, Univ Poitiers. **Topic** : Towards a profile model for Ontology Based Databases.
— **Atif Ali** (50%, 2006) : DEA Information System, Univ. Joseph Fourier Grenoble. **Topic** : Information Retrieval guided by an Ontology. Co-supervision with Michel Simonet.

Research Engineer Supervision

— **Bruno Thiao-Layel** (100%, 09/2016- 04/2018) : K-Ware platform for Drug-Safe project (funding Agence Nationale sur la Sécurité du Médicament).
— **Sebastien Cossin** (50%, 09/2016-01/2017) : Indicators extraction from texts forums in the context of the Drug-Safe project
— **Thimothée Jourde** (50%, 04/2017-07/2017) : Web forums scrapping for drugs misuse and information visualisation.

Professional Activities

PhD Defense Committee

— Haolin Ren, University of Bordeaux, France, March 2019
— Rabia Azzi, University of Paris 13, France, Nov. 2018
— Ali Goné Ngom, University of Gaston Berger Senegal, Aug. 2018
— Khadim Dramé, University of Bordeaux, France, December 2014

Invitations and Visiting Scholar

Some Invited Talks

— Mobile Data in Senegal, a Health Decision Enabler. Second International Symposium "Perspectives on ICT4D", 22 May 2015 Vrije University of Amsterdam
— Kepler Result for OAEI 2017, Ontology Matching Workshop, October 21st, 2017, Vienna, Austria
— Mobile Data as Public Health Decision Enabler : A Case Study of Cardiac and Neurological Emergencies. D4D Challenge at NetMob'2015, 10 Avril 2015, MIT (Cambridge, USA)
— December 2013 : ISIR (Université d'Osaka, Japon) invited by Kouji Kozaki on the topic Large Scale Ontology Matching

Visiting professor

— Febuary 2019 (5 days) : Tallin Technical University (TalTech), Estonia
— July 2016 (10d) : Laboratory VMASC, Old Dominion University (USA)
— December 2015 (7 days) : University Jendouba (Tunisia)
— September 2015 (10d) : scientific stay in Finland (winner of the SAMPO fellowship program). I visited and gave talks in 5 Laboratories, including Turku, Helsinki and Oulu.
— July 2015 (7d) : University of Ouagadougou, Burkina Faso, invited by Borlli Michel Somé.
— Avril 2015 (7d) : University of Jendouba (Tunisia), invited by Sami Zghal. 20h of classes "Représentation des Connaissances" Master 1 Informatique de Gestion
— May 2013 (7 d) : Laboratoire LIPAH, Faculté des Sciences Université de Tunis El Manar (Tunisie) invited by Sadok ben Yahia (co-editor of the speciaIssue Special of journal IJRIS)
— April 2013 (10j) : University of Aveiro (Portugal), invited by José Luis Oliveira (collaboration in the context of the EU-ADR european project)

Scientific Duties

Research and Development Projects

European Projects

— KaNNA (Marie Curie HORIZON 2020, 2019-2021), Herbs-Drugs interactions detection using artificial neural networks, Scientific director of the postdoc Marie-Curie H2020 of Goergeta Bordea, with Vincent Lepetit (Labri Bordeaux)
— EHVA (European HIV Vaccine Alliance, H2020) : 2016-2020. Le centre INSERM 1219 (Pr Rodolphe Thibeaut) involved in the Work Package *Data Integration and Down Selection of Vaccine Candidates* under the scientific coordination of Rodolphe Thiebaut (SISTM team of BPH INSERM 1219)
— EU-ADR (FP7) : Early detection of adverse drug events by integrative mining of clinical records and biomedical knowledge. Period : 2008-2012, 18 partenaires. Status : coordination of the ERIAS team participation for the last 2 out of 3 years duration of the project. Coordination : University Medical Center (EMC) Rotterdam.
— ARITMO (FP7) : Arrhythmogenic potential of drugs. Period : 2010-2012, 17 partners. Status : Participant ; Coordination : University Medical Center (EMC), Rotterdam.
— SeaLIfe (FP6) : A semantic grid browser for the life sciences applied to the study of infectious diseases. Period 2006-2009, 6 partners. Status : Postdoc on the project and coordination of the design and development of the SeaLife knowledge resource.
— INFACE (FP5) : Advanced Visual InterFACEs for timely Retrieval of Patient Related Information. Period 2002-2004. Status : participant. Coordination : GUY'S AND ST. THOMAS' HOSPITAL NATIONAL HEALTH SERVICE TRUST (UK)

