Using Natural Language Programming (NLP) Technology To Model Domain Ontology OTO by Extracting Occupational Therapy Concepts

Ahlam F. Sawsaa* and Joan Lu**

*/**School of Computing & Engineering, University of Huddersfield, UK,
13 Ivy Street, Moldgreen, Huddersfield, West Yorkshire, HD5 9 AE
*<a.sawsaa@hud.ac.uk>, **<j.lu@hud.ac.uk>

Ahlam Sawsaa is a post-doctoral researcher in the XML, Database and Information Retrieval (XDIR) Research Group at University of Huddersfield. She received her Ph.D. in computer science from the University of Huddersfield, UK in 2013. Her research interests include knowledge engineering, implementation and application of knowledge representation and reasoning, knowledge-based information retrieval. Particular interests center around tools, techniques and methodologies for developing logic based ontologies, in particular ontologies written in the Web Ontology Language OWL, domain modeling and rule languages, and semantic technologies, information extraction, and machine learning.

Joan Lu is professor in the Department of Informatics. She was a team leader of the IT Department in an industrial company before she joined the university. Her research interests include XML technology, object oriented system development, agent technology, data management system, information access/retrieval/visualization/representation, security issues and Internet computing. She serves as editor in chief for the *International Journal of Information Retrieval Research*.

Sawsaa, Ahlam F., Lu, Joan. **Using Natural Language Programming (NLP) Technology To Model Domain Ontology OTO by Extracting Occupational Therapy Concepts.** *Knowledge Organization. 41(6)*, 452-464. 33 references.

Abstract*:* Creation and development of formal domain ontology of occupational therapy (OTO) requires the prescription and formal evaluation of the results through specific criteria. UPON methontology of development ontologies was followed to create an OTO ontology, and was implemented by using Protégé-OWL. Accuracy of the OTO ontology was assessed using a set of ontology design criteria. This paper describes a software engineering approach to model domain ontology for occupational therapy resources (OTO) using Natural Language Programming (NLP) technology. The rules were written to annotate the domain concepts using Java Annotation Patterns Engine (JAPE) grammar. It is used to support regular expression matching and thus annotate OT concepts by using the GATE developer tool. This speeds up the time-consuming development of the ontology, which is important for experts in the domain who face time constraints and high workloads. The rules provide significant results: the pattern matching of OT concepts based on the lookup list produced 403 correct concepts and the accuracy was generally higher. Using NLP technique is a good approach to reducing the domain expert's work, and the results can be evaluated. This study contributes to the understanding of ontology development and evaluation methods to address the knowledge gap of using ontology in the decision support system component of occupational therapy.

Received: 18 December 2013; Revised: 4 November 2014; Accepted: 4 November 2014.

Keywords: domain ontology, concepts, occupational therapy, natural language processing, NLP

1.0 Introduction

Recently, unstructured data on the World Wide Web has generated significant interest, in the extraction of text, emails, web pages, reports and research papers in their raw form. Far more interestingly, extracting information from a specific domain using distributed corpora from the World Wide Web is a vital step towards creating cor-

pus annotation. The semantic web offers semantic annotations that describe web resources explicitly. These annotations are based on ontologies that represent domain knowledge through defining concepts and the semantic relations between those concepts. At the same time, ontologies have become the main components of the semantic web. The creator of the World Wide Web considers the ontology to be a critical part of the semantic web (Berners-Lee 2000). It requires standards of machine-processable representations of ontology. The standards for this purpose, such as Resource Description Framework (RDF) (Handschuh and Staab 2002), Web Ontology Language (OWL) (Gasevic et al. 2006), have been defined by the World Wide Web Consortium (W3C) (Sawsaa and Lu 2010; Semantic Web 2005). Consequently, ontology is a foundation that is central to the growth of the semantic web, that provides a common knowledge for correspondence and communication among heterogeneous systems. Furthermore, it is useful for different applications to share information among heterogeneous data resources (Alberto et al. 2002).

Information Extracting (IE) has received significant interest due to the number of web pages emerging on the Internet containing unstructured data (Jacob and Zhang 2013). Because of the amount of information available on the Internet, it has become necessary to have a tool for extracting it. It has been reported that specialists in the field of IE have worked to find suitable tools, such as wrappers, that classify interesting data and map them onto appropriate formats such as XML or relational databases (Moens 2006). Furthermore, some HTML-aware tools are based on inheriting the structural features of documents so as to extract the data. Natural Language Programming (NLP) is a technique used by a number of tools, such as GATE and JAPE, to extract the data in natural language documents (Cunningham et al. 2000). Tools like GATE use techniques such as part-of-speech tagging, filtering, or lexical semantic tagging to link relevant information, and identify relationships among phrases and sentence elements within text (Cunningham et al. 2000; Vlachidis et al. 2010). Each of these tools has advantages and disadvantages. A comparative analysis of the existing tools for data extraction is needed to assess their capabilities.

Occupational therapy needs a formal language (FL) to identify certain concepts in the field to make communication easier, but currently its FL is missing. This paper provides an initial definition of FL for the area of occupational therapy, together with a brief background of IE tools to justify the use of NLP technique. The NLP is used to extract concepts in the field and to speed up the process of building the Ontology of Occupational Therapy (OTO). During this study the CREOLE plug-ins from GATE in the IE system is used. It also shows how the JAPE grammar has been implemented by detailing the rules we use to annotate IS concepts. The paper is structured as follows: in Section 2, the background of ontology and IE is discussed. In Section 3, the methods used to develop domain ontology of OT and extracting OT concepts are discussed, and demonstrate how the OT ontology system was constructed. In Section 4 the implementation shows how the domain knowledge is acquired for creating the corpus, Gazetteer, and how the JAPE rule is implemented. In Section 5, discussion and evaluation are included. Finally, in Section 6, conclusions and suggestions for future work are made.

2.0 Background

2.1 Information extracting IE

A number of studies have shown that applications of IE can be used to annotate documents that are written in natural language. Certainly, the growing number of IE tools that can be used to annotate concepts, such as SHOE, Annota, Annozilla, MnM, Ontomat, COHSE, Melita, and GATE, make it easy to process machine-readable text (Srihari and Li, 1999). A comparison of these tools shows that they provide distinct methods of IE (Alberto et al. 2002; Calzolari 2013), as illustrated in Table 1.

