Knowledge Management Research & Practice (2009) 7, 206–217
© 2009 Operational Research Society. All rights reserved 1477–8238/09
www.palgrave-journals.com/kmrp/

Codifying collaborative knowledge: using Wikipedia as a basis for automated ontology learning

Tao Guo[1]
David G. Schwartz[2]
Frada Burstein[1] and
Henry Linger[1]

[1] Faculty of Information Technology, Monash University, Melbourne, Australia; [2] Graduate School of Business Administration, Bar-Ilan University, Israel

Correspondence: David G. Schwartz, Graduate School of Business Administration, Bar-Ilan University, Ramat-Gan, Israel.
E-mail: dschwar@mail.biu.ac.il

Abstract

In the context of knowledge management, ontology construction can be considered as a part of capturing of the body of knowledge of a particular problem domain. Traditionally, ontology construction assumes a tedious codification of the domain experts knowledge. In this paper, we describe a new approach to ontology engineering that has the potential of bridging the dichotomy between codification and collaboration turning to Web 2.0 technology. We propose to shift the primary source of ontology knowledge from the expert to socially emergent bodies of knowledge such as Wikipedia. Using Wikipedia as an example, we demonstrate how core terms and relationships of a domain ontology can be distilled from this socially constructed source. As an illustration, we describe how our approach achieved over 90% conceptual coverage compared with Gold standard hand-crafted ontologies, such as Cyc. What emerges is not a folksonomy, but rather a formal ontology that has nonetheless found its roots in social knowledge.

Knowledge Management Research & Practice (2009) **7,** 206–217.
doi:10.1057/kmrp.2009.14

Keywords: ontology; Wikipedia; Web 2.0; social knowledge; collaboration

Introduction

Many writers have pointed out (for example, Latour, 1987; Star & Griesemer, 1989; Lave & Wenger, 1991; Star, 1995) that work is a social activity that involves people interacting with objects in order to solve complex problems. Moreover, such work usually involves different groups each with a particular perspective on the domain. Any work domain is characterized by a body of knowledge (BoK) that is constructed from the practices of the community engaged in that work. This characterisation views the (re)construction of the BoK as dynamic, 'situated, collective and historically specific' (Bowker & Star, 1999).

In the context of knowledge management, ontology construction can be considered as a part of capturing of the BoK of a particular problem domain. With an increasing amount of research examining ways in which ontology can be used in knowledge management (Holsapple & Joshi, 2002; Sure *et al.*, 2002; Sure, 2003; Buchholz, 2006), there is no question that ontology is one of the core technologies for knowledge classification and codification.

Classification tends to standardise information or naturalise it (Bowker & Star, 1999) with the consequence that one group imposes its perspective on all other actors engaged in that work. Over a longer period of time such a classification becomes the BoK for the community (Latour, 1987), thus

Received: 19 September 2008
Revised: 21 April 2009
Accepted: 3 June 2009

effectively allowing the community to collectively forget the 'contingent, messy work' (Bowker & Star, 1999, p. 299) that underpinned the creation of the BoK that the classification represents. However, such classification systems are not capable of maintaining the integrity of the information in terms of the practices and social process of its creation. Moreover, they assume a strong structure that may not readily support the application of the information to the local needs and a context of a particular situation. This further extenuates the imposed standardisation.

Construction of an ontology requires categorisation work (Bowker & Star, 1999) that subsumes the multiple perspectives of actors and creates a knowledge object representing a component of BoK. Such work also needs to manage the intrinsic tension and dynamics of this goal. Classifications and ontology comprising categories are the means to share information between the different group perspectives, and through time and space.

The contradictions that exist between classification systems and the imperatives of practice need to be addressed in order to construct workable support tools that have validity through space and time. One approach to overcome such contradiction is through collaboration. Using the theoretical lens of *community of practice* (CoP) and its inherent concepts of boundary objects, situated learning and legitimate peripheral participation creates potential for a dynamic creation of BoK (Lave & Wenger, 1991; Bowker & Star, 1999). This approach means that a work domain can be explained from multiple perspectives and allows for the creation of boundary objects that span those perspectives and facilitates different groups to effectively operate in the work domain. As operation of a CoP is effectively a learning process, it presumes that the BoK that underpins the work of the community is continually evolving. One important addition to this lens is an ecological approach (Bowker & Star, 1999) that allows actors in the work domain to be active interpreters of the BoK without the 'primacy for any one viewpoint' (Star & Griesemer, 1989, p. 389). Thus, each actor is able to translate the BoK to their specific situation while maintaining the integrity of the BoK.

Collaborative knowledge creation efforts are assisted by the plethora of new collaborative technologies and platforms. Social technologies developed as part of Web 2.0 provide an appropriate infrastructure for collective construction of a BoK. The overriding objective of such collaboration is to give the community a voice that is not mediated through self-appointed gatekeepers of the BoK. A collaborative approach to creating a knowledge repository democratises the BoK by removing the category of 'expert'. Instead, knowledge authorship is attributed to actors engaged in a collective, social process whose contributions are based on their contexts and experience (Mika, 2005). Moreover, this model of authorship appropriates categorisation work by a social network in order to express community semantics that incorporate instances of the lived experiences of the community (Gillmor, 2004; Udell, 2004). As a result, 'folksonomies' are created as distributed classification systems by a group of individuals through their free tagging for personal retrieval and social sharing (Mathes, 2004; Vander Wal, 2004). Web 2.0 technologies provide the infrastructure to support a range of modalities for collaboration such as, for example, 'citizen journalism' (Glaser, 2006). Wikipedia1 is a good example of the largest collaboratively created knowledge repository on the Internet and is potentially a fertile source of social knowledge for domain ontology creation.

In this paper, we argue that traditional approaches to ontology construction that rely on expert input and published documentation are inconsistent with the dynamic needs to enable situated action. Such approaches require significant effort to address multiple practices and different perspectives of the work domain and often fail in supporting knowledge sharing in time and space. We study an alternative ontology learning technique, which should be more efficient, sufficiently accurate and workable from an engineering perspective.

We propose an innovative approach to ontology engineering that has the potential of bridging the traditional dichotomy between codification and collaboration through creative use of the knowledge management technology of Web 2.0. By shifting the primary source of knowledge from the expert to socially emergent bodies of knowledge created as a result of Web 2.0 development, we have identified the potential of using collaborative knowledge, rather than brittle expert knowledge, as the basis for ontology construction.

Our approach includes a semiautomated ontology learning capability from collective, collaborative and dynamically constructed BoK coded in Wikipedia. We demonstrate application of our approach to the creation of an ontology for venture capital that is based on Wikipedia entries. We compare this approach with handcrafted ontologies to highlight the utility and promising potential of our approach to address the contingencies of knowledge management to support dynamically changing productive, social and cultural practices within the work domain.