National Projects

- e-ATM : Enhancing Patient Safety In African Traditional Medicine. Period : 2016-2017. **Funding** : ANR MRSEI. Budget : 22Keuros. **Status** : scientific coordinator
- DoMINO/Drug-SAFE : Evaluation systématisée du médicament en population. Period : 2015-2018. **Status** : Scientific coordinator of a task **Funding** : French National Agency for Medicines and Health Products (ANSM), Budget for ERIAS 400keuros.
- ANR TECSAN RAVEL : Retrieval And Visualisation of Electronic health records. **Period** : 2012-2014, 6 partners (2 from Industry). Status : Our team leads 2 WPs. **Coordination** : University of Rennes.

Regional Projects

- PACHA : Development and validation of automated indicators of the relevance of oral anticoagulant prescription in adult medicine based on the hospital information system. **Periode** : 2016-2018. **Status** : Participant ; Scientific coordinator : Frantz Thiessard
- DIMMER : Detection of Drug Misuse from Social Networks and Forums. **Periode** : 2016-2017. **Funding** : PEPS IDEX Bordeaux, Budget 16KEuros, **Status** : Scientific coordinator
- K-Ware : Framework générique pour l'intégration données et connaissances : application à la cancérologie. Période : 2015. **Status** : scientific coordinator. Funding : Fédération de Recherche Santé Publique Population Société.
- BC3 : Base de Connnaissances Coeur Cerveau, Rhone Alpes regional Project (2001-2004). Status : research assistant, PhD funded by the project. Coordinator : Univ Lyon 1.

Expertise

- Projects Evaluator for the Medical Research Council (MRC), United Kingdom
- Expert projects evaluator for the French National Agency for Research (ANR)
- 2015-2017 : Member of the Advisory Board Simulating Religion Project, Institute for The Bio-Cultural Study of Religion. Virginia Modeling, Analysis and Simulation Center, Old Dominion University, USA.
- Since 2015 : Expert for the African Centre of Excellence in Mathematics, Computer Science and ICT (CEA-MITIC)
- 2014 : Expert for the National Center of Science and Technology Evaluation, Republic of Kazakhstan :
- 2009 : Expert for the French AERES, section des établissements. member of the committee for the evaluation of the ENSAM Art et Métiers ParisTech.

Conferences and Workshops Organisation

- 2016 : Co-President of the 6th French Speaking Conference on Ontologies (JFO), 13 - 14 October 2016, held in Bordeaux. Conference sponsorized by the National French Professional Asociation in Intelligence Artifical (AFIA)
- 2014, 2016 et 2017 : Co-Organisation of the International Workshop BigCVEn'2014 (Marrakech, Morocco), BigCVEn'2015 (Bangkok, Thailand), BigCVEn'2016 (Naples,

Italie) and BigCVEn'2017 (Jaipur, India) in conjunction with the IEEE conference SITIS (Signal Image Technology and Internet Based Systems)

— 2014 : Co-Chair of the Track Web Computing and Applications of the IEEE conference Signal Image Technology and Internet Based Systems (SITIS)

— 2010 : Poster and Demo chair of the e-Health'2010 international conference, Casablanca, Morocco.

