ONTOPRIMA: A PROTOTYPE FOR AUTOMATING ONTOLOGY POPULATION

Jawad Makki[1]

[1] Faculty of Information I, Lebanese University, Beirut, Lebanon

ABSTRACT

Ontology Population supports the process of building ontologies in the complex task of instantiating ontology. Performing this process manually is both expensive and time consuming; this logically leads to attempts of fully or partially automating the process of acquisition and absorption of knowledge in general and the process of Ontology Population in particular.

This paper presents OntoPRiMa (Ontology Population for PRIMA Project), a prototype based on a semi-automatic approach for ontology population in the domain of risk management. OntoPRiMa demonstrates the feasibility of the proposed approach in populating generic and small domain ontologies from textual corpus with many potentials and possible instances that are unknown and not realized. The prototype is based on Natural Language Processing (NLP) techniques for language processing, semantic web techniques (RDFS, RDF, Jena APIs) for knowledge modeling and representation and on domain expert's intervention for validating extracted instances. Furthermore, this paper presents an experiment for populating the ontology of the PRIMA project using a text corpus which is consist of chemical fact sheets issued by the U.S. Environmental Protection Agency. We have achieved high precision (above 80%) of knowledge acquisition in this experiment.

KEYWORDS

Ontology Population, Ontology-Based Information Extraction, Knowledge Acquisition, Semantic Web.

1. INTRODUCTION

In recent years, Ontologies have become a keystone technology for the knowledge representation and the semantic web. but with the fast growth and production of information and web contents many research efforts has been initiated and developed in order to provide optimal solutions in the field of information extraction and knowledge acquisition in general, and the field of ontology construction in particular.

Performing ontology construction tasks manually are expensive, complicated, laborious, tedious, and time consuming as well; moreover ontology evolution requires frequent updates [1], this logically has led to many attempts to fully or partially automate the process of creating and updating ontologies. Therefore there has been a growing interest in the (semi-) automatic learning and populating ontologies from data, in particular texts written in natural language.

Ontology Population is the process of defining and instantiating a knowledge base associated to ontology [2], it consists in learning the extensional aspects of a domain; and to be more specific it intends to learn instances of concepts as well as relations [3]. Whereas Ontology Learning refers to the process of acquiring (constructing or integrating) an ontology (semi-) automatically [4], it consists in adding by learning new concepts and new relations to the ontology.

DOI : 10.5121/ijwest.2017.8401

In this paper we are interested in Ontology Population from text. The work presented here builds upon our semi-automatic approach for Ontology Population via NLP Techniques in Risk Management previously published in [5] and [6]. We introduce OntoPRiMa (Ontology Population for PRIMA Project), a prototype that aims to semi-automate the process of ontology population in the domain of risk management by using the generic domain ontology of PRIMA [7].

PRIMA is a research and technological development project supported by the Information Society Technologies (IST) Programme of the European Union's Fifth Framework Programme. it represents the initial work of defining a Domain Generic Ontology, validated in industrial context, and kernel for further developments in the fields of ontology extension or content extension.

The PRIMA ontology is a domain generic ontology, it is a small-sized ontology but with many potentials and possible instances that are unknown and not realized or populated. The textual corpus (in which the ontology population process will use as source of knowledge) consists of technical texts in natural language written in English.

The prototype is based on NLP techniques for language processing, semantic web techniques (RDFS, RDF, Jena APIs) for knowledge modeling and representation and on domain expert's intervention for validating extracted instances.

We have experimented our prototype for populating the ontology of the PRIMA project using chemical fact sheets issued by the U.S. Environmental Protection Agency.

The remaining of this paper is organized as follows, where we present in section 2 the related works. Then we describe our proposed prototype OntoPRiMa in section 3. After that we apply an experiment In section 4. at last, we draw conclusions for further work and a future evaluation in section 5.

2. RELATED WORK

Ontology Population supports the process of ontology construction by building the Knowledge Base associated to the ontology. It consists of adding new instances of concepts and relations into an existing ontology. This process usually starts after the conceptual model of ontology is built.
Performing this process manually is both expensive and time consuming, which it logically leads to attempts of fully or partially automating the process of Ontology Population. Therefore some methods and approaches for automating the population of ontologies have emerged and considerably increased. These methods are utilizing various techniques starting from syntactic analysis, statistical techniques, named entities recognition, pattern-based extraction, and semantic analysis.

