Ontology Engineering for ICT Systems using Semantic Relationship Mining and Statistical Social Network Analysis

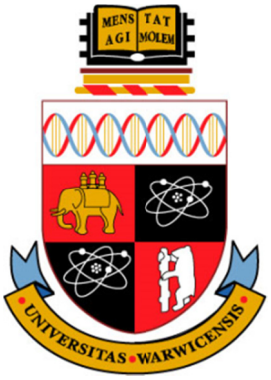
by

Xiao MA

A thesis submitted in partial fulfillment of the requirements for the

Degree of

Doctor of Philosophy in Engineering



Warwick Manufacturing Group, University of Warwick

March 2011

TABLE OF CONTENT

ACKNOWLEDGEMENTS

I would like to express my thanks to the Engineering and Physical Sciences Research Council (EPSRC) for providing the funding via the Warwick Innovative Manufacturing Research Centre (WIMRC) to make this research possible, and the Taj Medical Group for kindly partially sponsoring me to conduct this research.

I would like to thank all those who have supported me in their unique ways during this doctoral research:

1. My academic supervisor Dr. Jay Bal, who gives me endless patience, guidance and advice on the research, and shares his wisdom and experience on research life.
2. My industry supervisor Dr. Jagdish Jethwa for his guidance, advice and friendship.
3. The West Midlands Collaborative Commerce Market Place and the Taj Medical Group for generously offering their operational details and ontology related information.
4. Dr. Joseph Darlington and Mr. Ahmad Issa for their support on the research and thesis proofreading.
5. My parents CuiZhen Meng and YongSheng Ma for their support and encouragement during all these years that I have been away from them.
6. My wife Yu Yang for her endless support, encouragement during my doctoral research.
7. Dr. Mark Swift, Dr. Nikolas Armoutis, Rob McGongle, Tuhin Ahmed, Marios Mofatas, Xin Wang and Honghong Dai for individual support, academic advice and their great companionship.
8. Many others of my family, friends and staff at the University of Warwick whose names are deeply engraved in my memory.

DECLARATION

I, Xiao MA declare that except where acknowledged, the material contained in this thesis is my own work and that is has neither been previously published nor submitted elsewhere for the purpose of obtaining an academic degree.

FORWARD

This work was funded by WIMRC to investigate self-help systems for improving medical processes. Early work with the Taj Medical Group to investigate free(er) market medical processes in healthcare tourism identified good ontology as key to self-help systems. Generation of such ontology thus become the key research here.

ABSTRACT

In information science, ontology is a formal representation of knowledge as a set of concepts within a domain, and the relationships between those concepts. It is used to reason about the entities within that domain, and may be used to describe the domain. (Wikipedia, 2011)

This research takes two case study ICT applications in engineering and medicine, and evaluates the applications and supporting ontology to identify the main requirements for ontology in ICT systems. A study of existing ontology engineering methodology revealed difficulties in generating sufficient breadth and depth in domain concepts that contain rich internal relationships. These restrictions usually arise because of a heavy dependence on human experts in these methodologies.

This research has developed a novel ontology engineering methodology – SEA, which economically, quickly and reliably generates ontology for domains that can provide the breadth and depth of coverage required for automated ICT systems. Normally SEA only requires three pairs of keywords from a domain expert. Through an automated snowballing mechanism that retrieves semantically related terms from the Internet, ontology can be generated relatively quickly. This mechanism also enhances and enriches the binary relationships in the generated ontology to form a network structure, rather than a traditional hierarchy structure. The network structure can then be analysed through a series of statistical network analysis methods. These enable concept investigation to be undertaken from multiple perspectives, with fuzzy matching and enhanced reasoning through directional weight-specified relationships.

The SEA methodology was used to derive medical and engineering ontology for two existing ICT applications. The derived ontology was quicker to generate, relied less on expert contribution, and provided richer internal relationships. The methodology potentially has the flexibility and utility to be of benefit in a wide range of applications.

SEA also exhibits “reliability” and “generalisability” as an ontology engineering methodology. It appears to have application potential in areas such as machine translation, semantic tagging and knowledge discovery. Future work needs to confirm its potential for generating ontology in other domains, and to assess its operation in semantic tagging and knowledge discovery.

ABBREVIATIONS

CAP: College of American Pathologists

CTV3: Clinical Terms Version 3

DOGMA: Developing Ontology-Grounded Methods and Applications

EDI: Electronic Data Interchange

GB: GigaByte

HBS: Harvard Business School

HTML: Hyper Text Mark-up Language

ICT: Information Communication Technology

IHTSDO: International Health Terminology Standards Development Organisation

MAS: Manufacturing Advisory Service

MeSH: Medical Subject Headings

NACE: Statistical Classification of Economic Activities in the European Community

NHS: National Health Service of the United Kingdom

NPD: New Product Development

OWL: Web Ontology Language

RDF: Resource Description Framework

SEA: Semantic (S) relatedness oriented ontology engineering via retrieving information from the search Engine (E) index with assistance from social network analysis (A). It is the new ontology engineering methodology developed by this research.

SIC: Standard Industry Code (UK)

SLA: Service Level Agreement

SNA: Social Network Analysis

SNOMED CT: Systematized Nomenclature of Medicine-Clinical Terms

SNOP: Systematized Nomenclature of Pathology

TB: TeraByte. 1 TB = 1,024 GB

TMG: the Taj Medical Group

UMLS: Unified Medical Language System

UN: the United Nations

UNSPSC: the United Nations Standard Products and Services Code

W3C: the World Wide Web Consortium

WMCCM: West Midlands Collaborative Commerce Marketplace

WWW: World Wide Web

XML: Extensible Mark-up Language

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1: INTRODUCTION

Ontology describe the common entities and the relationships associated with a knowledge domain or a task. They are thus important when discussing concepts, activities and collaborations pertaining to that domain. ICT systems designed to handle processes and tasks can benefit hugely from suitable ontology that aids structure and decision making. ICT systems that lack suitable ontology behind them can suffer from issues such as an over reliance on domain experts, a large number of data gathering and refining steps, poor ability to infer new information and poor ability to address multi-disciplinary issues. This research explores how ontology for ICT systems can be created that are quick to generate, reliable in interpretation and require less input from domain experts. The exploration is conducted through a detailed analysis of two case study applications and their problems followed by a discussion as to the general applicability of the methodology derived.