Peer Review Activities & PC Member

— Journals : they include World Wide Web (WWW), Journ. of Biomed. Sem. (JBS), Arti. Intel. in Med. (AIIM), IMIA Year Book, Exp. Sys. Appli. (ESA)
— Conferences Program Committee : it includes MedInfo, AMIA, Digital Public Health, ICCCI, e-Health, Ontology Matching, IA&Santé, JFO, CNRIA

Publications

Synthesis

Guest Editor Activity : **04** | Book Chapters : **03** | Journal Articles : **15** | International Conferences and Workshops : **33** | National Conferences and Workshops : **06** | Posters and Demos : **09** | Patent : **01**

Publications list

Guest Editor

— **G. Diallo**, F. Mougin. Special Issue on 6th edition of French Speaking Conference on Ontologies (JFO). Journal on Data Semantics, Volume 7, Issue 4, pp 189–189. Guest Editors. 2018.
— A. Bon, V. de Boer, C. Guéret, **G. Diallo**, G. Gordijn. Proceedings of the 5th International Symposium "Perspectives on ICT4D" co-located with 10th ACM Web Science Conference (WebSci'18). Amsterdam, the Netherlands, May 27, 2018. CEUR Workshop Proceedings. Vol-2120
— N. Kamel, **G. Diallo**, S. Ben Yahia. Special Issue on : CIIA'2015 Advances in Computational Intelligence for Big Data. International Journal of Reasoning-based Intelligent Systems (IJRIS). Interdiscience Publishers. Guest Editors. 2017.
— **G. Diallo**, O. Kazar, F. Mougin. Actes des 6èmes Journées Francophones sur les Ontologies (JFO'2016), 13-14 Oct. 2016, Bordeaux, France

Book Chapters

— D. Tapucu, **G. Diallo**, Y. Ait Ameur, M. O. Unalir. Ontology-based Database Approach for Handling Preferences. In Data Warehousing Design and Advanced Engineering Applications : Methods for Complex Construction. Ladjel Bellatreche (eds), IGI Global, 2009. ISBN 978-1-60566-756.
— M. Simonet, R. Messai, **G. Diallo**, A. Simonet. Ontologies in the Health Field. In Data Mining and Medical Knowledge Management : Cases and Applications, P. Berka, J. Rauch, D. Abdelkader Zighed (eds) IGI Global, 2008. ISBN 978-1-60566-218-3.
— M-H. Pham, D. Bernhard, **G. Diallo**, R. Messai, M. Simonet. SOM-based Clustering of Multilingual Documents Using an Ontology. In Data Mining with Ontologies : Implementations, Findings and Frameworks. H.O. Nigro, S.C. Cisaro, D. Xodo (Eds.), IGI Global. ISBN : 978-1-59904-618-1, 2007.