The research paper [8] proposes a system that automatically extracts knowledge about artists from the Web, populates a knowledge base and uses it to generate personalized biographies. It uses syntactic analysis to get the Part-Of-Speech and employs Semantic analysis to perform named entity recognition and extract binary relations between two instances. It also applies a set of heuristics and reasoning methods in order to remove redundant instances from the ontology.
The project presented in [9] is an automatic approach that can extract instances of arbitrary given binary relations from natural language. It uses a deep syntactic analysis and statistical techniques to learn the extraction patterns for the relation.

Paper [10] describes a pattern-based method to automatically enrich a core ontology with the definitions of a domain glossary. Where paper [4] applies a method in the domain of cultural heritage. It is an automatic approach that extracts instances from semi-structured corpora (Art and Architecture Thesaurus) with the help of manually developed extraction patterns.

An information extraction system is presented in paper [11], this system automatically extracts information from heterogeneous sources (semi-structured data such as tables, unstructured text, images and image captions) and populates a knowledge base by using a standard rule-based information extraction system in order to extract named entities. These entities are converted into semantic structures with the help of special mapping declarative rules. Furthermore paper [11] addresses the problem of entity disambiguation by performing simple checks during instances creation.

The presented related works deal with ontology population from different aspects but we have noticed certain number of limitations in their methods and approaches. We will briefly discuss these limitations in the following points:

- They are non-portable or difficult to be portable from one domain to another since they are domain-dependent. This is the case of these works ([8], [10], [11])

- The instantiation of the knowledge base in the approach presented in [9] does not cover all the essential elements of ontology (the classes and properties), they are populating only the classes (concepts) or the properties (relations) but not both.

- Most of these methods are automatic and do not recognize the importance of human judgment as a strong and main element in the process of knowledge extraction and ontology construction. Consequently these methods do not provide decision support for the users or the domain experts.

- In sensitive domains such as risk management, knowledge extraction and ontology population cannot rely totally on automatic methods; Human control and validation are important in the process of retrieving and extracting knowledge from textual resources written in natural language.

3. OntoPRiMa

Based on our approach (NLP-based Ontology Population) presented in our previous articles [5], we are exposing in this paper the implementation of it in a prototype application called **OntoPRiMa** (Ontology Population for PRIMA Project). **OntoPRiMa** is mainly designed for the ontology population process within the PRIMA project [7].

The main objective of **OntoPRiMa** is to extract knowledge (more specifically to extract instances of concepts and instances of relations) from text in order to populate a given ontology.
The prototype is based on NLP techniques for language processing, semantic web techniques (RDFS, RDF, Jena APIs) for knowledge modeling and representation, and on domain expert's intervention for validating extracted instances.

The following sections present the used standards in the prototype, the OntoPRiMa architecture and its processes and functionalities.

3.1. STANDARDS & DEVELOPMENT TOOLS IN OntoPRiMa

OntoPRiMa is implemented in Java programming language. The ontology of PRIMA project (as the input ontology of the prototype) has been modeled and represented in RDF Schema (RDFS) [12] to insure semantic interoperability among semantic web technologies (Figure 1). It has been modeled using the free open source ontology editor Protégé [13].

```
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
       xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
       xmlns:prima="http://www.esi2.us.es/prima/ontology#"
       xml:base="http://www.esi2.us.es/prima/ontology#">
<rdfs:Class rdf:ID="Risk"/>
<rdfs:Class rdf:ID="Cause"/>
<rdfs:Class rdf:ID="Action"/>
<rdfs:Class rdf:ID="Impact"/>
<rdf:Property rdf:ID="Provoke">
      <rdfs:domain rdf:resource="#Cause"/>
      <rdfs:range rdf:resource="#Risk"/>
</rdf:Property>
<rdf:Property rdf:ID="reduce">
      <rdfs:domain rdf:resource="#Action"/>
      <rdfs:range rdf:resource="Cause"/>
</rdf:Property>
<rdf:Property rdf:ID="desc">
      <rdfs:domain rdf:resource="#Cause"/>
      <rdfs:range rdf:resource="http://www.w3.org/2000/01/rdf-
        schema#Literal"/>
</rdf:Property>
.....
.....
</rdf:RDF>
```