Codification for a domain ontology

Creation of the domain ontology is firmly located in the realm of codification. That is, the attempt to access human knowledge and formally model, then codify, that knowledge in some machine computable form. Balconi *et al.* (2007) suggest that codification of knowledge begins with articulation leading to the representation of knowledge in a language that may be understood by two or more actors. The codification process involves a determination, usually by one or more specific experts, of the relevance and relationship of a concept to the domain in question. Generally, a formal structured language is sought to represent the codified knowledge with the

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

overriding goal being to transform the competencies of a domain into propositional knowledge (Balconi, 2002).

The traditional approach to ontology construction assumes sourcing the domain knowledge of real world experts, authoritative documentation and established protocols to produce a broad, highly structured domain ontology through codification (Gómez-Pérez *et al*., 2004). This approach to creating and maintaining a useable ontology is a complex and time-consuming task. It is a process that belays the social foundation of domain knowledge and is based on an epistemology that is questionable in a sense that knowledge is assumed to be codifiable in advance.

On the practical side, the traditional approach, based on manual codification, creates a bottleneck and results in an ontology of questionable quality and long-term value (Uschold, 1996; Holsapple & Joshi, 2002; Sure, 2003; Pinto *et al*., 2004). Moreover, manual codification efforts are fraught with risks, such as incompleteness, brittleness and rigidity, and validity over time as it is becoming outdated without continuous maintenance and review.

Domain ontology differs from generic ontology. Generic ontology, usually referred to as upper ontology (Niles & Pease, 2001), is generic and abstract without the inclusion of specific domain concepts. By contrast, domain ontology has the following characteristics (Sabou, 2006):

1. addresses the specification of domain concepts
2. reflects conceptual terms accepted by a large domain community
3. has extensibility to the new knowledge
4. can be integrated with other domain ontology or generic ontology.

As we discussed above, creation of ontology always comes with the time constraint and the limited availability of domain experts. Therefore, an ontology learning technique is necessary to overcome these limitations.

Manual codification of domain knowledge from experts

Domain ontologies often represent codified knowledge of a small number of domain experts. The drawbacks of this approach are apparent, where the limited knowledge of a few domain experts can never represent comprehensively and precisely the broader societal knowledge of a certain domain within a given time frame. The disadvantages of employing a small group of experts to build ontology have been discussed widely (Ratsch *et al*., 2003; Pinto & Martins, 2004). Some alternative approaches tend to rely on collaboration where the main advantage is the distribution of human effort rather than any effective relief of costly labour involvement. Examples of these approaches include international research initiatives, such as Cyc (Lenat & Guha, 1990), open mind common sense (Singh *et al*., 2002) and verbosity (Von Ahn, 2006) that aimed to codify a common sense knowledge.

The problem of the knowledge acquisition bottleneck in the traditional approach has been addressed by introducing (semi)automatic learning techniques for building ontology from existing data (Cimiano, 2006). Mining knowledge from a document corpus is a popular approach to knowledge codification (Hearst, 1992). Such an approach involves collecting, extracting and acquiring knowledge from existing textual resources. In fact, there exists a great amount of domain corpus in various disciplines that can serve this purpose. However, the major challenge is guaranteeing the correctness, currency and consistency of such machine-acquired knowledge. The resulting knowledge requires significant fine-tuning and adaptation to be applicable to a real world project. To build a high-quality domain ontology, even with a predefined formal engineering methodology, the manual approach requires interaction with domain experts with deep understanding of ontology representation languages.

The following sections present a review of several state-of-the-art approaches to address the challenges of automated ontology creation. We use the term *ontology learning* to denote ontology creation from textual resources.

Ontology learning for automated knowledge codification

Buitelaar *et al*. (2005) suggest ontology learning as the process of knowledge acquisition from text. In the development of ontology, ontology learning enables (semi) automatic support for defining a structure and instantiating a knowledge base.

Ontology learning is a complex task that involves multiple axiomatisations at different levels (Cimiano, 2006). In other words, the expressivity-based categorisation of ontologies raises the requirements of ontology learning when moving from the level of terms through synonyms, concepts and concepts hierarchies to the level of relationships between the concepts, which results in its increased complexity. A 'layered cake' representing subtasks of ontology learning is depicted in Figure 1. This figure also shows the increasing complexity from the bottom to the top of each layer and the examples of the objectives of each subtask.

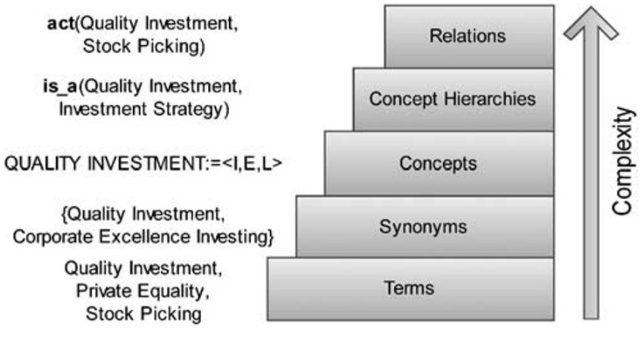

Figure 1 Ontology learning 'layered cake'.

The ontology learning task introduced above is a systematic challenge, as for each subtask there could be multiple solutions. However, critical in determining the approach for an ontology learning task and the possible quality of output is the question of what textual input is to be used, particularly whether to use unstructured or semistructured text. Thus, we shall categorise the different ontology learning approaches into two groups: ontology learning from unstructured text and ontology learning from semistructured text.

The ontology learning techniques, which are mostly derived from natural language processing (NLP) research, are based on statistical analysis of a large quantity of text corpus (Cimiano, 2006). The fundamental issue with this approach is the significant impact of the various quality of the corpus. To overcome this problem, it was proposed to use the World Wide Web as the much bigger source for text corpus than any existing one. For example, Cimiano *et al.* (2004) in PANKOW, while analysing an individual web page, use Google^{TM} web services as the query system for counting the hits (weight) of a generated hypothesis phrase. Moreover, web page indexes and outlines represent a certain degree of structure of text, hence using web pages can be useful for both types of ontology learning, e.g., unstructured as well as semistructured text.

Problems with ontology learning

Ontology learning has a large number of unsolved problems. For example, we first review two well-known ontology learning methodologies in order to analyse their common and different problems. The approach for a domain ontology learning proposed in this paper aims to overcome these problems.

Sabou (2005) suggests the approach of extracting domain concepts from the textual description of the services or the documentation of code in the Web service software. This approach is domain independent, which means that the technique is applicable in any application domain of the Web service. Also, the underlying text-mining techniques achieved remarkably good precision. However, even though the approach does not rely on any manually built training data, as with most machine learning techniques, the extraction patterns for identifying domain concepts and other training parameters must be manually predefined. Moreover, as the predefined extraction patterns have rather high coverage, further defining of more fine-grained patterns is necessary. WordNet, a lexical database of English terms that has been created and maintained at the Cognitive Science Laboratory of the Princeton University (Fellbaum, 1998) could be used for defining such extraction patterns. However, failing to utilise WordNet for synonym detection, this approach is rather weak and the integration to upper ontology requires additional effort. The absence of a formal engineering methodology also hampered the real world deployment of this approach.