1.1 RESEARCH MOTIVATION

The Semantic Web is a group of methods and technologies to allow machines to understand the meaning - or "semantics" - of information on the World Wide Web. Ontology are the “engine” of semantic web because they aid in the “understand the meaning” process by defining the relationships between different entities – an activity commonly called knowledge representation (Seidenberg and Rector, 2006).

A family of knowledge representation language for authoring ontology is established as the Web Ontology Language (OWL) (Dean and Schreiber, 2003). These languages are characterised by formal semantics and Resource Description Framework (RDF) / Extensible Markup Language (XML) - based serializations for the Semantic Web. OWL is endorsed by the World Wide Web Consortium (W3C) and has attracted academic, medical and commercial interest.

The ability to semantically model knowledge has turned ontology into useful tools for business use. The Yahoo index uses ontology to categorise web sites into a large taxonomy for web search purposes, while Amazon has built its product catalogue according to ontology on products and their features. Many other disciplines have recognised the value of ontology. For example, in healthcare, the Systematised Nomenclature of Medicine Clinical Terms (SNOMED CT) has been an attempt to standardise a "systematically organized computer processable collection of medical terminology" for use across health information systems, in order to solve the problem of poor communication between healthcare practitioners and patients (IHTSDO, 2009).

> *SNOMED Clinical Terms (SNOMED CT) is a comprehensive clinical terminology that provides clinical content and expressivity for clinical documentation and reporting. It can be used to code, retrieve, and analyze clinical data. SNOMED CT resulted from the merger of SNOMED Reference Terminology (SNOMED RT) developed by the College of American Pathologists (CAP) and Clinical Terms Version 3 (CTV3) developed by the National Health*

Service (NHS) of the United Kingdom. The terminology is comprised of concepts, terms and relationships with the objective of precisely representing clinical information across the scope of health care.

SNOMED CT provides the core general terminology for the electronic health record (EHR) and contains more than 311,000 active concepts with unique meanings and formal logic-based definitions organized into hierarchies. SNOMED CT is considered to be the most comprehensive, multilingual clinical healthcare terminology in the world.

Today, SNOMED CT is available in US English, UK English, Spanish and Danish. Translations into French, Swedish, Lithuanian, and several other languages are currently taking place. (IHTSDO, 2009)

SNOMED CT demonstrated that ontology can be built in multiple languages. Multilingual ontology may not be a simple translation of ontology from one language to others: as stated by SNOMED CT, "term-to-term translations may yield literal expressions that are often meaningless"(IHTSDO, 2009). Therefore in practice, ontology should firstly require a specific understanding of lexicons and semantics around terms in any individual language in order to be multilingual (Espinoza et al., 2008). For example, when Google entered the Indonesia market, its customised ads which match against web content failed to compete with similar ads from a local company "Sitii", because Google's matching mechanism was based on keyword relevant ontology in English, while Sitti collected terms and their relationships based on the Indonesian local language (Lacy, 2010). This shows that the same knowledge may require different ontological models from different ontology

building perspectives. For instance, a domain ontology may require domain specified terms and relationships, or an application dependant ontology should define application specified terms and relationships.

Ontology such as SNOMED CT are normally built to the classic ontology definition – "explicit formal specifications of the terms in the domain and relations among them" (Gruber, 1993). The ontology derived thus often lacks the broad coverage required to bridge the terminology gap between professionals and general users. There are parallels with expert systems that relentlessly focus on specific domains when a simpler answer exists from another domain. Secondly, the resources required to build such ontology (time and human expertise) are normally only affordable by large organisations. Thirdly, "explicit and formal" denies the fuzzy understanding that can exist between experts in different domains, and even within the same domain. A requirement that the whole area of Fuzzy Systems was developed to address. The need to address fuzziness at the edges of a domain becomes very important when trying to work across multiple disciplines. Systems that can deal with multi-domain knowledge will thus have a fuzzy area that links the different domains.

Figure 1.1 shows examples of the key joining terms between different concepts or domains: in this case "plastic surgery" is the fuzzy area that connects cosmetic surgeries (highlighted by dashed blue line) to the other medical practices. There is an

increasing need for multi-disciplinary systems, and it appears that existing research and approaches to generating the ontology required by such systems have not fully addressed the issues.