Figure 1.Part of PRIMA Ontology in RDFS

The populated ontology with the validated instances of classes and properties (the final output of this prototype) are represented using RDF [14] framework (Figure 2).

```
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
          xmlns:prima="http://www.esi2.us.es/prima/ontology#"
          xml:base="http://www.esi2.us.es/prima/ontology#">
<prima:Cause rdf:ID="cs-46">
    <prima:desc>Breathing small amounts of biphenyl over long periods
of time</prima:desc></prima:Cause>
<prima:Risk rdf:ID="rs-101">
    <prima:desc>damage to the nervous system</prima:desc>
</prima:Risk>
<prima:Cause rdf:about="#cs-46">
    <prima:Provoke rdf:resource="#rs-101"/>
</prima:Cause>
.....
</rdf:RDF>
```

Figure 2.RDF triplets from the populated PRIMA Ontology

As for the part-of-speech (POS) tagging, we used the TreeTagger tool [15] which is a language independent part-of-speech tagger. For semantic relation extraction, we relied on WordNet [16] in

order to expand some specific words with related terms and synonyms. To handle and interact with the ontology within a Java environment, we used the Jena API [17] which is a free and open source Java framework for building Semantic Web and Linked Data applications

3.2. OntoPRiMa ARCHITECTURE

The architecture of the proposed prototype "OntoPRiMa" consists of the following components as indicated in Figure 3.

- Ontology Handler
- Instances Recognition Rules
- Morphosyntactic & Semantic Analyzer
- Instance Extractor

3.2.1. Ontology Handler

The ontology handler component manages the access to ontology using Jena API [17] that deals with Linked Data, and makes the ontology and its elements accessible and manageable. This component allows the user to visualize and browse the ontology and its elements (triplets, concepts, relations, instances) within the interfaces of OntoPRiMa prototype.

3.2.2. Instances Recognition Rules

The set of instances recognition rules is stored in an XML database. We recall here that we have manually constructed and configured all these rules which allow us to trigger the creation of candidate instances of ontology concepts and relations.

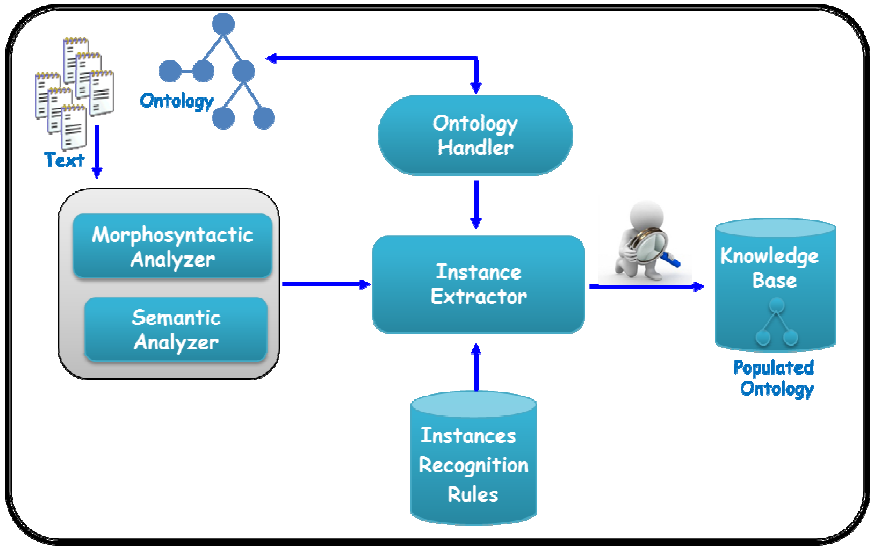

Figure 3.OntoPRiMa Architecture

3.2.3. Morphosyntactic & Semantic Analyzer

The morphosyntactic analysis is performed on the corpus (input text) using the Part-Of-Speech Tagger Treetagger [15]. The analyzer in this component annotates the text with part-of-speech

(noun, verb, adjective, etc.,) and lemma information (canonical form of words). As an output, the morphosyntactic analyzer provides an annotated corpus.