Table 1 shows that there are many tools such as SHOE, Annota Annozilla and KIM ontotext providing automatic annotation of extracted text from mark-up languages such as RDF OWL HTML, XHTML (Lassila and Swick 2009; Ibekwe-SanJuan 2010), written by different languages, e.g. Java and C. In comparing the GATE developer with these tools, GATE provides semi-automatic and automatic annotations in easy to use ways, similar to the MnM ontology editor. GATE can extract text from different formats, such as XML, HTML, XHTML, emails and PDF files, while MnM annotates HTML formats only.

Basically, the annotation of IS concepts is based on the GATE developer, which is an architecture tool for text engineering. It is a free open source tool developed by a team at the University of Sheffield, starting in the early 1990s. The first version was released in 1995, the second one was in 2002, and the most recent was released in 2010 (Cunningham et al. 2000).

GATE can run on any platform and supports JAVA 5.0. It has also been developed and tested on Linux, Windows and Mac OS X. It has a user interface to enable user editing and visualization and quick application development. Furthermore, it provides support for manual annotation, semi-automatic and semantic annotation, as well as ontology management. Moreover, GATE uses

Tools	Type	Degree of automation	Based on	Ease of use	Language written in	Advantages & Disadvantages
SHOE	Knowledge annotation	Automatic		+	Java	Allows users to mark up pages in SHOE, guided by ontologies or URL
Annota	Annotation schema W3C	Automatic	RDF mark-up XML,XHTML,CSS &Xpointer	+	C Available for Windows, Unix and MAC	Does not support IE; like an ontology server; makes annotation publicly available
Annozilla	Email annotation	Automatic	Mozilla	++		-
MnM	Ontology editor	Semi-automatic & automatic	HTML	+		Similar to Melita
Ontomat		Automatic	OWL	++		Used to create and maintain ontologies; uses OtoBroker as server
COHSE	Integration of text-processing components	Automatic	DAML+OIL	+	RDF	Uses ontology server to mark up pages in DAML+OIL and reuse as RDF
Melita	Annotation interface	Semi-automatic	Extensible mark-up language, Java, HTML	++		To retrieve structure and semi-structured annotations
KIM ontotext	Semantic annotation platform	Automatic	RDF	++		Semantic annotation, indexing, and retrieval of unstructured and semi-structured content.
GATE	Annotation tool	Semi-automatic & automatic	XML, HTML, XHTML, emails	+++	JAVA version 5	Comprises an architecture and framework. Based on NLP group

Table 1. Information Extracting Tools

CREOLE plug-ins as objects for language engineering. All of these are packaged as Java Archives and XML configuration data (Cunningham et al. 2000).

GATE is a tool used to take unseen texts and convert them into a fixed format such as XML or HTML. This data can then be displayed for users or stored in a database for analysis. Before talking about GATE in more detail, we should clarify the difference between information retrieval (IR) and IE (Crescenzi and Mecca 2004). IE helps the user to extract information from a huge amount of text for the purpose of fact analysis. Information retrieval (IR) is just pulling out documents containing relevant information according to a key word search. Information extracting (IE) can identify queries in a structural way and provides knowledge at a deeper level, while IR uses a normal query engine, which makes it hard to gain accurate answers, and provides knowledge at the standard level.

Consider an enquiry such as 'which UK airports are currently closed due to severe weather conditions?' Or, where an event took place and who it involved, such as 'where was Gordon Brown's last visit as prime minister?' (Chang et al. 2006; Srihari and Li 1999). IR would just provide a webpage containing the relevant information and the user would then need to search that webpage using various terms or concepts to analyse the information. IE, on the other hand, provides specific information about the enquiry, even if the information is not accurate, but you can only back up the correct information (Moens 2006). IE has been used for applications such as text mining, semantic annotation, question answering, opinion mining, decision support, rich information retrieval and exploration.

GATE has a comprehensive set of plug-ins, including Alignment, ANNIE, Annotation_Merging, Copy_Annots _Between_Docs, Gazetteer_LKB, Gazetteer_Ontology _Based, Information_Retrieval, Keyphrase_Extraction_ Algorithm, Language_Identification, Ontology_Tools, and WordNet.

GATE is based on ANNIE, which is a new IE system with core processing resources (GATE 2013). ANNIE relies on a finite state algorithm and JAPE grammar, and combines Tokenisor, which divide the text into simple tokens such as words, numbers, punctuation, Sentence Splitter (which splits the text into sentences), and abbreviations of the gazetteer list to help distinguish between sentences. The POS tagger, Name entity tagger, and JAPE transducer uses JAPE grammar to produce entities, Orthomatcher (co-references), to match rules to identify

the relations between names already found by the POS tagger. The Gazetteer is a list presenting a set of names, terms, etc., to identify entities based on the list. Among these modules, we used Tokenisor, Sentence Splitter, Gazetteer, and JAPE transducer (Handschuh and Staab 2002; Thakker et al. 2009).

GATE includes automatic and semi-automatic semantic annotation as well as manual annotation, which enables the user to create their own annotations. As a result, the GATE developer can be used to extract terms and concepts from a specific text effectively and efficiently. For this work, we annotate OT concepts from literature to be formalized into OTO ontology code.

2.2 Domain ontology

Ontology has been defined from different perspectives. Ontology in philosophy is a branch of metaphysics dealing with being or reality; otherwise, it is study of beings or existence. In the computer science and IA community it is defined as a specification of conceptualization (Tifous et al. 2007; Bhatt et al. 2009; Marcondes 2013). The ontology concept has been developed over time; in 1993 Tom Gruber described it as follows: "a formal explicit specification of shared conceptualization" (Gruber 1993). Practically, ontology is a method of representing items of knowledge even when these items are facts, ideas, and things. Smith (1996) asserted that ontology is formal theory within not only definitions but also supporting framework of axioms.

The domain ontology is a specific area of knowledge containing the main concepts and their relations. Gomez-Perez asserted that this kind of ontology has weaknesses including emerging upper-level ontology. It classifies its concepts according to different criteria, which leads to heterogeneity in knowledge. The domain ontology is the solution of specific concepts in each domain. The numbers of studies of ontology has grown rapidly in recent years. Pettey and Goasduff (2006) points out that integration of the semantic web could have the greatest impact on technology in the next few years particularly, in a specific domain such as biomedical ontologies which play a fundamental role in accessing the heterogeneous sources of medical information, and using, and sharing patient's data (Pettey and Goasduff 2006; Gòmez-Pérez et al. 2004; Sowa 2012).