Another approach for ontology learning is presented by STAR Lab (Reinberger & Spyns, 2005). The so-called unsupervised text mining proposed in this approach is based on the text mining of a literature collection. The outcome of this approach is a complex semantic network consisting of domain concepts and conceptual relationships between them. Ultimately, the semantic network is created for use by DOGMA (Jarrar & Meersman, 2008), an ontology engineering approach. The whole procedure from unsupervised ontology learning to ontology engineering is a promising method to a practical ontology development cycle. This method is closely related to our approach, differing primarily on the aspects of selection of source corpus and text-mining algorithms.

One of the common problems in the previous studies is that the widely used NLP technique of extracting concepts and relations relies on carefully and labour intensively predefined semantic patterns of common word usage. Quantity and quality of the selected text has significant impact on the result of extraction. As in the previous method, WordNet is not used during the mining process. Thus, it becomes hard to achieve consistency and coherency of an ontology that is generated through the aforementioned approaches. As we will demonstrate, a general and effective ontology learning approach based on social knowledge can act unseeded with no predefined domain restrictions. Moreover, as most formal methods of ontology learning, they ignore the dynamic nature of knowledge and a social mechanism of its construction.

In summary

Codification is a necessary condition to enable machine usable ontology construction. However, complete knowledge codification is problematic. Among others, Johnson *et al.* (2002) argue that the codification must distinguish between knowledge about the world (know what) and knowledge in the form of skills and competence (know how). They state that the dichotomy between codifiable and non-codifiable knowledge is problematic as it is rare that a BoK can be completely transformed into codified form without losing some of its original characteristics and that most forms of relevant knowledge are mixed in these respects. We agree with their statement that codification does not always imply progress and that simply extending the definition of what is codified and possible to codify will have limited practical implication unless the social and applied relevance of the knowledge being codified can be established and dynamically maintained. In the next section, we consider how the above-stated problems can be overcome by introducing collaboration in the process of ontology creation.

The collaborative approach to knowledge ontology construction

An alternative approach to codification is a collaborative approach that involves the joint creation, authorship, revision and refinement of a structured BoK as the basis for a domain ontology (Farquhar *et al.*, 1995; Farquhar *et al.*, 1997; Holsapple and Joshi, 2002; Sure *et al.*, 2002).

Such socially constructed knowledge can be termed collaborative knowledge as a specialisation of social knowledge.

As stated earlier, ontology codification relies on individual expertise and judgement as well as authoritative texts. Such knowledge represents a particular perspective, usually expressing the dominant paradigm of the domain. In contrast, collaboratively created knowledge is dynamic, changing with the inputs and experiences of those that participate in a given domain description.

The social nature of collaboration leads to an almost synonymous relationship between the terms *social knowledge* and *collaborative knowledge*. Yet, collaborative knowledge implies a common goal of collaboration to establish an agreed-upon base of knowledge, which may not be the case when dealing with social knowledge. Social knowledge can evolve devoid of common goals shared by participants in the social discourse.

In that sense, social knowledge can be more fluid, branching and abstract than the more focused collaborative knowledge efforts. Wikipedia represents the social and collaborative efforts of knowledge creation that is expressed in unstructured documents and semistructured forums. Such knowledge is dynamic, often changing as the consensus of the participants change.

The importance of collaboration to capture socially constructed knowledge has been recognised and used in a variety of domains (Ferneley *et al*., 2002; Adamides & Karacapilidis, 2006). However, those efforts are focused on dissemination and business processes, respectively. Cross *et al*. (2001) discuss how social and collaborative technologies can serve as a solid basis for having groups create and share agreed upon conceptualisations for knowledge management. Efforts to date have focused on enabling collaboration by using ontology as a knowledge management tool. At the same time, building an ontology through the input of multiple participants is based on collaboration to construct shared language and documented meaning of such collaboration. Consistent organisational knowledge management and implementation of organisational learning depends on the success to a large extent on the success in such collaboration.

The integration of collaborative or social knowledge into codified knowledge management efforts has received considerable attention. Different strategies, such as brainstorming and collaborative user profiling, have been explored as enhancing the effectiveness of codification (Ferneley *et al*., 2002). Perversely, a major obstacle in the path of automated ontology learning remains the inherent need to codify. However, codification of collaborative knowledge provides a 'best of both worlds' methodology in which the social nature of collaborative knowledge can be effectively and efficiently integrated into knowledge management systems based on codified knowledge. The emergence of social technologies provides an opportunity for collaborative knowledge codification, which can be used as alternative sources for domain ontology creation.

Wikipedia as collaborative knowledge

Ontology creation based on collaboratively created knowledge presumes on-going human interaction about the topic areas.

To meet the criteria of domain ontology, a selected semistructured data source should have the following features (Ponzetto & Strube, 2007):

1. domain independent – has a large coverage of human knowledge
2. up-to-date – reflects the latest knowledge
3. multilingual – to process information in different languages

The World Wide Web has all these features. As a system of interlinked documents, it can be seen as the biggest semistructured document repository of its kind. Text-mining techniques, to exploit this repository, are a very attractive and challenging research area. All the evidence shows enormous hidden knowledge value behind document interlinkages. In fact studies of the linkage and patterns of the World Wide Web as a knowledge base have already been carried out (Markert *et al*., 2003; Etzioni *et al*., 2004).

When using World Wide Web sources, Wikipedia, in particular, could be the prime candidate to illustrate how the core terms and relationships of a domain ontology can be distilled from its dynamic and temporal collaborative knowledge. Wikipedia is '… the world's largest encyclopaedia available on the Web at www.wikipedia.com.' and created collaboratively by a community of interests using wiki software … (http://www.pcmag.com/encyclopedia/). As such, for the purposes of this paper, it can be considered as a suitable example of a semistructured collaborative data source for the ontology creation. The promising features of Wikipedia, discussed below, support our position that it effectively represents collaborative knowledge and can therefore successfully play the role of a domain ontology source in the absence of direct access to domain expertise.

Comprehensive knowledge base

Despite the public criticism of Wikipedia's unauthoritative contents, academic research confirmed that Wikipedia's collaborative editing approach has achieved a remarkably high quality of its content (Giles, 2005). The coverage of human civilisation, in its English version, yields more than two million articles. Relatively speaking, it is not only a significant achievement made by an individual website or the Internet, but a valuable summary of human intellectual history collected through the effort of a large group of people, with a varying level of expertise and social status.

Effective data storage

Unlike other forms of published compendia, Wikipedia is in electronic form and thus has the advantage of underlying database technology, as it stores knowledge in an efficient and logical structure as shown in Figure 2. Each page in Wikipedia has different attributes and elements. Among these data models, most of the links, such as redirects, category links and page links, are cross-referenced.

From this figure, it is clear that Wikipedia fits the description of a semistructured text for ontology learning.

Adaptable application programming interface (API)

The web-based software running behind Wikipedia, Wikipedia API was introduced in August 2006 with the new release of MediaWiki (http://www.mediawiki.org). It facilitates accessing its backend database at a high level abstract (http://www.mediawiki.org/wiki/API). The various APIs facilitate the collaborative construction of the entries and also provide the query modules that allow third-party applications to retrieve most of the data related to articles from Wikipedia's database for further processing.