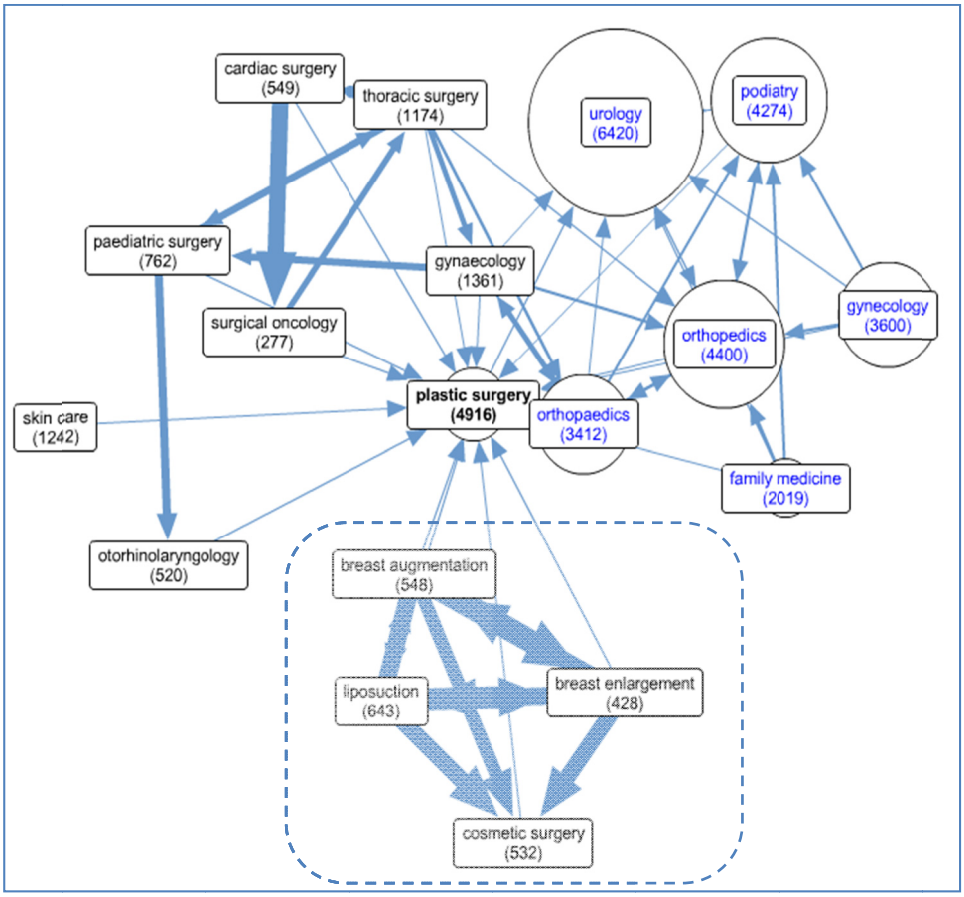


Figure 1.1: Example of fuzzy area in ontology

For example, the United Nations (UN) has provided ontology to support a "multi-sector standard for efficient, accurate classification of products and services" (Granada_Research, 2001). However when it was implemented in practice, regional committees found it hard to re-use the UN version, so they proposed different ontology. Sometimes, individual countries proposed their own ontology: for example

"The North American Industry Classification System" or the "Standard Industrial Classification (UK)" to ensure clarity due to cultural or language issues, and a lack of sufficient detail in the original definitions.

It appears that current methods of ontology engineering do not generate multi domain ontology that meets real needs. Swartout and colleagues conclude that researchers in this area have not been successful in building up-to-date ontology, reducing domain expert involvement and providing the breadth of coverage to enable multi-disciplinary applications (Swartout et al., 1997). Ontology are often a key tool in the analysis and translation of specialist language, and are essential in building systems for multi-disciplinary working, or ones that allow non-specialists to access them. They are widely used in artificial intelligence, the Semantic Web, systems engineering, software engineering, biomedical informatics, library science, enterprise information systems and information architecture, as a form of knowledge representation about the world or some part of it. Examples of work in this area are:

- In water management, traditional system modelling initiatives follow textbook approaches in a single domain. A lack of mutual understanding among modelling teams undermines the reproducibility of modelling work. Multi-disciplinary ontology was introduced to link the generic part of modelling with various water management domains for different types of users and different levels of modelling complexity. (Scholten et al., 2007)

- In collaborative engineering design, ontology-based modelling via the Semantic Web is proposed, to enable communication and knowledge reuse among multi-disciplinary organisations. Original domain specified (engineering) knowledge representation was reconfigured as a new multi-agent distributed design ontology, and this ontology facilitates semantic access and information retrieval across different disciplines. (Zhang and Yin, 2008)

The Web 2.0 has also created a massive increase in user generated content. Davies and colleagues have pointed out that people need techniques to enable them to make sense of the huge volumes of information with speed and reliability (Davies et al., 2006). These technique for structured information representation in a given application environment assist the user in the sense making process. To be specific, people require the ability to select relevant content, and to filter out any irrelevances, as well as to summarise and extract the essence of numerous articles and to clarify the relationships between such pieces of information. This requires the "Semantic Web" to provide a framework that enables data sharing and re-use across applications, and to describe (or tag), manage and process information semantically (Berners-Lee, 2000). Using the output from the semantic web, systems can "understand" the meaning (semantics) of natural language, so as to assist the decision making processes. At the heart of the Semantic Web is the use of "Ontology" that establish human oriented terminologies, summarise concepts and reason from their relationships (Davies et al., 2003).

This research seeks methods to derive multi-disciplinary ontology for any sector or specialism quickly, reliably and economically compared with current methods. The methods developed are derived from an analysis of the knowledge processing needs for two sectors, medical tourism from The Taj Medical Group, and virtual organisation formation (opportunity finding and partner selection) in the engineering sector from the West Midlands Collaborative Commerce Marketplace.

- The Taj Medical Group is a medical tourism agency that helps organise travel and treatment options for patients. They partner with a range of hospitals, specialist and clinics. At the time of the research the Taj Medical Group was the largest provider of medical tourism to India. A problem they had was that because of the amount of time and resources required to match each individual patient with the best provider, the business was not very profitable. They required help in reducing the experts' workload in matching patient needs with providers' capabilities.
- The West Midlands Collaborative Commerce Marketplace (WMCCM) is an ontology driven marketplace for engineering businesses. The system automatically matches tender opportunities with company competency, so as to provide focused opportunities. It does this through a semantic analysis of tender content to classify tenders against a general engineering ontology that was custom created for the system. Company competencies for the 400+ profiled companies are also classified against the same ontology. Thus, the capability

needs of a tender can be matched against the company(s) that can provide them, and an appropriate partnership suggested. The WMCCM ontology was built partly top down and partly bottom up by domain experts.

These case studies are described in more detail in chapter two.