Journal Articles

- F. Mougin, D. Auber, R. Bourqui, **G. Diallo**, I. Dutour, V. Jouhet, F. Thiessard, R. Thiébaut, P. Thébault. Visualizing omics and clinical data : which challenges for dealing with their variety ? *Methods*. Volume 132, 1 January 2018, Pages 3-18. https ://doi.org/10.1016/j.ymeth.2017.08.012.
- K. Dramé, F. Mougin, **G. Diallo**. Large scale biomedical texts classification : a kNN and an ESA-based approaches. J. Biomedical Semantics. Vol :7(40), 2016. DOI Information : 10.1186/s13326-016-0073-1. [IF 1.620]
- R. Gini, M. Schuemie, J. Brown, P. Ryan, E. Vacchi, M. Coppola, W. Cazzola, P. Coloma, R. Berni, **G. Diallo**, J. L. Oliveira, P. Avillach, G. Trifirò, P. Rijnbeek, M. Bellentani, J. van Der Lei, N. Klazinga, M. Sturkenboom. Data Extraction And Management In Networks Of Observational Health Care Databases For Scientific Research : A Comparison Among EU-ADR, OMOP, Mini-Sentinel And MATRICE Strategies. eGEMs (Generating Evidence & Methods to improve patient outcomes). Vol. 4 (2016) :1
- M. Sawadogo, M.S. Borlli, **G. Diallo**. Alignement Interactif d'Ontologies avec ServOMap. Technique et Science Informatiques. VOL 35/1 - 2016 - pp.55-74 - doi :10.3166/tsi.35.55-74.
- **G. Diallo**. An effective method of large scale ontology matching. J. of Biomedical Semantics, vol :5(44), 2014. DOI :10.1186/2041-1480-5-44. [IF 1.620]
- K. Dramé, **G. Diallo**, F. Delva, JF. Dartigues, E. Mouillet, R. Salamon, F. Mougin. Reuse of termino-ontological resources and text corpora for building a multilingual domain ontology : An application to Alzheimer's disease. J. of Biomed. Informat., vol. 48 :171-182. 2013. doi :10.1016/j.jbi.2013.12.013. [IF 2.447]
- P. Lopes, T. Nunes, D. Campos, L. Furlong, A. Bauer-Mehren, F. Sanz, MC. Carrascosa, J. Mestres, J. Kors, B. Singh, E. Mulligen, J. van der Lei, **G. Diallo**, P. Avillach, E. Ahlberg, S. Boyer, C. Diaz, J.L. Oliveira. Gathering and Exploring Scientific Knowledge in Pharmacovigilance. PLOS One, vol. 8(12)2013. doi :10.1371/journal. pone.0083016
- B. Kamsu-Foguem, **G. Diallo**, C. Foguem. Conceptual graph-based knowledge representation for supporting reasoning in African traditional medicine. Engineering Applications of Artificial Intelligence, 26(4) :1348-1365. 2013.
- M. Ba, **G. Diallo**. Large-scale biomedical ontology matching with ServOMap. IRBM, 34(1) : Digital Technologies for Healthcare Issue. Pages 56-59. 2013
- P. Avillach, J-C. Dufour, **G. Diallo**, F. Salvo, M. Joubert, F. Thiessard, F. Mougin, G. Trifiro, A. Fourrier-Reglat, A. Pariente, M. Fieschi. Design and Validation of an Automated Method to Detect Known Adverse Drug Reactions in MEDLINE : a Contribution from the EU-ADR Project. Journal of American Medical Informatics Association (JAMIA). 2013 ;20 :446-452 doi :10.1136/amiajnl-2012-001083.
- J.L. Oliveira, P. Lopes, T. Nunes, D. Campos, S. Boyer, E. Ahlberg, E. Mulligen, J. Kors, B. Singh, L. Furlong, A. Bauer-Mehren, M. Carrascosa, J. Mestres, P. Avillach, **G. Diallo**, C. Diaz Acedo, J. van der Lei. The EU-ADR Web Platform : Delivering advanced pharmacovigilance tools. Pharmacoepidemiol Drug Saf. 22(5) :459-67. 2013. doi :10.1002/pds.3375.
- A. Bauer-Mehren, EM. van Mullingen, P. Avillach, MC. Carrascosa, R. Garcia-Serna, J. Pinero, B. Singh, P Lopes, JL. Oliveira, **G. Diallo**, J. Mestres, E. Ahlberg Helgee, S. Boyer, F. Sanz, JA. Kors, LI. Furlong. Automatic filtering and substantiation of drug safety signals. PLoS Computational Biology 8(4) : 2012.

e1002457. doi :10.1371/journal.pcbi.1002457.
— **G. Diallo**, M. Simonet. Towards Dynamic Ontologies Repository Building. SIGHIT Record 2(1) : 27 (2012).
— H. Oliver, **G. Diallo**, E. de Quincey, D. Alexopoulou, B. Habermann, P. Kostkova, M. Schroeder, S. Jupp, K. Khelif, R. Stevens, G. Jawaheer, G. Madle. A user-centred evaluation framework for the Sealife semantic web browsers. BMC Bioinformatics. Suppl 10 :S14
— M. Simonet, **G. Diallo**, G., Palmer, A. Simonet. (2003). Ontologies pour l'organisation des connaissances - vers une ontologie anatomo-fonctionnelle du cerveau. Santé et Systémique (Hermès-Lavoisier), 7(1) :44-75.