In this sense, Wikipedia allows 'legitimate peripheral participation' (Lave & Wenger, 1991), in the sense that through collaborative contribution to its content, 'newcomers' can contribute to the construction of BoK of the CoP and enhance their expertise as a result of such collaboration. Thus, Wikipedia is an inclusive social technology allowing anyone to participate in multiple roles and, especially, as authors and editors.

The above description of Wikipedia justifies its use as a collaborative knowledge base for ontology learning. In the next section, we describe our innovative approach of ontology creation from semistructured sources as found in relevant Wikipedia pages. It also provides an illustration of the application of the proposed approach to creating ontology for venture capital as the domain, as one of the authors has expertise and considerable experience in that area.

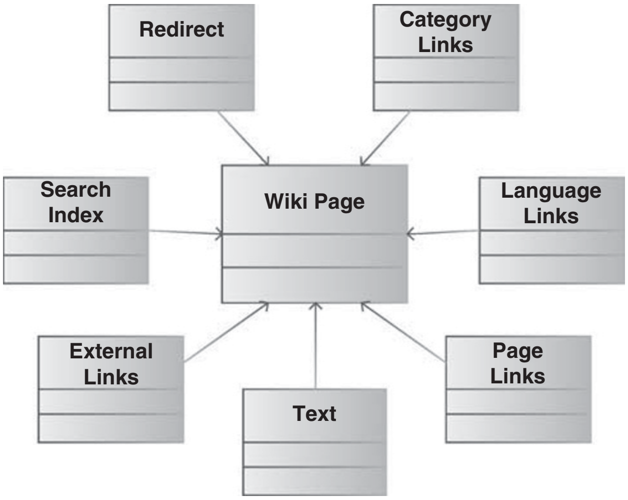

Figure 2 An overview of Wikipedia database structure (adapted from http://www.mediawiki.org/wiki/File:Mediawiki-database-schema.png).

Proposed approach: from Wikipedia to ontology base

In using Wikipedia as a collaborative knowledge base, our aim is not a folksonomy, which is a form of taxonomy created by a collaborative effort, but rather a formally codified ontology with concepts and relationships formally described, that has nonetheless found its roots in collaborative social knowledge. As one of the major goals of creating an ontology for knowledge management is to provide relevant domain knowledge structures, we expect that socially created terms would be highly relevant to end users, as these were derived as a result of collaboration of multiple authors and editors, who over a period of time create, review and cross-reference knowledge objects as part of their contribution to Wikipedia creation.

The ontology learning framework consists of three steps as shown in Figure 3. Although the actual algorithms used are beyond the scope of this paper, we provide the example of concept extraction to indicate the nature of the technical strategy.

Following this process, we first define the concepts and then find the hierarchical relationships among them. Hierarchical relation among concepts forms the *backbone taxonomy* (Guarino & Welty, 2000) of an ontology. The extraction of hierarchical relation employs machine learning and NLP techniques for information extraction approach as suggested by Suchanek *et al*. (2006). For the last step, building arbitrary relationships, the involvement of the domain expert is still necessary. However, the process is more efficient as we provide candidate concepts to domain experts for building *ad hoc* relations derived by

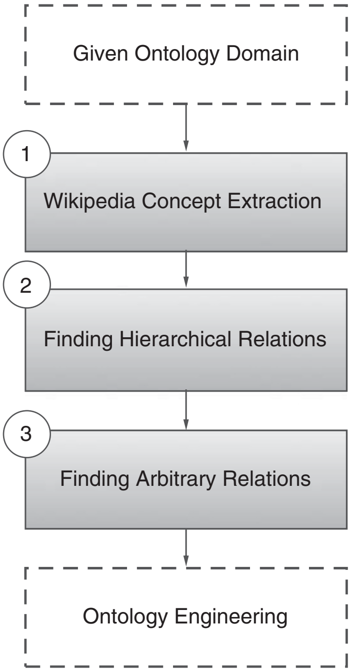
Given Ontology Domain
1
Wikipedia Concept Extraction
2
Finding Hierarchical Relations
3
Finding Arbitrary Relations
Ontology Engineering

Figure 3 Workflow of ontology learning.

our system as part of the ontology learning process from Wikipedia pages. A system generates a list of candidate terms by ranking the closeness of the candidate term with the subject term of the paper.

Here, we illustrate how the concepts are extracted in a chosen domain. In Wikipedia, a 'domain', as the sphere of concepts, is modelled as a category. When a domain is given for ontology development, we first translate the name of domain into the name of a Wikipedia category. In this context, 'Venture Capital' is given as our example domain. As a first step, we find a category named 'Category:Venture capital' in Wikipedia. The page of 'Category:Venture Capital' shows that there are 29 articles categorised under the category, 'Angel Capital', 'Angel Investor' and 'Carried Interest' and so on. A typical form of noise during the concept extraction is the 'List of' page, such as in the 'Venture Capital' category, 'List of Chicago venture capital companies' and 'List of venture capital firms', which do not represent any substantial concepts in the domain. Therefore, removing all pages with 'List of' as the start of the name can avoid such noise.

Clearly, the coverage of any domain ontology should not be based on a basic descriptive article, which in this example contains barely 29 concepts. In fact, in the real world, most of the concepts belong to more than one subdomain of knowledge. To discover the correlated categories to 'Venture Capital', we applied an algorithm, which sorts page categories by the number of pages contained in each. By eliminating the first category, which should be the domain category itself, we came up with a list of category names sorted by closeness to the domain category. In this example, the resulting list appeared to be:

- Private equity: 11,
- Economics and finance stubs: 6,
- Financial terminology: 4,
- Indian company stubs: 1, and so on.

The algorithm then traverses the list. From the list we derived that 'Private Equity' is closest to 'Venture Capital', and thus, we expand our concept extraction to the 'Private Equity' category. In order to achieve optimal results, the concept extraction process is applied in an iterative way as depicted in Figure 4.

Figure 4 also illustrates the extraction of four core corresponding elements of concepts from ontology learning: *concept name*, *concept definition*, *concept synonyms* and *relevant concepts*. We use *concept definition* to acquire hierarchical relations. *Relevant concepts*, then, can be used to support building *ad hoc* relations, a complex issue that will not be dealt within the scope of this paper.

In our approach, we take advantage of Wikipedia API, which stores information appropriate for performing query and editing tasks on any Wikipedia entry. We used its query functionalities for information retrieval about ontology concepts and relationships between them. Among a number of generic formats, we selected XML as the output format with the consideration of a wide range of programming libraries for manipulation. The following table lists all the properties used by our application to retrieve concepts from Wikipedia API (Table 1).