A significant growth of ontology projects shows that a lot of ontologies have been developed and designed based on different methodologies. The methodology is still a big problem in this area. Many methodologies have been proposed since the 1990s. Deokattey et al. (2010) described a method for developing a domain ontology in the multidisciplinary area of accelerator driven systems in nuclear physics. "The merging system was used and a new knowledge organization tool was developed using an existing larger tool; the International Nuclear Information System (INIS) thesaurus." Furthermore, it is a shared belief that ontology receives a lot of recognition from various research fields. Although there are some well-known domain ontologies, such as CYC, the Standardized Nomenclature for Medicine (SNOMED, a clinical terminology) (Jepsen 2009), Toronto Virtual Enterprise (TOVE) (Laboratory 2011) "TOVE Ontology Project," and the GENE ontology (GO) (Gasevic et al. 2006), study of the ontology area is still immature and improvements are needed (Lassila and Swick 2009). GO was developed by the National Human Genome Research Institute in 1998. It presents a controlled vocabulary of gene and gene product attributes. It contains 30,000 concepts and is organized as follows: cellular component, molecular function, and biological process. It is regularly updated and is available in several formats (Gasevic et al. 2006; GeneOntology 2009; Jepsen 2009).

2.3. Occupational Therapy (OT)

Occupational therapy (OT) is a treatment using certain activities to improve physical, mental and social performance. The main objective of the treatment is development of the individual's personal autonomy, both social and professional. The therapist is one of a multi-disciplinary team working with a stroke patient, for example. OT needs a formal language to identify certain concepts in the field to make communication easier between therapists and between people and machines. "Occupational therapists seem to have difficulties of finding appropriate words with which to think about and express the nature and purpose of their practice" (Breines 2005). Although many attempts have been made to find a full agreement on specific concepts in the area, the problem still remains of identifying the key concepts and their relationships (Cunningham et al. 2000). However, ontology in a specific domain provides concepts and the relationships between these concepts. Furthermore, it offers shared vocabularies in a consistent mode. More crucially, in natural language, therapists use different meanings for one concept, for instance "occupation, function." If you looked in a dictionary you might find more than one definition.

Ontology specifies a formal definition to avoid vagueness and ambiguity, to decide the exact meaning. If you looked in the *Oxford Dictionary of English* for the meaning of occupation you will find many definitions. At the same time, many therapists, social and health staff, use different words and phrases to define occupation, such as daily activities, everyday life, persons of any age, individual and sociocultural value, leisure and so on; so ontology provides a single definition and more accurate restriction by axioms makes communication more effective. However, "ontology incorporates both a subject representational

vocabulary and a bibliographic description format, and can be made compatible with any digital information resource in a library or any webpage from the Internet" (Deokattey et al. 2010).

3.0 Methods employed

3.1 Designing at theoretical model with NLP

The construction ontology model is a real scenario and complicated task. Moreover, there is no unique method of modelling ontologies. Modelling ontological knowledge needs ontological engineering to ensure implicit knowledge is formalized in different ways. The ontological engineering provides methodologies for constructing ontology modelling. The development of OTO ontology followed Methontology (Gòmez-Pérez et al. 2004). Methontology is a general methodology framework. The OTO ontology model was designed to represent a knowledge domain of occupational therapy. The designed model was informed by the background information gathered and annotated by Natural Language Programming (NLP) technique. The model developing process consisted of the following activities: Specification and Conceptualisation. The outcome of these activities was a model prototype that has been tested and evaluated.

The OTO ontology is designed to support the semantic web and fill the knowledge gap between the therapeutical knowledge that is represented in OTO. It contains a set task to help the therapists assess the patient. The therapists need different types of knowledge to understand their professional roles. It also helps therapists to explain what they are doing for patients and among staff. This knowledge assists in carrying out the function of occupational therapy in different settings, and understanding clients' problems, what types of treatment are needed, and what service will be provided (Kielhofner 2009). For example, referral of a patient to a therapist worker requires some details about the current treatment and the main purpose of the appointment - perhaps he/she just needs evaluation or follow ups. The ontology aims to provide concrete details that are needed.

Otherwise, therapists need knowledge to:

- Define the nature, purpose, scope, and value of occupation therapy
- Understand problems and their solutions
- Relate their work to other fields such as medical science.

3.2 NLP technique used to extract OT concepts:

We present an automatic extraction method based on ANNIE using a JAPE grammar that extracts concepts from XML files and HTML text. Our JAPE rule extracts concepts as follows: the first entity detected is Information service {*Type=Token, start=867, end=837, id=4210, majorType=concept*} labelled as *Occupational Therapy.concept***.**

```
Phase: one
Input: Lookup Token
Options: control = applet

Rule: concept1
Priority: 20
(
({Token.string == "occupational"})
{Token.string == "therapy"}
({Lookup. major Type == "concept"})
) : occupational
-->
: occupational. concept = {Rule=concept1}
```

In these rules, we specify a string of text {Token.string == } that must be matched, specifying the attributes of the annotation by using operators such as "==", and then annotating the entities according to the correct labels. Furthermore, using a control field such as all, applet, brill gives the right results. The next example shows how regular expressions could be annotated as showing concepts related to (Therapy) metacharacter (dot, *, [ ], |), {Token.string == "therap (ies)"}.

Our method of creating OTO ontology is based on UPON (Unified Process for Ontology) methodology. UPON is based on a software engineering process. It is mainly for large-scale domain ontology but provides useful guidelines for small ontologies. UPON consists of cycles, phases, iteration and workflows and is distinguishable by its use case driven, iterative and incremental nature (De Nicola et al. 2009).

The following steps are proposed in this methodology: Inception, Elaboration, Construction, and Transition. With each iteration different workflows come into play, and a richer, more complete version of the ontology is produced.

Requirements workflow—specifying the semantic needs and user view of the knowledge to be encoded in the ontology—starts with:

- Determine the domain interest, which is occupational therapy.
- Define the main reason for developing ontology and identify intended users. In this case developing OTO is to provide a model of occupational therapy resources that could be used within the domain to assists therapists in their work. OTO ontology will provide a formal representation of the concepts and describe the relations between those concepts.

– Writing the sketches outlining the sequence of activities that take place in a particular scenario. In OTO, ontology is not required because of the simplicity of the sequence activities.
– Create application lexicon by collecting terms from the domain knowledge. Most of the terminology is collected from online resources. Table 2 shows the lexicon for OTO ontology.

Occupation	Role	Ability
Occupational mapping	Routine	Activity
Occupational performance	Setting	Assessment
Occupational components	Skills	Autonomy

Table 2. Lexicon for OTO ontology

Identify competency questions for which ontology must provide the answer. The competency question is essential to be identified for use in the workflow to evaluate the ontology.