International Conferences and Workshops

— BMJ. Some, G. Bordea, F. Thiessard, S. Schulz, **G. Diallo** : Design Considerations for a Knowledge Graph : the WATRIMed use case. Studies in Health Technology and Informatics, vol. 259 :59 - 64. IOS Press . DOI :10.3233/978-1-61499-961-4-59.
— S. Cossin, L. Lebrun, G. Lobre, R. Loustau, V. Jouhet, R. Griffier, F. Mougin, **G. Diallo**, F. Thiessard : Romedi : an open data source about French drugs on the semantic web. MEDINFO'2019, 26-30 Aug 2019, Lyon, France.
— S. Bouasker, S.Ben Yahia, **G. Diallo** : An Insight into Biological Data Mining based on Rarity and Correlation as Constraints. In SAC'19 Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing Pages 3-10, Limassol, Cyprus, 08-12 April 2019.
— K. Reby, S. Cossin, G. Bordea, **G. Diallo**. SITIS-ISPED in CLEF eHealth 2018 Task 1 : ICD10 coding using deep learning. 19th Working Notes of CLEF Conference and Labs of the Evaluation Forum, CLEF 2018 (2018) CEUR Workshop Proceedings, vol.2125.
— E. Jimenez-Ruiz, T. Saveta, O. Zamazal, S. Hertling, M. Röder, I. Fundulaki, A.C. Ngonga Ngomo, M.A. Sherif, A. Annane, Z. Bellahsene, S.Ben Yahia, **G. Diallo**, D. Faria, M. Kachroudi, A. Khiat, P. Lambrix, H. Li, M. Mackeprang, M. Mohammadi, M. Rybinski, B.S. Balasubramani, C. Trojahn. Introducing the HOBBIT platform into the Ontology Alignment Evaluation Campaign. Proceedings of the 13th International Workshop on Ontology Matching co-located with the 17th International Semantic Web Conference (ISWC 2018).
— S. Cossin, V. Jouhet, F. Mougin, **G. Diallo**, F. Thiessard. IAM at CLEF eHealth 2018 : Concept Annotation and Coding in French Death Certificates. 19th Working Notes of CLEF Conference and Labs of the Evaluation Forum, CLEF 2018 ; CEUR Workshop Proceedings, vol.2125.
— M. Kachroudi, **G. Diallo**, S. Ben Yahia. OAEI 2018 results of Kepler. Proceedings of the 13th International Workshop on Ontology Matching co-located with the 17th International Semantic Web Conference (ISWC 2018). CEUR Workshop Proceedings, vol.2125 :173-178
— M. Kachroudi, **G. Diallo**, S. Ben Yahia. On the Composition of Large Biomedical Ontologies Alignment. ACM Proceedings of the 7th International Conference on Web Intelligence, Mining and Semantics (WIMS'17). pp :24 :1–24 :10, June 19-22th 2017, Amantea, Italy. Doi :10.1145/3102254.3102284.
— S. El Hajjami, M. Berrada, M. Harti, **G. Diallo**. Towards An Agent-Based Approach For Multidimensional Analyses of Semantic Web Data. IEEE Proceedings of