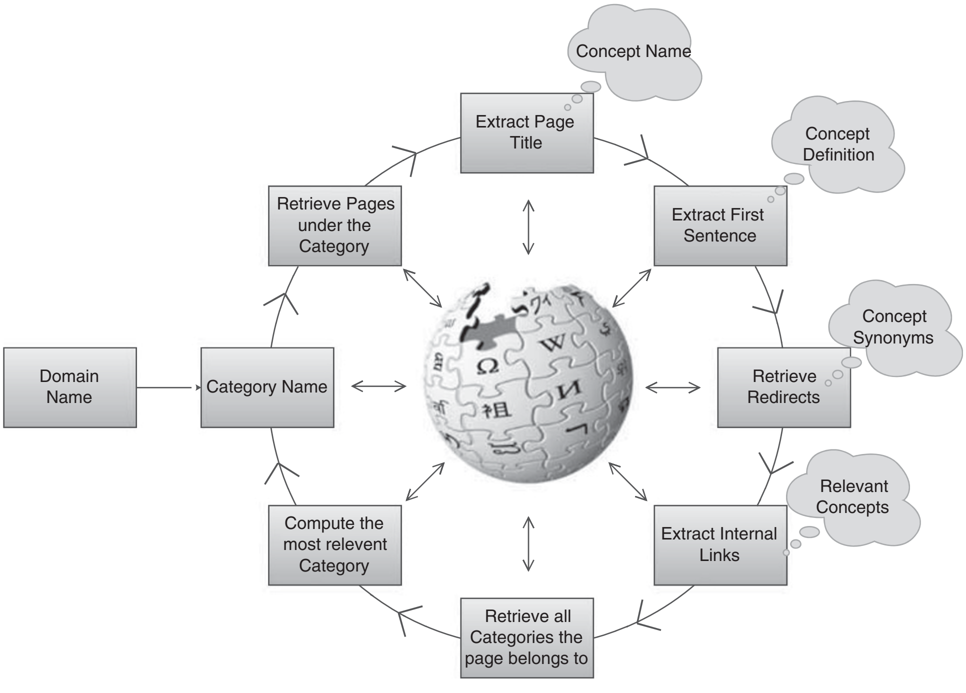

Figure 4 Iterative process of concept extraction.

Table 1 Wikipedia API properties

Property name	*Usage*
Namespace	Namespace indicates the partition of different type of information. For example, '0' represents Wikipedia articles, whereas, '4' denotes internal information about Wiki.
Categories	Categories return all the categories, which a given article belongs to.
Links	Links list all the links in a Wikipedia article.
Backlinks	Backlinks return all the other contents linking to a given article.
Category members	Giving a category name, all the members can be listed. Use with the combination of namespace property can list member articles or subcategories.

Application and evaluation of the proposed approach

The proposed ontology learning method is not fully automatic, but a semiautomatic approach, with minimal human effort in order to deliver high-quality domain ontology. Compared to traditional ontology development, even with modern ontology engineering methodology (Jarrar *et al.*, 2003), to generate an ontology of the size of 200 concepts with consensus, requires a dozen domain experts to work intensively in a group. In the proposed approach, consensus can be achieved by harvesting public knowledge, which does not require direct domain expertise. Rather, basic training to understand a few semantic relations is expected to be sufficient for anyone to become an ontology engineer.

In our test case, a graduate computer science student, who had no domain knowledge about venture capital, business or marketing, was employed as an ontology engineer to perform an ontology development task. Specifically, he developed a web interface for mining Wikipedia API in the way described in the previous section. As an illustrative example, he looked at developing an ontology for 'business/marketing plan'. Wikipedia was used intensively and exclusively for semiautomatic knowledge acquisition by both machine and engineer.

At the first stage, Wikipedia's web interface was used to acquire the necessary background knowledge about the domain by navigating through related articles. Once the big picture was formed, the ontology engineer started the process of building the ontology from the root concept – marketing plan. From there, by using our application, he made his own decisions to fetch the related concepts either from categories or articles. Most of the concepts included in the resulting ontology were identified by the ontology engineer while reading key articles from Wikipedia. When building the ontology, some of the required concepts were already locally acquired from Wikipedia. The missing concepts were then iteratively fetched during the building process, as represented in Figure 3. The relations between the concepts extracted in this way, were used as a guide for the domain expert, who was responsible for the final decisions to accept them in the resulting ontology.

The initial resulting ontology tree is shown in Figure 5 with 49 concepts and 40 relations. It was produced using the proposed approach in a 1-h session using no other sources, but Wikipedia for seeding concepts.

Social ontology comparison

For evaluation of the result by comparison with Gold standard ontology, we first selected ResearchCyc, the research version of the Cyc knowledge base (Lenat & Guha, 1990). ResearchCyc consists of more than 300,000 concepts and three million assertions developed in 25 years, starting from 1984, with the aim of capturing all common sense knowledge as an ontology. Despite criticism of its over complexity, scalability and incompleteness (Bertino *et al.*, 2001), Cyc is by far the biggest manually crafted ontology with human-controlled quality. Hence, it is rather common to see Cyc being used widely as Gold standard for ontology learning evaluation (Ponzetto & Strube, 2007; Suchanek *et al.*, 2007; Zirn *et al.*, 2008). Creating Cyc was a monumental effort in social ontological engineering.

The second Gold standard used for comparison was WordNet (Fellbaum, 1998), which is the most widely used lexical resource for NLP. According to the latest statistics, WordNet 3.0 has 155,287 terms within 117,659 synsets (synonym sets), and using WordNet as a machine-readable dictionary, some ontology learning techniques are aimed at expanding WordNet automatically (Snow *et al.*, 2006).

Both Gold standards have wide coverage of common sense knowledge. Therefore, we compared the coverage of concepts extracted from Wikipedia with the existing concepts in the two Gold standards in order to generate an index to measure the result of conceptual coverage (CC). We use C_{all} to denote all the concepts in the developed ontology. $C_{extracted}$ denotes the number of concepts resulted from the proposed approach. In the same way, C_{GS} denotes the number of relevant concepts in the selected Gold Standard ontologies. Hence, the CC indexes are calculated as follows.

For extracted concepts,

$$CC_{extracted} = \frac{C_{extracted}}{C_{all}}$$

For Gold standard concepts,

$$CC_{GS} = \frac{C_{GS}}{C_{all}}$$

Gold standard comparison results

The initially generated ontology using the proposed approach produced 49 core concepts from Wikipedia.