1.2 PROBLEM DESCRIPTION

The creation or identification of a suitable ontology for any sector or field is thus important in aiding communication, enhancing collaboration and automating processes. Ontology engineering, as a subject, is an existing area of research in knowledge management. However, creating or arriving at a suitable ontology (ontology engineering) for an industry remains a major problem, and an increasingly important area of research.

For instance, SNOMED CT still suffers many drawbacks, such as unclear identification from different healthcare aspects, top level ontology alignment with other widely accepted ontology, and ambiguous relationships towards non-expert terminology, etc. (Schulz et al., 2009)

In taking the West Midlands Collaborative Commerce Marketplace (WMCCM) as an example, first hand experience with WMCCM has revealed insights on certain problems. WMCCM matches tender opportunities with company competency, by categorising tender (Figure 1.2) and company information (Figure 1.3) against its

ontology covering the engineering sector and associated activities. Poor categorisation results in wasted tender scrutiny and poor partnership possibility suggestions.

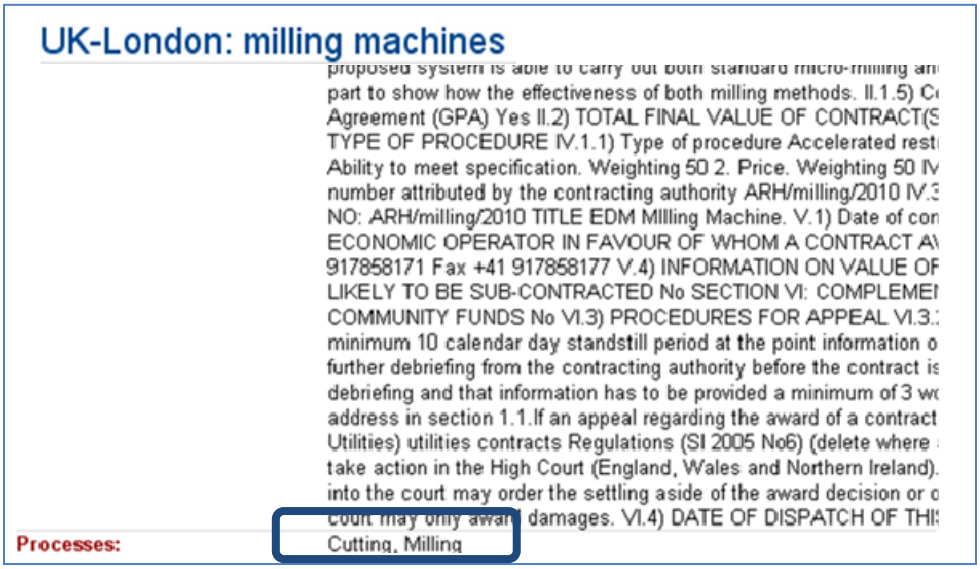


Figure 1.2: Example of tender matching in WMCCM system via processes

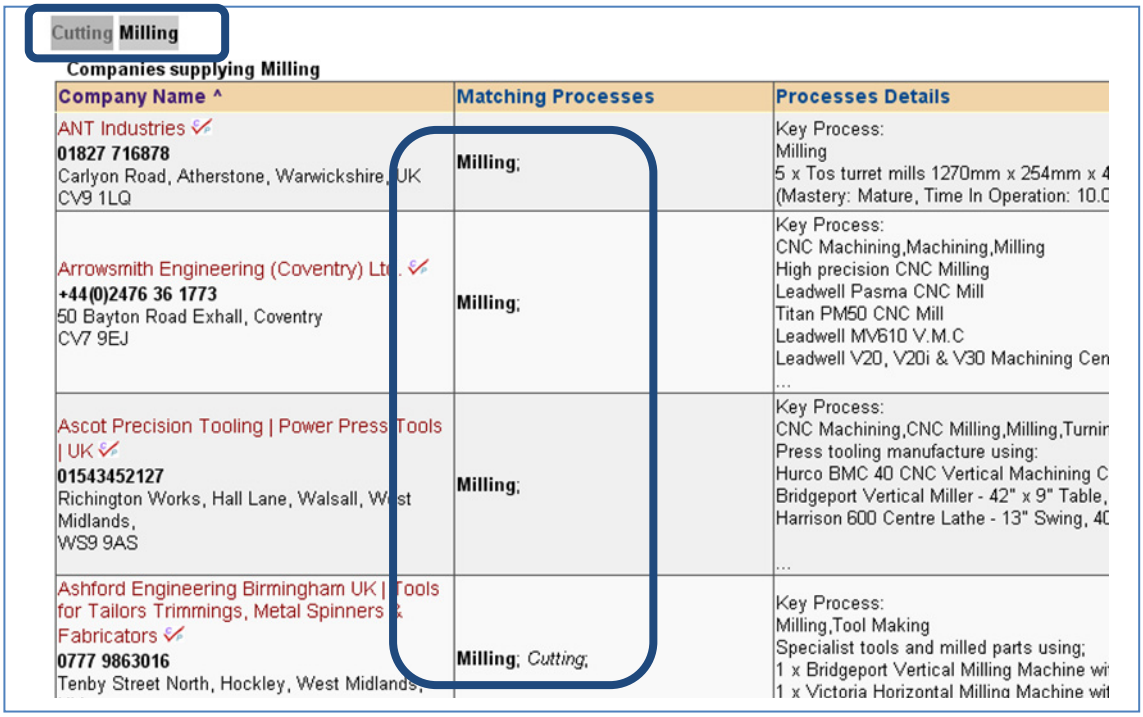


Figure 1.3: Example of company matching in WMCCM system via processes

A key factor in this and other ontology matching within WMCCM is the quality (coverage of the domain and internal relationship) of the ontology. For WMCCM, its ontology was generated experimentally, following a mixture of top down derivation and bottom up synthesis. WMCCM expert staff collected first hand data from professionals – namely academic and industry experts. The data was the words that companies actually used when discussing their capabilities, and unified these terms with those derived from standards (sourced from books or government classifications). Figure 1.4 shows part of the Ontology structure adopted by WMCCM.