The semantic analysis is performed on the ontology and the results of the morphosyntactic analyzer (annotated corpus). It uses the large lexical database semantic "Wordnet" [16] in order to group the words into sets of synonyms. The interfaces of OntoPRiMa prototype allow the user to visualize the results of these two analyzers.

3.2.4. Instance Extractor

This module extracts the candidate instances and associates a confidence weight to each of them. These instances are validated by an expert via the validation interface where the expert can accept, modify and/or reject the obtained results (the candidate instances). The validated instances are modeled in RDF/XML format, and they are stored in the Knowledge Base of the ontology in order to achieve the ontology population process.

3.3. WORKFLOW FOR ONTOLOGY POPULATION IN OntoPRiMa

Here we are presenting briefly the ontology population processes that OntoPRiMa relies on them. For more information, these processes are deeply described in our previous work [5] & [6].

The ontology population processes in OntoPRiMa pass in three phases:

1. Initialization Phase
2. Semantic Relation Instances Extraction
3. Ontology Population Process

The initialization phase in OntoPRiMa is based on annotating texts with POS and lemma information; it allows filtering the texts of the corpus based on the morphosyntactic categories. It is also dedicated to construct manually the set of Instances Recognition Rules.

After the initialization phase, the ontology is automatically traversed. For each object property or relation R_{ab} (Figure 4) that links two classes C_a and C_b (domain C_a and range C_b of the property) the steps and actions of the second and third phases are triggered.

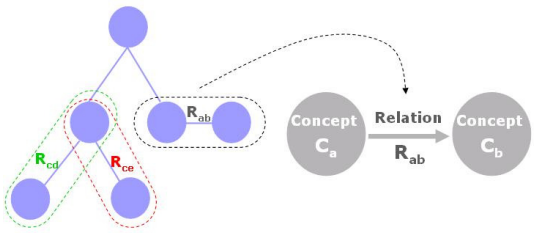

Figure 4. Traversing relations (object properties) during ontology processing

The second phase (Semantic Relation Instances Extraction) aims to extract all possible instances of relations by relying on the predictive power of verbs.

The final phase is devoted for ontology population, it consists of five steps:

1. Triplets Extraction
2. Candidate Instance Extraction

3. Evaluation and Automatic weighting
4. Expert Validation
5. Ontology instantiation

The main algorithm (Triplet Extraction Algorithm) that gathers the complete process that OntoPRiMa relies on is presented in Figure 5.

Main Algorithm : the complete process of the approach
Triplet Extraction Algorithm

```
Populate_Ontology (O , C , Rules)

Variables
Input :
      O : Generic Ontology in RDFS
      C : Corpus of texts in naturel language
      Rules : = {R1,R2,…,Rn} Set of instance recognition rules
Output:
      PO : Populated Ontology with RDF triplets
Begin
      Repeat
         Rab = getCurrentRelation(O)    //Object Property
         Ca =  getDomain (Rab,O)      // LeftConcept
         Cb =  getRange (Rab,O)        //RightConcept

         Call POS_Tagging(C)
         LVab = get_ListVerbs(C,Rab)
         L_Triplet = get_triplets(C,LVab)
         L_Inst_Triplet = get_Instance_triplets(L_Triplet, Rules)

         // Calculation of the confidence weight for each triplet
         Repeat
           Insert (triplet,get_weight(triplet,Ca,Cb,Rab,O)) in
           List_triplet_weight;
           Take next triplet (Instancea,Vab, Instanceb);
         Until reading all the triplet in L_Inst_Triplet;
         Validation by the expert for the List_triplet_weight
         //ontology population
         Repeat
           Call Create_instance_concept(Instancea,Ca,O)
           Call Create_instance_concept(Instancea,Cb,O)
           Call Create_instance_relation(Vab,Rab,Instancea,Instanceb)
           Call GenerateNewRDFtriplet (PO,Instancea,Vab,Instanceb)

           Take next triplet (Instancea,Vab,Instanceb)
         Until reading all the validated triplet in L_Inst_Triplet;

         Take next Relation
      Until reading all the relations (properties)
            of the Ontology O;
END
```

Figure 5.Triplet Extraction Algorithm

4. EXPERIMENTS AND RESULTS

An experiment was designed to validate our proposed approach with the functionalities of the prototype OntoPRiMa in the domain of risk management. The experiment was consisted of populating the ontology of the PRIMA project using 62 chemical fact sheets as textual corpus (written in English natural language) issued by the U.S. Environmental Protection Agency (EPA)[1].