the second International Conference on Intelligent Systems and Computer Vision. April 2017, Fès, Maroc. DOI : 10.1109/ISACV.2017.8054933
— A. Khiat, **G. Diallo**, Beyza Yaman, E. Jiménez-Ruiz, M. Benaissa. ABOM and ADOM : Arabic Datasets for the Ontology Alignment Evaluation Campaign. Proceedings of On the Move to Meaningful Internet Systems Conferences -ODBASE'2015. Lecture Notes in Computer Science, vol 9415. pp 545-553. 2015.
— E. Mutafungwa, F. Thiessard, MP. Diallo, R. Gore, V. Jouhet, F. Mougin, S. Ben Yahia, C. Karray, N. Kheder, R. Saddem, J. Hämäläinend, **G. Diallo**. Mobile Data as Public Health Decision Enabler : A Case Study of Cardiac and Neurological Emergencies. Proceedings of the Orange D4D Challenge, 4th International conference on the scientific analysis of mobile phone datasets (NetMob'2015), 7-10 April 2015, MIT Media Lab, Cambridge, MA, USA.
— N. Kheder, **G. Diallo**. ServOMBI at OAEI 2015. Proceedings of the 10th International Workshop on Ontology Matching collocated with the 14th International Semantic Web Conference (ISWC 2015), Bethlehem, PA, USA, October 12, 2015. CEUR Workshop Proceedings 1545. pp 200-207.
— K. Dramé, F. Mougin, **G. Diallo**. A k-nearest neighbor based method for improving large scale biomedical document indexing. Proceedings of the 6th International Symposium on Semantic Mining in Biomedicine (SMBM), 6th-7th October, 2014, Aveiro, Portugal.
— K. Dramé, F. Mougin, **G. Diallo**. Query expansion using external resources for improving information retrieval in the biomedical domain. CEUR WS Lab Proceedings for CLEF 2014. Vol-1180 :189-194
— Kammoun, **G. Diallo**. ServOMap results for OAEI 2013. In : The 8th International Workshop on Ontology Matching (OM-2013), collocated with the 12th International Semantic Web Conference (ISWC'2013), CEUR Workshop Proceedings. Vol 1111, 2013.
— **G. Diallo**, M. Ba. Effective method for large scale ontology matching. CEUR-WS Proceedings of the 5th International Workshop on Semantic Web Applications and Tools for Life Sciences (SWAT4LS 2012), Vol-952. 2012.
— M. Ba, **G. Diallo**. ServOMap and ServOMap-lt results for OAEI 2012. The 7th International Workshop on Ontology Matching (OM 2012), collocated with the 11th International Semantic Web Conference (ISWC'2012), CEUR Workshop Proceedings. Vol 946.
— M. Ba, **G. Diallo**. Knowledge Repository as Semantic Similarity Enabler. IEEE Proceedings of the 8th International Conference on Signal Image Technology & Internet Based Systems (SITIS'12). 26-29 Nov 2012, Naples, Italy.
— K. Drame, **G. Diallo**, F. Mougin. Towards a bilingual Alzheimer's disease terminology acquisition using a parallel corpus. Stud Health Technol Inform. 2012 ;180 :179-83
— F. Thiessard, F. Mougin, **G. Diallo**, V. Jouhet, S. Cossin, N. Garcelon, B. Campillo, W. Jouini, J. Grosjean, P. Massari, N. Griffon, M. Dupuch, F. Tayalati, E. Dugas, A. Balvet, N. Grabar, S. Pereira, B. Frandji, S. Darmoni, M. Cuggia. RAVEL : Retrieval And Visualization in ELectronic health records. Stud Health Technol Inform. 2012 ;180 :194-8.
— R. Gini, M. Coppola, P. Ryan, G. Righetti, I. Peri, R. Berni, P. Avillach, P. Coloma, G. Trifirò, **G. Diallo**, J. van der Lei, M. Sturkenboom, M. Schuemie. Frameworks for data extraction and management from electronic healthcare databases

for multi-center epidemiologic studies : a comparison among EU-ADR, MATRICE, and OMOP strategies. 24th European Medical Informatics Conference - MIE2012. 26-29th August 2012, Pisa, Italy.

— **G. Diallo**. Efficient building of local repository of distributed ontologies. IEEE Proc. of the 7th International Conference on Signal Image Technology & Internet Based Systems (SITIS'11). K Yetongnon, R Chbeir and A Dipanda eds. Nov 28-Dec 1st 2011, Dijon, France.

— E. de Quincey, H. Oliver, P. Kostokova, G. Jawaheer, G. Madle, **G. Diallo**, D. Alexopoulou, M. Shroeder, B. Habermann, K. Khelif, S. Jupp, R. Stevens. Mining for Pattern of Semantic Link Usage : Do Domain Users Actually Like Semantic Browing ? International Workshop on Intelligent Web Intereaction (IWI'09), Milano, Italy. 2009.