– What is the name of a specific treatment?
– Does the required treatment need a long or short time to apply?
– Is the treatment needed pre-assessment?

Modelling related use cases correspond to paths through the ontology to answer competency question. These models assist users in searching in the ontology to find answer to the competency questions, add and remove elements in the modelled ontology, as illustrated in Figure 1.

The use case graph shows how the ontology can be searched. Therapists, patients, and service administrative staff can search the ontology by problem, classifications of main components; for example searching on clients need, intrapersonal experiences or practice decisions.

Before creating the ontology, we had to collect the OT concepts for the domain model. Our approach consisted of annotating these concepts based on the JAPE grammar, using the GATE software. The annotation process began by:

– Creating a corpus of documents and a lexicon of OT, with JAPE rules used to extract OT concepts. GATE

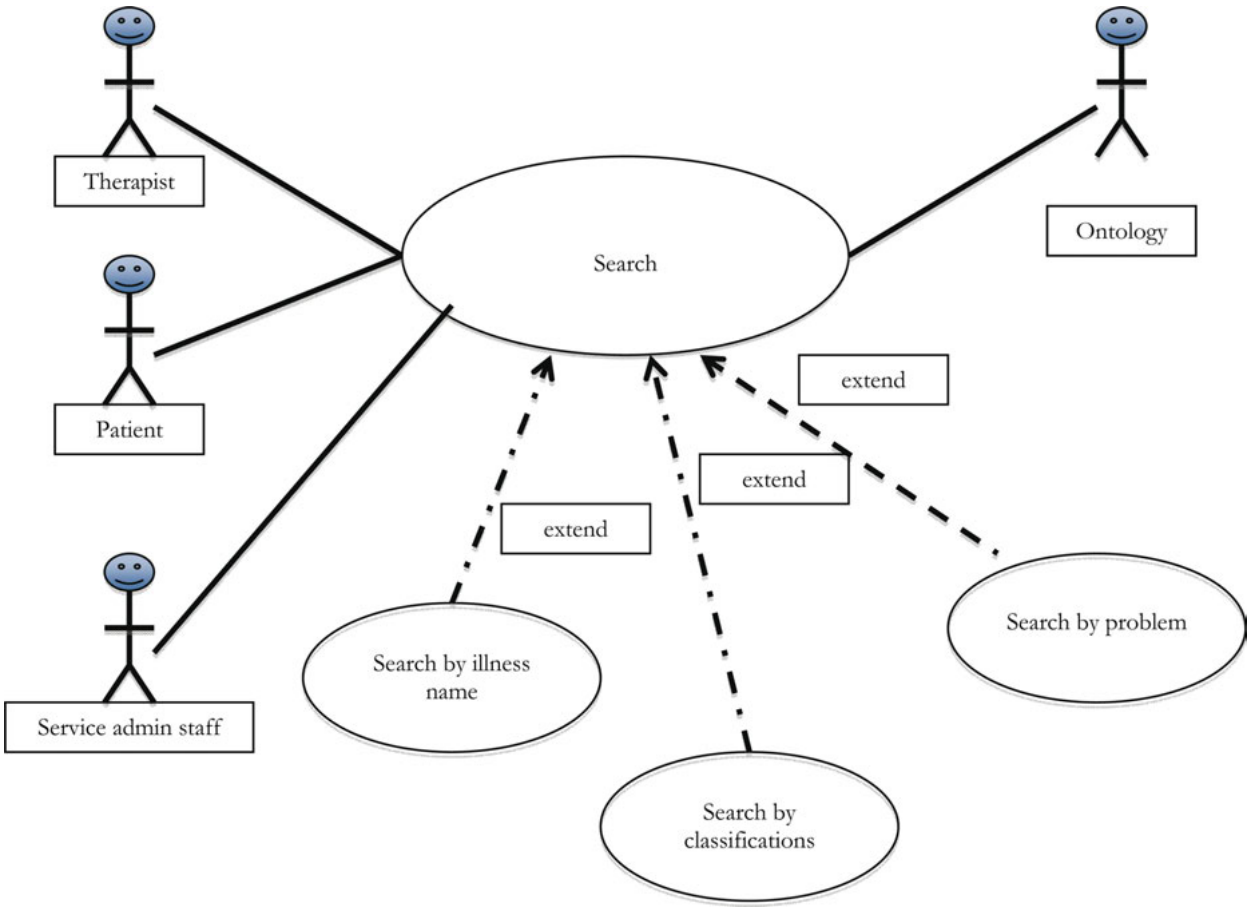

Figure 1. Use case for searching the OTO ontology

provides facilities for loading corpora for annotation from a URL or uploading from a file. The process generally started as follows:

– We compiled OT knowledge from different resources and various publications.

– We analysed the data to ensure it covered the whole field.

– We transferred the information resources into an XML file to form the corpus.
– We uploaded the corpus into the GATE software to start running ANNIE.
– We annotated the concepts based on JAPE grammar, which is run within ANNIE.
– Testing and evaluation, as illustrated in Figure 2.

Analysis workflow—refinement and structuring of the ontology requirements identified in previous workflow needs to be completed by following these steps:

Acquiring domain resources to create a domain lexicon through collecting domain terms such as:

Therapist	Patient
Client	Therapy action
Therapy	Initial assessment
Referral or reason for contact	Information gathering
Activities of daily living	Play activities
Working and productive activities	Internal influence
Performance components	Values

However, to attend to specific problems the relationships that provide more precise details should be identified; in OTO ontology, many taxonomic relationships have been built, e.g., Is-A, hasA. Also, some relations among concepts are non-taxonomic relations used to present within the relation concept.

From these terms, the potential relations between the concepts will be created of OTO ontology. Therapists should have professional beliefs and values about the nature of people and the nature of health and beliefs about the nature and purpose of OT. The OT ontology provides clear relationships between clients, whether individual, group or organisation, each of which has to be influenced by internal and external factors.