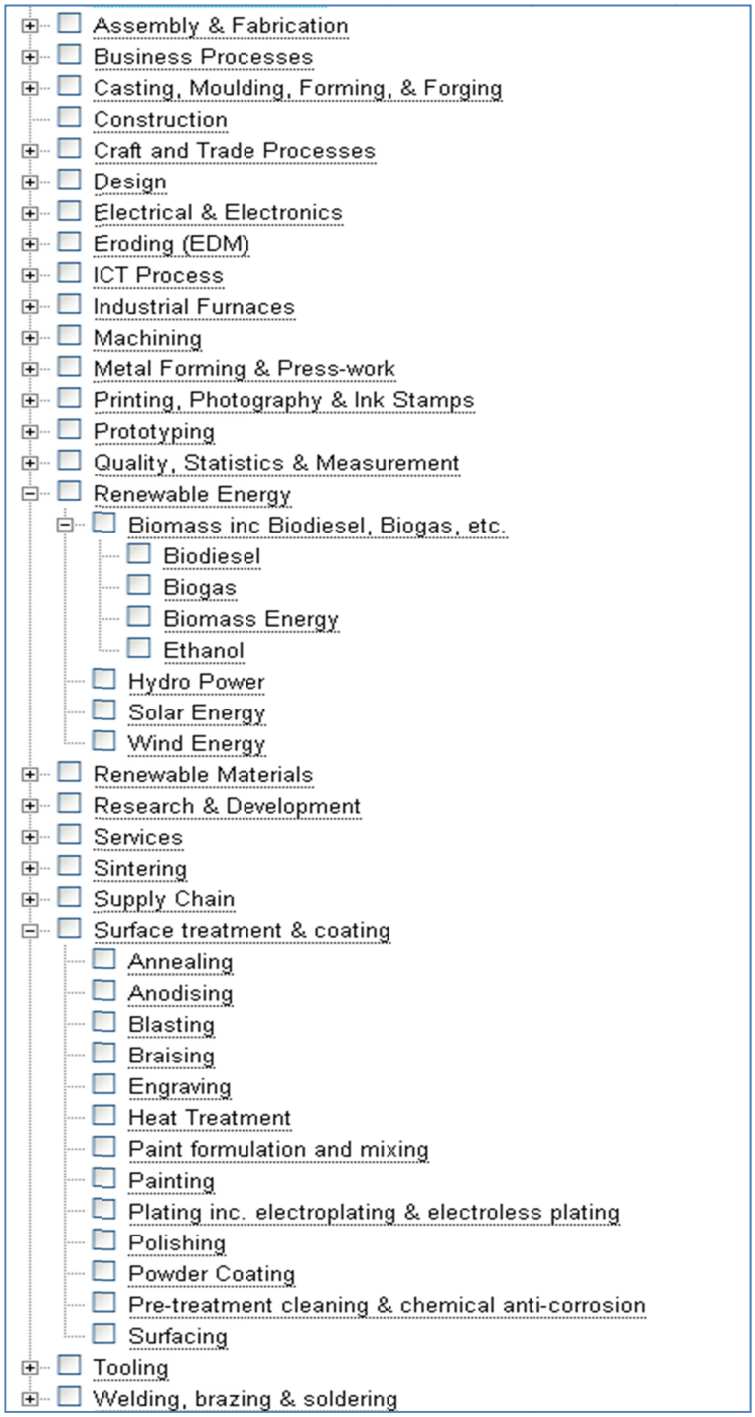


Figure 1.4: WMCCM Ontology[1]

[1] It is a three level structure, only the renewable energy and surface treatment sections are expanded

Currently, there exists a range of ontology for the engineering sector. In the UK, the SIC (Standard Industrial classification) has been available for many years and is regularly updated. The current version dates from January 2008, and has been aligned with NACE (Statistical Classification of Economic Activities in the European Community) (National Statistics, 2010). At the same time, a higher level ontology also facilitates the engineering sector's trading activities – the United Nations Standard Products and Services Code (UNSPSC).

However, there are problems in using the SIC type of ontology in practice since they rarely meet the full needs of the application environment and revisions arise so as to reflect changes in the world economy, industry and business change, and so, logically, the need to categorize its activities. The biggest changes in recent years are mainly due to the growth and development of new services in the area of ICT (Information and Communications Technology). Thus, the coverage of fast changing sectors such as Engineering or Medicine can be patchy, and there can be many gaps in the newer technologies and processes. This, in automated systems, can lead to poor categorisation of information and an inadequate level of performance from the "smart" systems using the ontology.

Secondly, existing ontology are structured according to their application environments and it is not easy to modify them to suit other applications. Therefore,

it is not a wise option to select and modify an existing ontology. Thus efficiency improvement in the application of ontology through reuse is not a sensible option.

SNOMED CT has included different collections from a variety of clinical information. In order to make use of such information, current health information systems link such collections to clinical knowledge bases for information retrieval, exchange, etc. However, a lack of satisfactory foundation ontology would not only reduce the effectiveness of these functions, but could be a threat to patient safety if the exchange of clinical information is not aligned.

In a similar way, Electronic Data Interchange (EDI) standards have developed many variants for different sectors as a result of terminology differences in different industries (Sommer, 2003). However, many of these ontology variations do not work with each other, as they are developed from different viewpoints and thus not aligned with each other. With the increasing level of multi-disciplinary and inter sectional working, it is desirable to merge existing sectorial or discipline based ontology to enable cross disciplinary collaboration. Noy and Musen identify this as an important research interest: to incorporate ontology, so as to reuse information from each of them (Noy and Musen, 1999). Thus more research is required on methods for building new cross disciplinary ontology that incorporate information from disparate discipline or sector ontology

It is also apparent that researchers in this area have not been successful in building ontology that (Swartout et al., 1997):

- is based on up-to-date information;
- reduces the demands on the domain experts;
- provides enough coverage to enable multi-disciplinary usage;
- is broad enough for translation between specialist and non-specialist knowledge understanding.