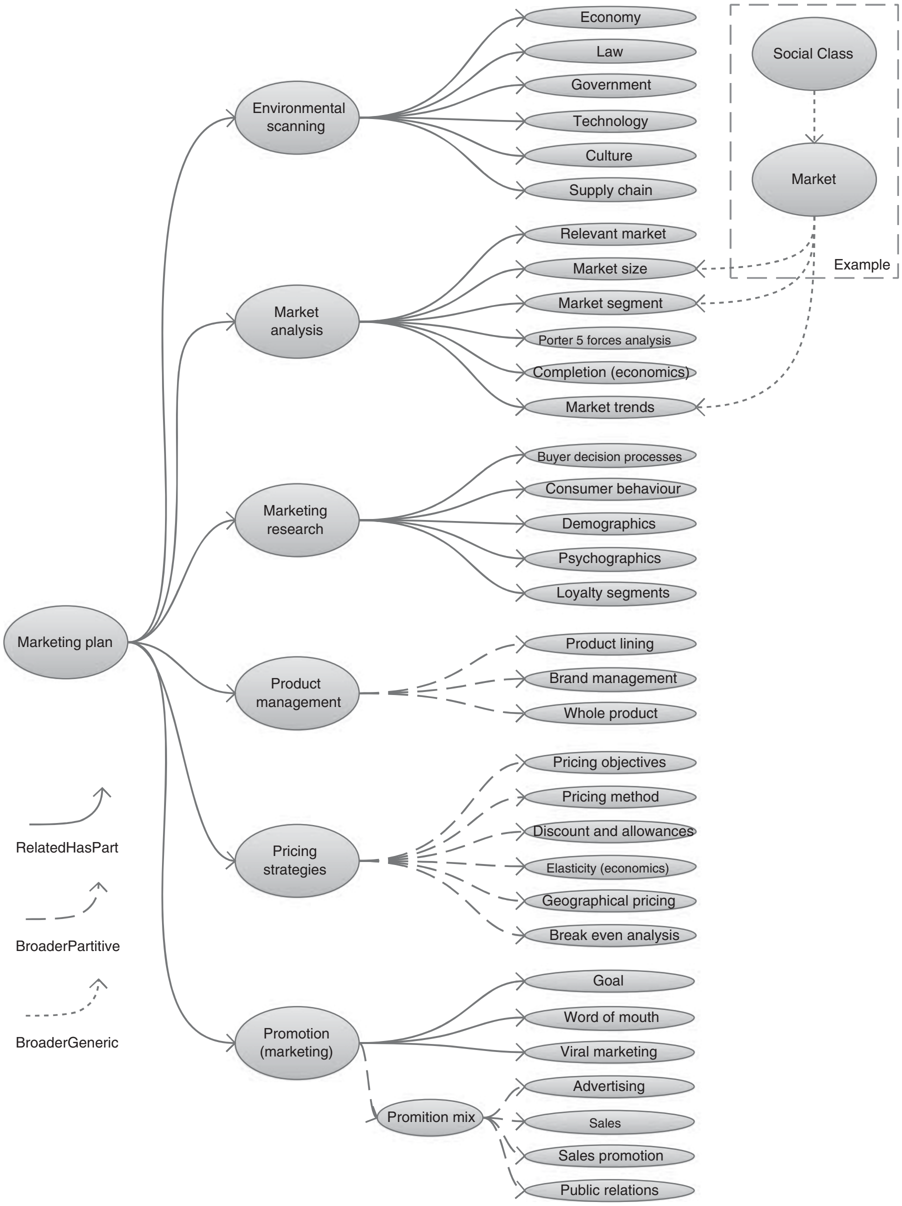

Figure 5 ‘Marketing plan’ ontology.

Table 2 Evaluation of the proposed approach to conceptual coverage (CC)

	CC
Gold standards	
ResearchCyc	0.33
WordNet	0.41
Proposed approach	
Wikipedia	0.94

As Table 2 shows, the Wikipedia semiautomatic ontology learning approach covers 94% of expected concepts in our test case domain.

During the evaluation, some trivial concepts, such as *Porter 5 forces analysis*, were successfully discovered and extracted from Wikipeida. However, disappointingly, of many omitted concepts, *marketing plan* and *supply chain* could not be located in either ResearchCyc or WordNet.

As demonstrated in this example, the proposed approach allowed to reduce the time and effort in creating an ontology of a particular domain capitalising on the existing publicly available collaborative database documented in Wikipedia. It only requires a limited involvement of the domain expert to confirm the quality of the learnt relationships and concepts in a semiautomated manner.

Conclusions and future research

Collaborative knowledge creation efforts are always at least implicitly focused on generating a common ontology. With the wealth of new collaborative technologies and platforms, there should be some better ways to support collaborative efforts and then use the resulting documentation as the basis for the (semi)automated creation of an ontology, which would then have the desired collaborative characteristics. Using an existing and dynamic collaborative corpus created by enthusiasts of Web 2.0 technologies is still a relatively unexplored area. The advantage of such an approach is a seamless and relatively effortless process for which an ontology is the result of ongoing collaboration and knowledge management without an *a priori* goal of ontology creation.

In this paper, we demonstrate an innovative approach to managing collaborative knowledge. We have taken a step forward in bridging the gap between codification and collaboration by utilising Wikipedia as a repository of dynamic and timely knowledge maintained by a variety of communities and social mechanisms. This emphasis on the collective construction of knowledge ensures that the ontology created from such a repository provides a voice for the community in the BoK and processes around that BoK. The ontology is then an artefact that itself becomes part of the discourse in the community around the BoK and plays a significant role in the dynamic evolution of the BoK and the ontology (Burstein *et al.*, 2006). Our approach is significant in that it assumes collaboration in knowledge creation, and provides a voice to the community in knowledge codification and encourages the dynamic evolution of the BoK by automating some of the codification processes. The automation is itself significant from a community perspective. Actors are removed from the time consuming and onerous task of ontology construction, while at the same time they are given the authority to make judgements about the generated ontology. Thus, the ontology not only reflects the collective knowledge of the community of interest but also emphasises the collective nature of the codification process.

We have proposed and illustrated an application of a semiautomatic approach to collaborative ontology learning that shows promising results when compared to two Gold standard hand-crafted ontologies with over 90% CC reached in 1-h effort by a non-expert. Our emerging ability to incorporate such knowledge in ontologies as the basis for knowledge management tools will result in richer, more precise, and more relevant knowledge codification, in an ever-changing world in which access to social knowledge plays an increasingly important role. As we advance testing of the ontology learning component, we expect that the impact on a broader engineering methodology will be substantial, and yet, much more work is needed in this area. Using additional meta-knowledge characteristics of the collaborative corpus as provided by the Wikipedia, API also opens up a number of interesting directions as mentioned above.

We plan to extend our work to some different problem domains to help uncover unique ontology learning requirements and lead to refinement of the approach. We have begun testing the approach by isolating subareas of the Venture Capital Investment domain and generating automated ontologies for each subareas. Initial expert-assessed results have shown broad coverage of domain concepts, but some additional testing is still needed to confirm the approach.

There is also additional work to be done in the usage of the Wikipedia API. The automated analysis we described used a rather small proportion of data provided by the API. Future work will investigate other information available from Wikipedia API, e.g., the page view counter and editing counter that can be used in weighing the importance of an article. In addition, language links can help to develop a multilingual ontology (Lauser *et al.*, 2002; De Bo *et al.*, 2003).

References

Adamides E and Karacapilidis N (2006) Information technology support for the knowledge and social processes of innovation management. *Technovation* **26(1)**, 50–59.

Balconi M (2002) Tacitness, codification of technological knowledge and the organization of industry. *Research Policy* **31(3)**, 357–379.