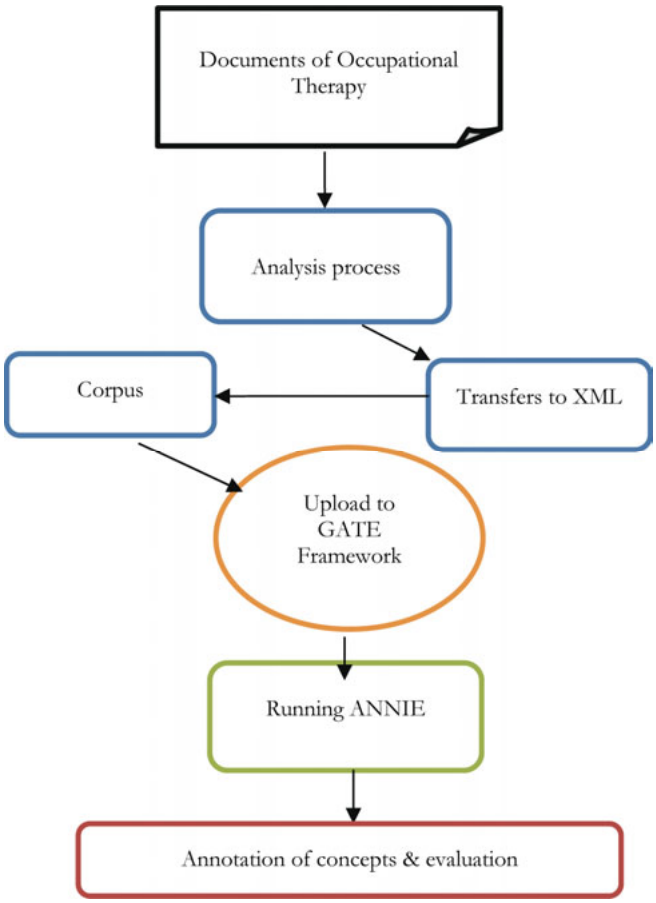

Figure 2. Annotation workflow

Meanwhile, each client has a history, experience, language, social and culture, physical, thoughts, beliefs, values, aspirations, needs, problems, interest, behaviours, abilities and environment. They connect with a therapist who has values, identity, attitudes and capability. More relationships are depicted in Figure 3.

– Modelling the application activities using UML diagrams to explain those activities involved in searching, adding, and moving elements in the ontology.
– Building the reference glossary by defining the terms.

Term	Definition
Occupational mapping	First stage in the process of developing a national framework, as it analyses the professional areas covered.
Therapy	Treatment intended to relieve or heal a disorder
Client	A person or group that uses professional advice or services.
Occupation	A job or profession: people in professional occupations

Table 3. Reference Glossary

Design workflow—refinement of entities, actors and processes identified in the analysis workflow, including identification of their relationships.

Implementation workflow—selecting and encoding the ontology in formal language.

4.0 Implementation

Implementation identifies the main goal of the OTO ontology, which is seeking to provide information for therapists that helps them to:

– Provide valuable information about the patients.
– Provide information on daily service and activities.
– Enable effective occupational performance.

Knowledge acquisition is started by:

– Creating the corpus. It has 300 files in the XML format, containing text relevant to the IS field.
– Creating the IS Lexicon: This is a list of OT terms that have to be identified as Major Type, Minor Type, etc. For example:

Productivity: major type= concept
Personal Care: minor type= term
Activities: major type= concept

Next we use JAPE rules:

Using JAPE rules allows us to extract concepts and identify tokens that contain the concepts in the correct

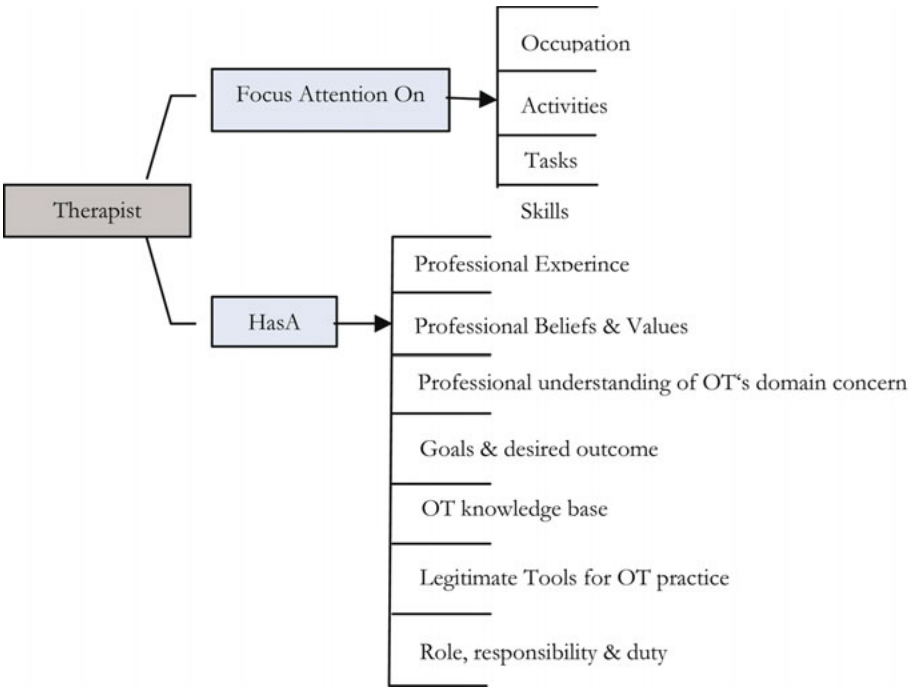

Figure 3. Relationships in OTO ontology

order, and then look up the concepts in the Gazetteer. JAPE rules create a phase based on Java for creating specific grammar. Each JAPE rule consists of Left-Hand-Side (LHS) that contains the pattern that must be matched, and Right-Hand-Side (RHS) that details the annotations that are to be created (Cunningham, et al. 2000).

The JAPE grammar was used to support regular expression matching, as this is how annotation is achieved in GATE. Annotation can also be carried out using other CREOLE plug-ins such as Gazetteer, for which it is necessary to create a list of concepts to be annotated. By clicking on the ANNIE Gazetteer, all the lists appear, including the OT list, as shown in Figure 4. The sub-list presents the main concepts, such as action, task, interactions, social role and developmental.

The next step was uploading the corpus to the application framework, using the JAPE grammar and Gazetteer to match and annotate concepts from the corpus. All terms were gathered, along with the relevant information, to be described in the lexicon. We collect the terms from the point of view of occupational practice and therapists; see Table 4.

Occupation	Role	Ability
Occupational mapping	Routine	Activity
Occupational performance	Setting	Assessment
Occupational components	Skills	Autonomy

Table 4. Lexicon for OTO ontology

5.0 Building OTO computational model

A conceptual model of the OT in natural language needs to be modelled. The primary output of this stage is OTO ontology, which is structured in the appropriate ontology editor, such as Protégé. The OTO ontology is structured in natural language to be suitable for data modelling and knowledge representation. It indents for expression of unambiguous and complete specification of domain concepts with relations between them, and organises them in super-types and sub-types of hierarchy. Furthermore, ontology in Protégé can be exported to different formats such as RDF and XML; Figure 5 shows part of OTO ontology in OWL language. The Organization concept is a superclass, and a subclass of Client.