This work sets out to address these issues. The approach selected (described in-depth in chapters 3 and 5) is to use the Google search engine index as a data source to directly extract ontological information. There are very few studies in ontology engineering based on a search engine index (discussed in section 3.1). The work undertaken contributes new techniques that reduce some of the problems and issues identified in the ontology engineering methodology review conducted in chapter two.

1.3 RESEARCH QUESTION AND OBJECTIVES

From the issues discussed above, the research question posed is as follow:

"Is it possible to quickly, reliably and economically generate ontology for a specific area or areas that can provide the breadth and depth of coverage required for automated systems."

The approach chosen to address this question is one of creating a methodology. To generate such a methodology, several objectives have been identified:

- To examine and discuss current ontology engineering problems in medical and engineering sector, accompanied by the first hand data collected from TMG and WMCCM.
- To devise methods to ***quickly, reliably and economically*** create ontology for these domains and evaluate them.
- To discuss how general the approaches are and to create a generalisable ontology engineering methodology

1.4 STRUCTURE OF THIS WORK

The structure adopted for this thesis is illustrated in Figure 1.5

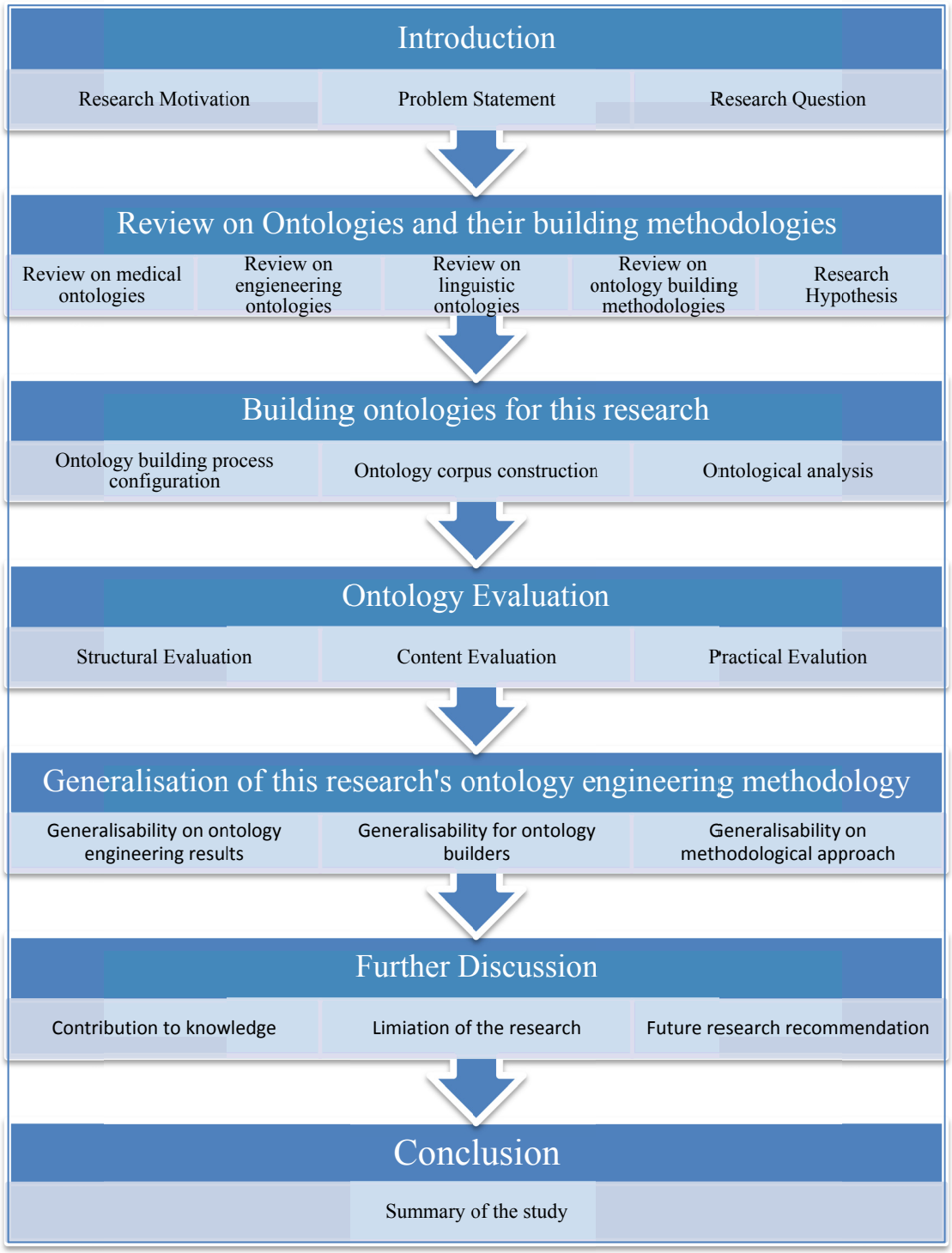


Figure 1.5: Structure of the research

CHAPTER 2: THE ONTOLOGY FIELD

The goal of this research was to explore methods to generate ontology for use in ICT systems that were quick and reliable to generate and required less of a contribution from domain experts than current approaches. The research approach chosen was to look at two problem domains, devise approaches to meet their needs and then explore the generalisability of the approaches devised. This chapter starts by describing the two problem domains and then discusses some of the current ontology engineering approaches suggested in the literature.

2.1 MEDICAL SECTOR CASE STUDY

Ontology in the medical sector are built to enable healthcare information (such as patient data, diagnosis and care regimes) share, reuse and transfer in medical information systems. They normally contain detailed medical terminology (clearly defined medical domain concepts) and related terms around them.