The 62 sheets contain 9150 sentences, approximately 123,250 words including 8,668 occurrences of verbs (Table 1).

Table 1. Statistical figures of the used EPA corpus.

Nb. of sheets	Nb. of sentences	Nb. of words	Nb. of verbs
62	9150	123250	8668

In this experiment we have populated specifically the two RDFS classes “Cause” (as domain) and “Risk” (as range) of the property «Provoke» that links them.

```
<rdf:Property rdf:ID="Provoke">
      <rdfs:domain rdf:resource="#Cause"/>
      <rdfs:range rdf:resource="#Risk"/>
</rdf:Property>
```

Figure 6.“Provoke” Property definition in PRIMA ontology

After launching the extraction process of list of verbs associated with the Property "Provoke" (Table 2), the prototype produces a list gathered from the most frequent verbs in the corpus and the synonyms of the verb "Provoke".

Table 2.List of extracted verbs associated with the Property "Provoke"

Verbs	Nb. of occurrences	frequency of occurrence
expose	441	0.051
occur	298	0.034
cause	178	0.021
increase	178	0.021
report	175	0.020
produce	159	0.018
release	127	0.015
make	126	0.015
expect	121	0.014
do	107	0.012
affect	105	0.012
show	102	0.012
indicate	99	0.011
result	86	0.010
irritate	62	0.007

synonyms of verb “provoke”
Provoke
Arouse
Elicit
Enkindle
Evoke
Fire
Kick_up
Kindle
Molest
Raise
Stimulate

[1] U.S. Environmental Protection Agency www.epa.gov/chemfact

Using the above produced list of verbs, the ontology population processes in remaining phases produce 660 candidate instances triplet and assign the confidence weights for each triplet of them. Above 80% of these triplets are evaluated as acceptable by the domain expert.

For example, for the following sentence: *"Breathing small amounts of biphenyl over long periods of time has caused damage to the liver and damage to the nervous system"* OntoPRiMa generates two candidate triplets in RDF format as mentioned below in figure 7.

```
<prima:Cause rdf:ID="cs-46">
<prima:desc>Breathing small amounts of biphenyl over long periods
of time</prima:desc>
    <prima:Provoke rdf:resource="#rs-101"/>
<prima:Provoke rdf:resource="#rs-102"/>
</prima:Cause>
<prima:Risk rdf:ID="rs-101">
<prima:desc>damage to the liver</prima:desc>
</prima:Risk>
<prima:Risk rdf:ID="rs-102">
<prima:desc>damage to the nervous system</prima:desc>
</prima:Risk>
```

Figure 7 Generated triplets by OntoPRiMa in RDF Format

As a result, we have achieved high precision (above 80%) of knowledge acquisition in this experiment (Table 3).

Table 3.Results & precision of obtained in the experiment

Nb. of candidate instances triplet	**Accepted instances by the expert**	**Precision**
660	540	81.8%

5. CONCLUSION

In this paper, we have implemented our semi-automatic approach for ontology population from texts (published in our previous work [5]) in a prototype application called OntoPRiMa. . OntoPRiMa is dedicated for information extraction and more precisely for ontology population within the framework of the PRIMA project.

The results of the presented experiments in the domain of risk management show that ontology population tasks with OntoPRiMa have reached high precision (above 80%) of accepted Instances triplets. This percentage is satisfactory results encouraging us to go further in populating the remaining elements in PRIMA ontology and in populating other generic ontology in other domain without extensive reworking.

To conclude, the main contribution of this paper can be summarized with the following points:

- We have implemented OntoPRiMa, a prototype for semi-automatic ontology population from texts that relies on a combination of NLP techniques, semantic web technologies and automatic weighting for evaluation and supporting the decisions by experts.
- The adapted approach by OntoPRiMa is independent of the data (independent of the size of the textual corpus), but it remains sensitive to the linguistic structuring of the domain corpus.
- OntoPRiMa is practical and useful for experts and able to provide decision support.

- Finally, OntoPRiMa ensures a real portability from one domain to another, so it is domain independent.