In the formal model of OTO ontology that covers the whole domain, the main classes are categorised based on fundamental considerations about the role of concepts in the area. The ontology is implemented by organising concepts according to is-A, part-of and hasA relations with axioms. That gives specific definitions of these concepts.

The OTO ontology allows the users to explore the ontology structure by browsing the upper level of the tree. The upper level provides a general understanding of the OT domain, whereas the deeper levels can be reached when they are navigated to through multiple levels of the tree.

The upper-level of classes contains abstract entities created based on taxonomy of OT and the philosophical approach of science definition. Formally, the OTO model in-

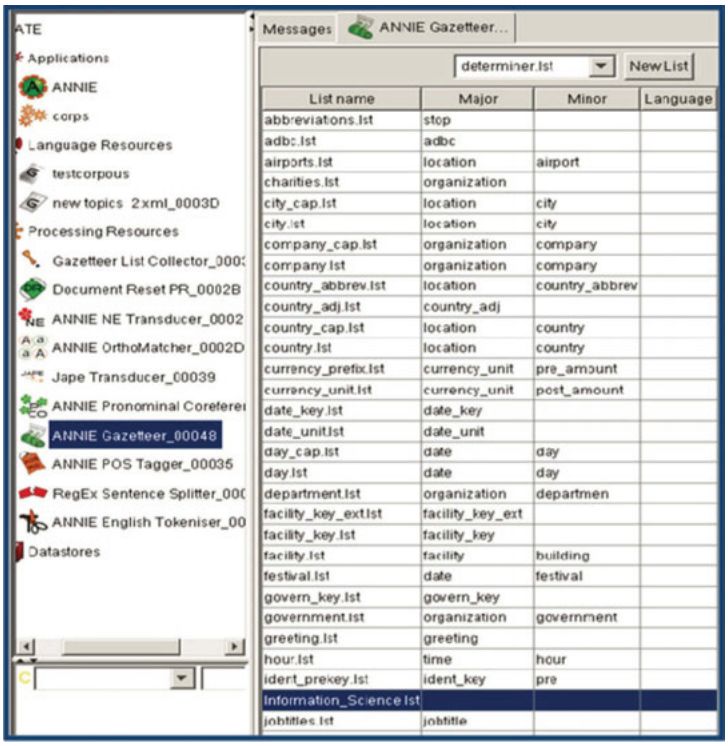

Figure 4. Screenshot of OT Gazetteer

cludes fourteen levels of representation, which provide the foundation of knowledge framework for the OTO ontology. The OTO ontology root classes are: Action, Client, PerformanceArea, PerformanceComponent, Service, Therapist, and Therapy Time, Space. The root classes are hierarchically specialized; each sub class is grouped under a main class, for instance "TherapyAction," was grouped under the Therapy class, as shown in Figure 6. The OTO ontology structure is extendable and flexible.

5.1 OTO Components

5.1.1 Classes

Classes in OTO ontology (also called concepts) are a type of object in the real world, e.g. the class "Action" models the class of all tools that are used in the domain to facilitate doing and providing services. Classes in OTO ontology are defined to be unique by their definitions. Classes have too many relationships to each other. The relation type indicates that a class has a relationship with other subclasses by specific relations.

5.1.2 Axioms

Ontology has axioms which are basic statements; these axioms represent a basic knowledge, e.g. <owl:Class rdf: about="ActivityOfAdailyLiving#"> <rdfs:subClassOf rdf: resource="PersonalHygiene#"/> ActivityOfAdailyLiving class is a subclass of the PersonalHygiene class – it is an axiom.

5.1.3 Properties

In the OTO ontology, relationships are called properties in OWL and some other description logic languages. The attributes are created in object properties - Owl: Object property - and data property view - Owl:Data Type property. The object property is the relationship between instances, whilst data property describes the relationships between instances and data values, which link an instance to RDF or to XML schema. The data property is similar to the object property unless it can be just functional in characteristic, not inverse in description, as shown in Figure 7.