— **G. Diallo**, K. Khelif, O. Corby, P. Kostkova, Gemma Madle. Semantic Browsing of a Domain Related Resources : The Corese-NeLI Framework, Proc. Int. Workshop on Web Information Retrieval Support Systems (held with the 2008 IEEE/-WIC/ACM Int. WI'08 Conference), 9-12 December 2008, Sidney, Australia. AR : 24%.

— P. Kostkova, **G. Diallo**, G. Jawaheer. User Profiling for Semantic Browsing in Medical Digital Librairies. 5th European Semantic Web Conference (ESWC'08), 01-05th of June 2008, Tenerife, Spain.

— **G. Diallo**, P. Kostkova, G. Jawaheer, S. Jupp, R. Stevens. Process of Building a Vocabulary for the Infection Domain. 21st IEEE International Symposium on Computer-Based Medical Systems, 17-19th June 2008, Jyvaskyla, Finland.

— P. Kostkova, **G. Diallo**, G. Jawaheer. Application of User Profiling on Ontology Module Extraction for Medical Portals. In Proceedings of MedInfo Workshops, MedSemWeb 2007, "What Semantics Do We Need for A Semantic Web for Medicine ?", Brisbane, Australia, August 2007.

— **G. Diallo**, M. Simonet, A. Simonet. An Approach to Automatic Ontology-based Annotation of Biomedical Texts. In Proceedings of IEA/AIE'06, Springer Lecture Notes in Artificial Intelligence, vol. 4031, 2006, pp 1024-1033.

— **G. Diallo**, M. Simonet, A. Simonet. Bringing Together Structured and Unstructured Sources : the OUMSUIS Approach. International Workshop on Knowledge Systems in Bioinformatics (KSinBIT'06), in conjunction with OTM'06, Lecture Notes in Computer Science Vol. 4277 pp.699-709, Springer-Verlag, 2006.

— M. Simonet, R. Patriarche, D. Bernhard, **G. Diallo**, S. Ferriol, P. Palmer. Multilingual Ontology Enrichment for Semantic Annotation and Retrieval of Medical Information. In Proceedings of MEDNET'06, Toronto, Canada, October 17th-20th, 2006.

— R. Patriarche, S. Gedzelman, **G. Diallo**, D. Bernhard, C.G Bassolet, S. Ferriol, A. Girard, M. Mouries, P. Palmer, M. Simonet. A Tool for Textual and Conceptual Annotation of Documents. In Innovation and the Knowledge Economy , Volume 2, Issues, Applications, Case Studies (Proc. of Intl. conference eChallenges 2005) ; Paul Cunningham and Miriam Cunningham (eds) ; IOS Press, pp.865-872. ISSN 1574-1230, ISBN 1-58603-563-0. 2005.

— M. Simonet, D. Bernhard, **G. Diallo**, S. Gedzelman, R. Patriarche. An environment for ontology design and enrichment from texts. In Proceedings of SWAP05 - Semantic Web Applications and Perspectives 2nd Italian Semantic Web Workshop, Trento(Italy), 15-16 December, 2005. CEUR Workshop Proceedings, ISSN

1613-0073.

— S. Gedzelman, M. Simonet, D. Bernhard, **G. Diallo**, P. Palmer. Building an ontology of Cardio-Vascular diseases for Concept-Based Information Retrieval. IEEE Proceedings Computers in Cardiology, Vol.32 : pp.255-258, 2005.

Posters and Demos

— BMJ. Some, G. Bordea, F. Thiessard, **G. Diallo** : Enabling West African Herbal Traditional Medicine Digitalization : the WATRIMed Knowledge Graph. MEDINFO'2019, 26-30 Aug 2019, Lyon, France.