Healthcare ontology deal with both types of users: healthcare professionals who are expert in the field and patients who are non-specialists. Such ontology are expected to bridge the gaps between these two user groups. A case study of the Taj Medial Group (TMG) was used to explore the communication between experts and non-specialist (patients), and to help define the issues better.

2.1.1 TMG

The Taj Medical Group (TMG) is a leading medical tourism facilitator, and has arranged treatment for over eight hundred international patients from the UK and other countries. A major problem faced by TMG was the level of resources required for the process of matching their patients' enquiries with the capabilities of different medical treatment providers. The business suffered from low efficiency as a result of the resources required for each enquiry but a low conversion rate from enquiries to paying patients.

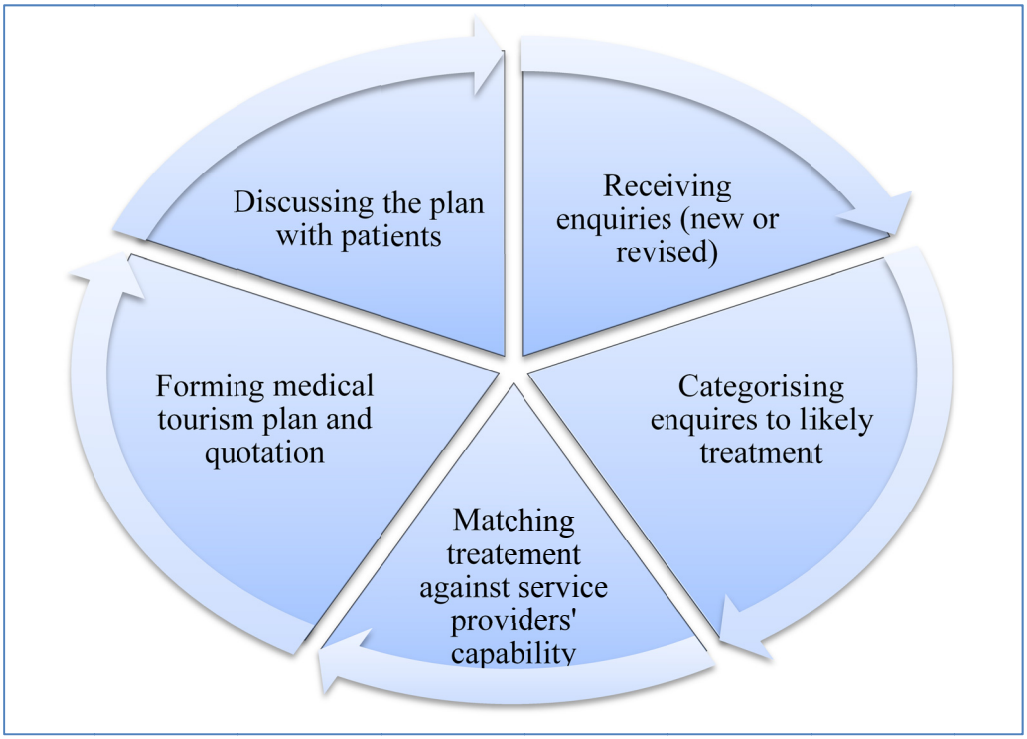


Figure 2.1: Illustration of TMG's business process cycle

Figure 2.1 illustrates TMG's business process cycle from receiving an enquiry to finalising a medical tourism contract. Advertising generated many customer

enquiries, which needed to be manually categorised and matched to likely treatment procedure. Following this, the available service providers were manually matched against the enquiries based on their treatment profile and capability. Finally the recommended medical tourism solution and quotation were fed back to the initial enquirer. The whole cycle may be repeated as many times as required to get the patients' agreement on the treatment plan proposed by TMG.

This whole process required expert company resources, and made the business fundamentally unprofitable in the eyes of the author. Amazingly at the same time, many new businesses were entering this market with a similar business model.

To enhance the process, TMG proposed a grouping of their service providers and the customer enquiries against the same ontology, so that smart automatic matching could be achieved. TMG's services were divided into eleven main groups, as Figure 2.2 shows (top level ontology): *General Surgery (abdominal surgery), Cardiology (Cardiac Surgery), Neurology (neurosurgery), Dentistry (Dental Care), ophthalmology, Cosmetic Surgery, orthopaedics (orthopaedic surgery), Comprehensive Health Checks, Otolaryngology (head and neck surgery), Paediatrics (paediatric surgery), and Vascular surgery*.