Balconi M, Pozzali A and Viale R (2007) The ‘codification debate’ revisited: a conceptual framework to analyze the role of tacit knowledge in economics. *Industrial and Corporate Change* **16(5)**, 823–849.

Bertino E, Catania B and Zarri GP (2001) *Intelligent Database Systems*. Addison-Wesley Longman Publishing Co., Inc. Boston, MA.

Bowker G and Star L (1999) *Sorting Things Out: Classification and Its Consequences*. MIT Press, Cambridge, MA.

Buchholz W (2006) Ontology. In *Encyclopaedia of Knowledge Management* (Schwartz DG Ed), pp 694–702, IGI Reference, Idea Group Inc., Hershey, PA.

Buitelaar P, Cimiano P and Magnini B (2005) *Ontology Learning from Text: Methods, Evaluation and Applications*. IOS Press, Amsterdam.

Burstein F, Mckemmish SM, Fisher JL, Manaszewicz R and Malhotra P (2006) A role for information portals as intelligent decision support systems: Breast Cancer Knowledge Online experience. In *Intelligent Decision-making Support Systems: Foundations, Applications and Challenges* (Gupta JND, Forgionne GA and Mora M, Eds), pp 359–383, Springer-Verlag, London, UK.

Cimiano P (2006) *Ontology Learning and Population from Text: Algorithms, Evaluation and Applications*. Springer-Verlag New York Inc., Secaucus, NJ.

Cimiano P, Handschuh S and Staab S (2004) Towards the self-annotating web. In *Proceedings of the 13th International Conference on World Wide Web*, May 17–20, pp 462–471, ACM, New York, NY.

Cross R, Parker A, Prusak L and Borgatti S (2001) Knowing what we know: supporting knowledge creation and sharing in social networks. *Organ Dynamics* **3(2)**, 100–120.

De Bo J, Spyns P and Meersman R (2003) Creating a ‘dogmatic’ multilingual ontology infrastructure to support a semantic portal, in on the move to meaningful Internet systems 2003: OTM 2003 workshops. *Lecture Notes in Computer Science* **2889**, 253–266.

Etzioni O, Cafarella M, Downey D, Kok S, Popescu AM, Shaked T, Soderland S, Weld DS and Yates A (2004) Web-scale information extraction in know it all: (preliminary results). In *Proceedings of the 13th international conference on World Wide Web*, May 17–20, pp 100–110, ACM, New York, NY.

Farquhar A, Fikes R and Rice J (1997) Ontolingua server: A tool for collaborative ontology construction. *International Journal of Human–Computers Studies* **46(6)**, 707–727.

Farquhar A, Fikes R, Pratt W and Rice J (1995) Collaborative ontology constructions for information integration. Technical Report, KSL-95–63, Stanford University Knowledge Systems Laboratory, Stanford University, Palo Alto, CA.

Fellbaum C (1998) *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

Ferneley E, Berney B and Rezgui Y (2002) Information retrieval algorithms for knowledge management – the challenge continues. In: *Proceedings of the European Conference on Information and Communciation Technology Advances and Innovation in the Knowledge Society*, eSMART 2002 in collaboration with CISEMIC 2002 Conference, Salford, Vol. 1, pp. 168-177.

Giles J (2005) Special report: Internet encyclopedias go head to head. *Nature* **438(15)**, 900–901.

Gillmor D (2004) We the Media. Sebastopol, CA: O'Reilly Media http://www.authorama.com/book/we-the-media.html.

Glaser M (2006) Your guide to citizen journalism. Public Broadcasting Service http://www.pbs.org/mediashift/2006/09/your-guide-to-citizen-journalism270.html.

Gómez-Pérez A, Fernández-López M and Corcho O (2004) *Ontological Engineering: With Examples from the Areas of Knowledge Management, E-Commerce and the Semantic Web*. Springer, London, UK.

Guarino N and Welty C (2000) A formal ontology of properties. In *Proceedings of EKAW-2000: The 12th International Conference on Knowledge Engineering and Knowledge Management*, Vol. 1937, pp 97–112, Springer-Verlag, London, UK.

Hearst MA (1992) Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th Conference on Computational Linguistics* Vol. 2, Nantes, France, pp 539–545, Association for Computational Linguistics, Morriston NJ.

Holsapple CW and Joshi KD (2002) A collaborative approach to ontology design. *Communications of the ACM* **45(2)**, 42–47.

Jarrar M and Meersman R (2008) Ontology Engineering – The DOGMA approach Lecture Notes In Computer Science archive. *Advances in Web Semantics I: Ontologies, Web Services and Applied Semantic Web Section: Part I Ontologies and Knowledge Sharing*, pp 7–34, Springer-Verlag; Berlin, Heidelberg.

Jarrar M, Verlinden R and Meersman R (2003) Ontology-based customer complaint management. In *Proceedings of the Workshop on Regulatory Ontologies and the Modeling of Complaint Regulations* , LNCS, 2889, pp 594–606.

Johnson B, Edward LE and Lundvall B-Å (2002) Why all this fuss about codified and tacit knowledge? *Industrial and Corporate Change* **11(2)**, 245–262.

Latour B (1987) *Science in Action: How to Follow Scientists and Engineers through Society*. Open University Press, Milton Keynes, UK.

Lauser B, Wildermann T, Poulos A, Fisseha F, Keizer J and Katz S (2002) A comprehensive framework for building multilingual domain ontologies: Creating a prototype biosecurity ontology. International Conference on Dublin Core and Metadata Application Archive. In *Proceedings of the International Conference on Dublin Core and Metadata for e-Communities: Supporting diversity and convergence table of contents*, pp 113–123, Dublin Core Metadata Initiative, Florence, Italy.

Lave J and Wenger E (1991) *Situated Learning: Legitimate Peripheral Participation*. Cambridge University Press, Cambridge.

Lenat DB and Guha RV (1990) *Building Large Knowledge-Based Systems: Representation and Inference in the Cyc Project*. Addison-Wesley, Longman Publishing Co., Inc. Boston, MA.

Markert K, Nissim MK and Modjeska NN (2003) Using the web for nominal anaphora resolution. *Proceedings of the European Chapter of the ACL (EACL) Workshop on the Computational Treatment of Anaphora* (Dale R, Van deemter K and Mitkov R, Eds), April 12–17 Budapest, Hungary pp 39–46.

Mathes A (2004) Folksonomies – cooperative classification and communication through shared metadata. *Computer Mediated Communication (LIS590CMC)*. University of Illinois, Urbana-Champaign, Illinois.

Mika P (2005) Ontologies are us: a unified model of social networks and semantics. In *Proceedings of the International Semantic Web Conference 2005 (ISWC 2005) Lecture Notes in Computer Science (LNCS) 3729*, pp 522–536, Springer-Verlag, Galway, Ireland.

Niles I and Pease A (2001) Origins of the IEEE standard upper ontology. *Working Notes of the IJCAI-2001 Workshop on the IEEE Standard Upper Ontology*, pp 37–42, Seattle, WA.

Pinto HS and Martins JP (2004) Ontologies: how can they be built? *Knowledge and Information Systems* **6(4)**, 441–464.