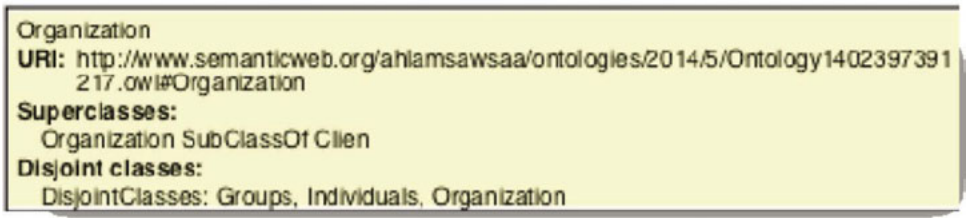

Figure 5. OTO ontology in OWL language

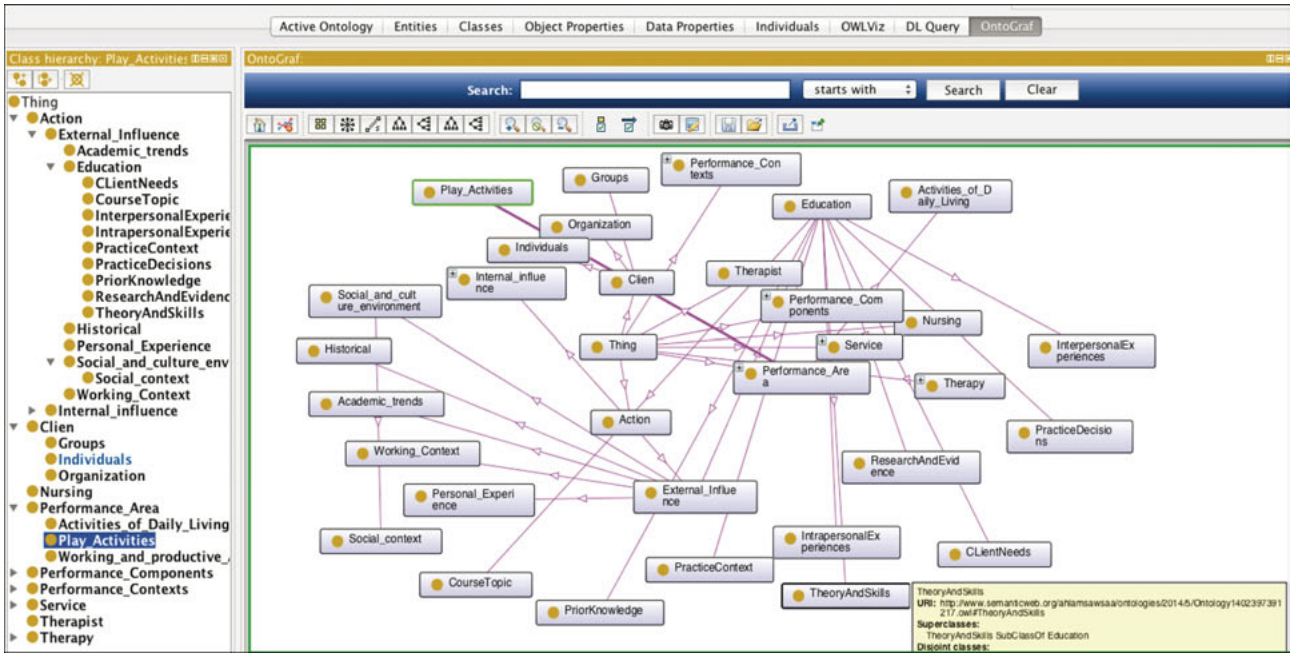

Figure 6. OTO ontology in Protégé

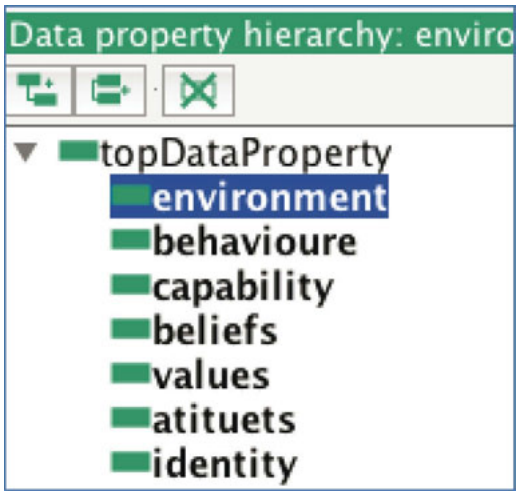

Figure 7. Data Property

5.1.4 Individuals

Ontology Instances in Protégé are called individuals of classes that are created in the individuals view. Each instance can be described in the description tab as the Type and name of the same individual. The instance 'Institute of Speech and Language, and Communication' is described under types as a ActionPlanning and the same name is SLCNs. Attribute is allocated in data property assertion, and the relations under the object property assertion.

6.0 Evaluation and Discussion

The OTO ontology is based on IE techniques for extracting the main concepts in the domain. Our extraction of OT concepts, using JAPE grammar and regular expression based on the GATE developer for automated IE, provides significant results. The main idea behind using JAPE and regular expression is to identify OT terminology as tokens, for example occupational therapy, activity and daily living, from a large text. The term identification relies on looking up a list of OT terms from the Gazetteer. For example, we could look up collecting information, analysis of information, interpreting information or social norms, social expectation, cultural norm, cultural expectation. These concepts can be collected to be the main component of the OT glossary and to structure in semi-formal hierarchy before creating the computational model of the OTO ontology.

We extracted the OT concepts from a corpus of 300 documents, obtained specifically for this purpose. We ran the ANNIE application, using document reset, Tokenisor, Sentence Splitter, Gazetteer, POS tagger, JAPE transducer and Orthomatcher. The annotation set that appeared in the display panel and the concepts are highlighted in the annotation default; each annotation has a different colour after running ANNIE and highlighting the matching concepts. The results show that our approach successfully annotates concepts. We recalled 541 of the *Occupation* concept, 275 of the *Therapy* concept and 35 of the *Habit* concept. Each annotation starts from a specific point and ends at a different point based on how many tokens it has. The *Occupation* concept starts at point (557) and ends at (566), while the *Therapy* concept starts at (624) and ends at (636), with its features {major Type=concept}.

In this study, the data were evaluated based on evaluation metrics for precision, recall and the F-measure, which are common metrics in IR field that can be defined as follows:

$$Precision = \frac{Correct + \frac{1}{2}\,Partial}{Correct + Spurious + Partial}$$

$$Recall = \frac{Correct + \frac{1}{2}Partial}{Correct + Missing + Partial}$$

$$F - Measure = \frac{(\beta 1 + 1)\mathrm{P} * \mathcal{R}}{(\beta 2\ \mathrm{P}) + \mathcal{R}}$$

The statistics of the corpus show that the pattern matching of OT concepts based on the lookup OT list was 403, correct concepts and accuracy were generally higher, and there were no partially correct results (0), or missing false positives, as illustrated in Figure 8.

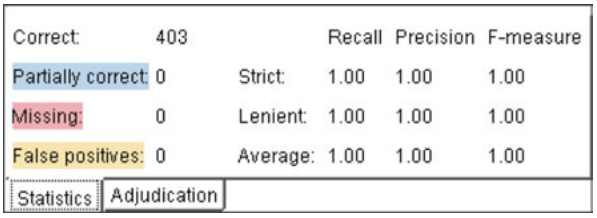

Figure 8. The Accuracy results

The OTO ontology model represents OT therapist searching needs, and has potential to help social studies researchers find primary sources. The OTO ontology followed Methontology (Fernández-López et al., 1997) ontology methodology engineering as a general framework to construct the ontology model, and it was evaluated based on ontology quality criteria to ensure its clarity and completeness.

7.0 Conclusion and future work

The development of OTO ontology is a stage towards creating a shareable and reusable OT system. Ontology is

a formal model for therapists and social staff organising vision in the information hyperspace of the domain knowledge. This paper has described a method using NLP techniques to extract concepts for the purpose of developing OTO ontology. Furthermore, the development of the IE system should save time and effort in labelling the most common concepts. In total, we extracted 664 concepts that are classes of the OT, and 650 subclasses making up the main components of the ontology skeleton. The IE technique can be applied to many different formats, such as XML, HTML documents, URLs or emails. The OTO ontology was developed based on UPON methodology. In developing the ontology, the process of ontology development has been illustrated with the ontology capturing concepts and relation associated with different aspects of the OT domain.

Ontology is at the heart of the semantic web. It defines concepts and relations that make global interoperability possible. In future work, we plan to enhance more concepts to develop OTO, to be applied in specific applications to solve specific problems related to the OT domain. Overall, the OTO ontology will increase the value of the data stored in the ontology that will allow it to be shared and reusable across applications.