— S. Cossin, L. Lebrun, A. Niamkey, F. Mougin, M. Lambert, **G. Diallo**, F Thiessard, V Jouhet : SmartCRF : a prototype to visualize, search and annotate a patient's electronic health record in an i2b2 clinical data warehouse. MEDINFO'2019, 26-30 Aug 2019, Lyon, France

— B. Thia-Layel, V. Jouhet, **G. Diallo**. Joint Handling of Semantic Knowledge Resources and their Alignments. Proceedings of the 13th International Workshop on Ontology Matching co-located with the 17th International Semantic Web Conference (ISWC 2018). Ceur Workshop Proceedings, Vol-2288 :226-227

— H. Akremi, S. Zghal, **G. Diallo**. Modeling of uncertainty : Fuzzification of medical ontology. WIMS '16 : ACM Proceedings of the 6th International Conference on Web Intelligence, Mining and Semantics. Nîmes, France — June 13 - 15, 2016. pp 32 :1-32 :4. doi>10.1145/2912845.2912879. Poster Paper

— Z. Manbone, M. Savadogo, B. M. J. Some, **G. Diallo** :DServO : A Peer-to-Peer-based Approach to Biomedical Ontology Repositories. MedInfo 2015 : 1103. Poster Paper

— **G. Diallo**, A. Kammoun. Towards Learning Based Strategy for Improving the Recall of the ServOMap Matching System. In : CEUR Workshop Proceedings SWAT4LS'2013, Vol-1114, (4 pages). 2013.

— **G. Diallo**, M. Ba. Matching à large échelle d'ontologies : l'approche ServOMap. 13eme Conférence Francophone sur l'Extraction et la Gestion des Connaissances 29 janvier - 01 février 2013, Toulouse, France.

— **G. Diallo**, N. Grabar, F. Thiessard, N., Garcelon, J. Grosjean, M. Dupuch, S., Bento-Pereira, B. Frandji, SJ. Daroni,M. Cuggia. Towards complex queries on data from complex patients. In : AMIA'2012 Annual Symposium, Nov.2012, Chicago, USA.

— **G. Diallo**. Towards decentralized and cooperative repositories of distributed ontologies. ACM Proceedings of the 4th International SWAT4LS Workshop :Semantic Web Applications and Tools for Life Sciences (SWAT4LS,11). pp 8-9, 2012.

National conferences and workshops

— M. Kachroudi, **G. Diallo**, S. Ben Yahia. Initiating Cross-Lingual Ontology Alignment with Information Retrieval Techniques. Actes des 6èmes Journées Francophones sur les Ontologies (JFO), 13-14 Oct. 2016, Bordeaux, France

— B. Thiao-Layel, V. Jouhet, **G. Diallo**. K-Ware : vers une gestion conjointe de ressources sémantiques et leurs alignements. Actes des 6èmes Journées Francophones sur les Ontologies (JFO), 13-14 Oct. 2016, Bordeaux, France.

— H. Akremi, S. Zghal, V. Jouhet, **G. Diallo**. FONTO : Une nouvelle méthode de fuzzification d'ontologies. Actes des 6èmes Journées Francophones sur les Ontologies (JFO), 13-14 Oct.. 2016, Bordeaux, France.
— M. Savadogo, BMJ. Some, **G. Diallo**. Alignement interactif d'ontologies avec ServOMap.5ème Journées Francophones sur les Ontologies (JFO), 14-16 Nov. 2014, Hammamet, Tunisie
— K. Dramé, **G. Diallo**, F. Mougin. Construction d'une ontologie bilingue de la maladie d'Alzheimer à partir de textes médicaux. Atelier IC pour l'Interopérabilté Sémantique dans les applications en e-Santé, 26 Juin 2012, Paris, France.
— D. Tapucu, **G. Diallo**, S. Jean, Y. Ait Ameur, M. O. Ünalir, N. Belaidi. Définition et Exploitation des Préférences au Niveau Sémantique. 3èms Journées Francophones sur les Ontologies (JFO'09), Poitiers, France. 3-4 Décembre 2009.

Patent

— L. Letenier, C. Goehrs, S. Cossin, B. Thiao-Layel, F. Thiessard, A. Pariente, **G. Diallo**, V. Jouhet. Dispositif et un procédé de génération d'une base de données relative aux médicaments. Num. demande France 1661257. 21 Novembre 2016.

Table des matières