Pinto HS, Staab S and Tempich C (2004) DILIGENT: towards a fine-grained methodology for distributed, loosely-controlled and evolving engineering of ontologies. *Proceedings of the 16th European Conference on Artificial Intelligence (ECAI)* In (de Mantras RL and Saitta L, Eds), pp 393–397, IOS Press, Valencia, Spain.

Ponzetto SP and Strube M (2007) Deriving a large scale taxonomy from Wikipedia. In *Proceedings of the 22nd National Conference on Artificial Intelligence* pp 1440–1445, Vancouver, Canada.

Ratsch E, Schultz J, Saric J, Lavin PC, Wittig U, Reyle U and Rojas I (2003) Developing a protein interactions ontology. *Comparative and Functional Genomics* **4(1)**, 85–89.

Reinberger ML and Spyns P (2005) Unsupervised text mining for the learning of DOGMA-inspired ontologies. *Ontology Learning from Text: Methods, Evaluation and Applications and Evaluation.* In (Buitelaar P, Cimiano P and Magnini B, Eds), pp. 29–43, IOS Press, Amsterdam.

Sabou M (2005) Learning Web service ontologies: An automatic extraction method and its evaluation. In *Ontology Learning from Text: Methods, Evaluation and Applications Frontiers in Artificial Intelligence and Application Series* (Buitelaar P, Cimiano P and Magnini B, Eds), pp 125–139, Vol. 123, IOS Press, Amsterdam.

Sabou M (2006) Building Web service ontologies. p 187, PhD thesis, SIKS Dissertation Series, UK.

Singh P, Lin T, Mueller E, Lim G, Perkins T and Zhu W (2002) Open mind common sense: knowledge acquisition from the general public. In *Proceedings of the First International Conference on Ontologies, Databases, and Applications of Semantics for Large Scale Information Systems*, LNCS 2519, pp 1223–1237, Springer-Verlag, London, UK.

Snow R, Jurafsky D and Ng AY (2006) Semantic taxonomy induction from heterogenous evidence. In *ACL '06: Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the ACL*, pp 801–808, Association for Computational Linguistics, Morristown, NJ.

Star L (Ed) (1995) *Ecologies of Knowledge: Work and Politics in Science and Technology*. SUNY Press, Albany, NY.

Star L and Griesemer J (1989) Institutional ecology, 'translations' and boundary objects: amateurs and professionals in Berkeley's museum of vertebrate Zoology, 1907–39. *Social Studies of Science* **19(3)**, 387–420.

Suchanek FM, Ifrim G and Weikum G (2006) 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*, Sydney, Australia, July 2006, Association for Computational Linguistics, pp 18–25.

Suchanek FM, Kasneci G and Weikum G (2007) Yago: A core of Semantic Knowledge. In *WWW'07: Proceedings of the 16th International Conference on World Wide Web*, pp. 697–706, New York, NY, ACM Press.

Sure Y (2003) Methodology, tools and case studies for ontology-based knowledge management. Unpublished Doctoral Dissertation, Karlsruhe University, Germany.

Sure Y, Erdmann M, Angele J, Staab S, Studer R and Wenke D (2002) Ontoedit: Collaborative ontology development for the semantic Web. In *Proceedings of the 1st International SemanticWeb Conference (ISWC2002), June 9–12, 2002* , LNCS 2342, pp221–235 Springer, Sardinia, Italia.

Udell J (2004) Collaborative knowledge gardening. *InfoWorld.* http://www.infoworld.com/article/04/08/20/34OPstrategic_1.html (accessed 24 June 2009).

USCHOLD M (1996) Building ontologies: towards a unified methodology. *16th Annual Technical Conference of the British Computer Society Specialist Group on Expert Systems*, pp 75–90, SGES Publications, Cambridge, UK.

VANDER WAL T (2004) Folksonomy, http://vanderwal.net/folksonomy.html (accessed 24 June 2009).

VON AHN L (2006) Games with a purpose. *Computer* **39(6)**, 92–94.

ZIRN C, NASTASE V and STRUBE M (2008) Distinguishing between instances and classes in the Wikipedia taxonomy. In *Proceedings of the 5th European Semantic Web Conference* (HAUSWIRTH, M KOUBARAKIS M and BECHHOFER S, Eds), LNCS, berlin, Heidelberg, June 2008 Springer Verlag.

About the authors

Tao Guo graduated with a research degree from the Faculty of Information Technology, Monash University, Australia, where he received his Master of Computer Science degree. He started programming when he was 12 years old as influenced by his father, who was a software engineer at the time. His research interests include ontology, semantic web, text- mining, web application architecture, and agile software development. He received his Bachelor's degree from the Department of Electronic Information and Control Engineering, Beijing University of Technology. With his great interest in computer science, after working for two years as a software developer, he left his country and began his academic journey in Australia. He has years of experience with .NET, Java and Python, and currently is looking for a research engineer position.

David Schwartz is a senior lecturer at the Graduate School of Business Administration at Bar-Ilan University, Israel. Since 1998 he has been serving as Editor-in-Chief of the journal *Internet Research*. Dr. Schwartz's research has appeared in publications such as *IEEE Intelligent Systems, International Journal of Human-Computer Studies, IEEE Transactions on Professional Communications, Kybernetes*, and the *Journal of Organizational Behavior*. His books include *Cooperating Heterogeneous Systems, Internet-Based Knowledge Management and Organizational Memory*, and the *Encyclopedia of Knowledge Management*. He has been a visiting scholar at Columbia University, Department of Biomedical Informatics and Monash University, Faculty of Information Technology. His main research interests are knowledge management, ontology, internet-based systems, and computer-mediate communications. Dr. Schwartz received his Ph.D. from Case Western Reserve University; MBA from McMaster University; and B.Sc. from the University of Toronto, Canada.

Frada Burstein is an experienced researcher and educator in the area of decision support, knowledge management and systems with more than 20 years of professional experience in the area both nationally and internationally. She gained her Ph.D. in Decision Support Systems (DSS) from the Soviet Academy of Sciences. Professor Burstein has published extensively in academic journals and collections of papers. She has also been a member of the scientific, program and organizing committee or chaired many international workshops and conferences on DSS and knowledge management. She is an executive committee member for the Australian Council of Professors & Heads of Information Systems, advisory board member for the Association of Information Systems Special Interest Group in DSS and a secretary for the IFIP WG 8.3 on Decision Support and Knowledge Management Systems. Professor Burstein is cunently Associate Dean Research Training for the Faculty of Information Technology, Monash University.

Henry Linger is the deputy director of the Knowledge Management Research Program at Monash University and a senior lecturer in the Faculty of Information Technology. He has been a research sssociate at Defence Science and Technology Organisation for the past 10 years. Henry conducts research in the area of knowledge work, knowledge management, organisational learning and the design of systems to support professional work. His research involves national and international collaborations addressing a broad range of domains including biology, immunology, epidemiology, food safety, meteorology, defence and clinical and management aspects of healthcare.