References

Alberto, H., Berthier, A., Altigran, S., and Juliana, S. 2002. A brief survey of web data extraction tools. *Sigmod Record* 31: 84-93.

Berners-Lee, Tim. 2000. "Semantic web on xml." Presentation at XML 2000, Washington DC, December 06 2000. http://www.w3.org/2000/talks/1206-xml2k-tbl/slide1-0.html.

Bhatta, Mehul, Rahayub, Wenny, Sonib, Sury Prakash and Woutersb, Carlo. 2009. Ontology driven semantic profiling and retrieval in medical information systems. *Web semantics: science, services and agents on the World Wide Web* 7: 317-31.

Breines, Estelle. 2005. *Occupational therapy: activities for practice and teaching*. London: Whurr.

Calzolari, Nicoletta. 2013. *The people's web meets NLP*. New York: Springer.

Chang, Chia-Hui, Kayed, Mohammed, Girgis, Moheb Ramzy and Shaalan, Khaled. 2006. A survey of web information extraction systems. *IEEE transactions on knowledge and data engineering* 18: 1411-28.

Crescenzi,Valter and Mecca, Giansalvatore. 2004. Automatic information extraction from large websites. *Journal of the ACM* 51: 731-79.

Cunningham, Hamish, Maynard, Diana and Tablan, Valentin. 2000. *JAPE: a java annotation patterns engine. (2nd ed.). Research memorandum Cs-00-10*. Department of Computer Science, University of Sheffield, November 2000. Accessed January 2013. http://gate.ac.uk/sale/tao/splitch13.html#x18-32300013.

De Nicola, Antonio, Missikoff, Michele and Navigli, Roberto. 2009. A software engineering approach to ontology building. *Information systems* 34: 258-75.

Deokattey, Sangeeta, Neelameghan, Arashanipalai and Kumar, Vijai. 2010. A method for developing a domain ontology: a case study for a multidisciplinary subject. *Knowledge organization* 37: 173-84.

Fernández-López, Mariano, Gómez-Pérez, Asunción and Juristo, Natalia. 1997. Methontology: from ontological art towards ontological engineering. *Proceedings of the AAAI'97 Workshop on Ontological Engineering, Spring Symposium Series*. Stanford, pp. 33-40.

Gaševic, Dragan, Djuric, Dragan and Devedžic, Vladan. 2006. *Model driven architecture and ontology development*. Berlin: Springer.

GeneOntology. 2009. Welcome to the gene ontology website! Accessed April 20 2013, http://www.geneontology.org/.

Gómez-Pérez, Asunción, Fernandez-Lopez, Mariano and Corcho, Oscar. 2004. *Ontological engineering: with examples from the areas of knowledge management, e-commerce and the semantic web*. London: Springer.

Gruber, Thomas R. 1993. Toward principles for the design of ontologies used for knowledge sharing. *International journal of human-computer studies* 43: 907-28

Handschuh, Siegfried, Staab, Steffen. 2002. Authoring and annotation of web pages in CREAM. *Proceedings of the 11th International World Wide Web Conference, Honolulu, Hawaii, May 7-11 2002*. pp. 10-24.

Ibekwe-SanJuan, Fidelia. 2010. Semantic metadata annotation: tagging Medline abstracts for enhanced information access. *Aslib proceedings* 62: 476-88.

Jacob, Elin and Zhang, Guo. 2013. Role of virtual boundaries in knowledge sharing and organization. In Smiraglia, Richard P., ed., *Proceedings of the 3rd North American Symposium on Knowledge Organization* 4: 122-30.

Jepsen, Thomas C. 2009. Just what is an ontology. *IT professional* 11: 22-27.

Kielhofner, Gary. 2009. Conceptual foundations of occupational therapy practice. Philadelphia: F. A. Davis Co.

Laboratory, M. 2011. Tove ontology project. Toronto: University Of Toronto. Accessed February 2013, http://www.eil.utoronto.ca/enterprise-modelling/tove/.

Lassila, Ora and Swick, Ralph R. 2009. Resource Description Framework (RDF) model and syntax specification. Cambridge W3C Recommendation 22 February 1999. *World Wide Web Consortium*. Accessed January 2013, http://www.w3.org/tr/rec-rdfsyntax/.

Marcondes, Carlos Henrique. 2013. Knowledge organization and representation in digital environments: rela-

tions between ontology and knowledge organization. *Knowledge organization* 40: 115-22.

Moens, Marie-Francine. 2006. *Information extraction: algorithms and prospects in a retrieval context*. New York: Springer.

Pettey, Christy and Goasduff, Laurence. 2006. "Gartner's 2006 Emerging Technologies Hype Cycle Highlights Key Technology Themes." *Gartner, Inc.* Accessed December 2014. http://www.gartner.com/newsroom/id/495475

Sawsaa, Ahlam and Lu, Joan. 2010. Ontology of information science based on owl for the semantic web. *Proceeding of International Arab Conference on Information Technology (ACIT'2010)*. Benghazi, Libya: University of Garyounis, pp. 145-61.

Semantic Web. 2005. Semantic web activity statement W3C. Accessed February 2013, http://www.w3.org/2001/sw/activity.

Smith, Barry. 1996. Mereotopology: A theory of parts and boundaries. *Data and knowledge engineering* 20: 287-303.

Sowa, John F. 2012. Ontologies, last modified November 29 2010. http://www.jfsowa.com/ontology/index.htm.

Srihari, Rohini and Li, Wei. 1999. Information extraction supported question answering. In *Proceedings of the Eighth Text Retrieval Conference (TREC-8),* pp. 310-19. Accessed January 2013, http://www.dtic.mil/cgibin/gettrdoc?ad=ada460042&location=u2&doc=gettrdoc.pdf.

Thakker, Dhaval, Osman, Taha and Lakin, Phil. 2009. Gate JAPE grammar tutorial. Accessed September 2013. https://gate.ac.uk/sale/thakker-jape-tutorial/GATE%20JAPE%20manual.pdf.

Tifous, Amira, Ghali, Adil El, Dieng-Kuntz, Rose, Giboin, Alain, Christina, Christina and Vidou, Géraldine. 2007. An ontology for supporting communities of practice. *Proceedings of the 4th international conference on Knowledge capture, Whistler, BC, Canada*. New York: ACM, pp. 39-46.

Vlachidis, Andreas, Binding, Ceri, May, Keith and Tudhope, Douglas. 2010. Excavating grey literature: a case study on rich indexing of archaeological documents by the use of natural language processing techniques and knowledge based resources. *Aslib proceedings* 62: 466-75.