REFERENCES

[1] G. Wohlgenannt, S. Belk and M. Schett, "A Prototype for Automating Ontology Learning and Ontology Evolution," in 5th International Conference on Knowledge Engineering and Ontology Development (KEOD-2013), Vilamoura, Portugal, 2014.

[2] P. Buitelaar, P. Cimiano and B. Magnini, "Ontology Learning from Text: An Overview," in Ontology Learning from Text: Methods, Applications and Evaluation, 2005.

[3] H. Cunningham, K. Bontcheva and Y. Li, "Management and Human Language: Crossing the Chasm," Journal of Knowledge Management, vol. 9, no. 5, pp. 108-131, 2005.

[4] G. Petasis, Karkaletsis, Vangelis, Paliouras, Georgios, Krithara, Anastasia and Zavitsanos, Elias, "Ontology Population and Enrichment: State of the Art," in Knowledge-Driven Multimedia Information Extraction and Ontology Evolution: Bridging the Semantic Gap, Berlin, Heidelberg, Springer Berlin Heidelberg, 2011, pp. 134-166.

[5] J. Makki, A.-M. Alquier and V. Prince, "Ontology Population via NLP Techniques in Risk Management," International Journal of Humanities and Social Sciences (IJHSS, ISSN: 2070-3783), vol. 3, no. Summer 2009, pp. 212-217, 2009.

[6] J. Makki, V. Prince and A.-M. Alquier, "Semi Automatic Ontology Instantiation in the domain of Risk Management," in IIP2008 - 5th International Conference on Intelligent Information Processing, Beijing China, 2008.

[7] "PRIMA Project Risk Management IST-1999-10193," Information Society Technologies (IST), [Online]. Available: http://www.esi2.us.es/prima/.

[8] S. Kim, H. Alani, W. Hall, P. Lewis, D. Millard, N. Shadbolt and M. Weal, "Artequakt: Generating Tailored Biographies from Automatically Annotated Fragments from the Web," in the 15th European Conference on Artificial Intelligence, Lyon, France, 2002.

[9] F. Suchanek, G. Ifrim and G. Weikum, "LEILA: Learning to Extract Information by Linguistic Analysis," in Proceedings of the 2nd Workshop on Ontology Learning and Population: Bridging the Gap between Text and Knowledge – OLP 2006, pp. 18 – 25, Sydney, Australia, July 2006.

[10] R. Navigli and P. Velardi, "Enriching a Formal Ontology with a Thesaurus: an Application in the Cultural Heritage Domain," in Proceedings of the 2nd Workshop on Ontology Learning and Population: Bridging the Gap between Text and Knowledge – OLP 2006, pp. 1 – 9, Sydney, Australia, July 2006.

[11] P. Buitelaar, P. Cimiano, S. Racioppa and M. Siegel, "Ontology-based Information Extraction with SOBA," in roceedings of the International Conference on Language Resources and Evaluation, pp. 2321-2324, May 2006.

[12] D. Brickley and R.V. Guha, "RDF Schema 1.1," W3C, 25 February 2014. [Online]. Available: https://www.w3.org/TR/rdf-schema/.

[13] M. A. Musen, "The Protégé project: A look back and a look forward," AI Matters, vol. 1, no. 4, pp. 4-12, 2015.

[14] Richard Cyganiak, DERI, NUI Galway and David Wood, "RDF 1.1 Concepts and Abstract Syntax," W3C , 25 February 2014. [Online]. Available: https://www.w3.org/TR/2014/REC-rdf11-concepts-20140225/.

[15] H. Schmid, "TreeTagger - a part-of-speech tagger for many languages," University of Stuttgart, [Online]. Available: http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/.

[16] "WordNet a lexical database for English," Princeton University, 2010. [Online]. Available: http://wordnet.princeton.edu/.

[17] "Apache Jena - Jena Ontology API," The Apache Software Foundation, [Online]. Available: http://jena.apache.org.

Authors

Dr. Jawad Makki received the Ph.D. degree in computer science from the University of Toulouse I, France, in 2010. Since 2011, he has been with the Lebanese University where he is currently a senior lecturer in Computer Science and Information Studies at the Faculty of Science and the Faculty of Information. His main research interests are Semantic Web & Ontologies, Knowledge Representation, NLP & Information Extraction, and Social Network Analysis.