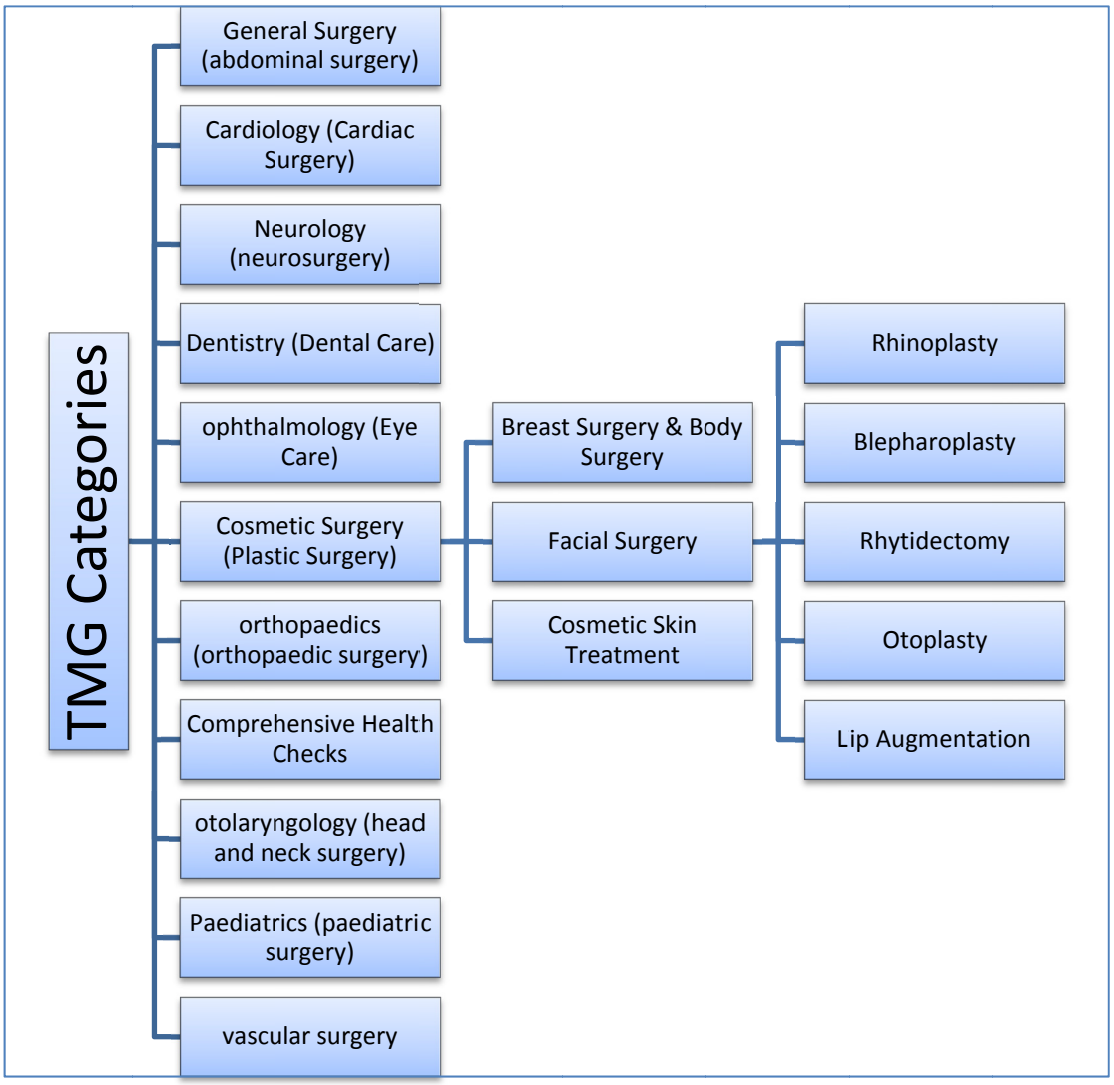


Figure 2.2: Illustration of TMG's ontology structure

It is noticed that TMG also used alternative descriptions (thesaurus in most cases) of their proposed concepts. For instance, "dentistry" is also described as "dental surgery", and "otolaryngology" is also named "head and neck surgery". The inclusion of non-expert terms reflects TMG's efforts to present "patient friendly terminology" in order to connect generic terms to their medical specialist terms.

Within each category, more detailed information relating to the medical procedure is structured (50 categories in lower level ontology). For example in Figure 2.2, under

the category of "Cosmetic Surgery", TMG present their services provided within this category, including: *Breast Surgery & Body Surgery***,** *Facial Surgery and Cosmetic Skin Treatment*. Additionally, typical procedures (instances) within each of these categories are listed as third level categories (240 instances in total). For instance, Facial surgery is further divided into *Rhinoplasty, Blepharoplasty, Rhytidectomy, Otoplasty and Lip Augmentation*. As in the higher level, TMG also utilised non-expert terminology to describe the categories at this level. Once again taking the 'Facial Surgery' category as an example, the procedures involved are associated with their natural language descriptions as:

Rhinoplasty - Nose Re-shaping

Blepharoplasty - Upper / Lower Eyelid Surgery)

Rhytidectomy - Face & Neck Lift

Otoplasty - Ear Pinning Surgery

For each of these sections, detailed explanations of surgeries and related treatments were also translated into non-professional language so that patients can be better informed before they take any action. For example, Otoplasty was described as:

"Otoplasty is a procedure used to modify the size and shape of the ears. In most cases is carried out to set prominent ears back closer to the head or to reduce the size of large ears. Protruding ears are often a family characteristic; in some cases one ear is more prominent than the other and many people feel self-conscious about them. During ear pinning your surgeon makes a cut behind your ear, close to the groove between your ear and your head. Adjustments are then made to the cartilage so that your ear lies closer to your head. If the lobe of

your ear is especially large you may choose to have a small procedure to reduce its size. Ear pinning is most often performed during childhood (after the age of five), but can also be performed on adolescents and adults. …"(TMG, 2009)

A similar approach has been adopted by the National Health Service (NHS) to provide services to patients for self-help health checks (NHS, 2010).

However, a study that analysed TMG's customer enquiries over a four year period revealed that their current categorisation was able to match less than 50% of the terms that appeared in the enquiries, although TMG has tried to collect terms that were used by patients. The TMG ontology covered most health providers' capabilities, but the function of bridging the professional terminology and non-expert vocabulary still failed. It was surmised that a lack of a broad coverage on non-expert terminologies led to such a failure.

There were also many occasions when enquiries were not fully interpreted, and hence, they were not allocated to all necessary categories, sometimes even to the wrong categories. For example, spinal surgery enquires generally involve a cardiac check; however, TMG's service ontology did not provide such a connection. It appeared that a lack of internal links among entries led to insufficient reasoning.

Such issues were not unique to the TMG's ontology, GALEN (Rector et al., 1995) was another example of medical terminology ontology lacking rich internal relationship.

2.1.2 GALEN

GALEN (Rector et al., 1995) was implemented to serve as a classification and coding system to help in categorising information, in much the same way as TMG's categorisation. GALEN also proposed the same hierarchy structure.

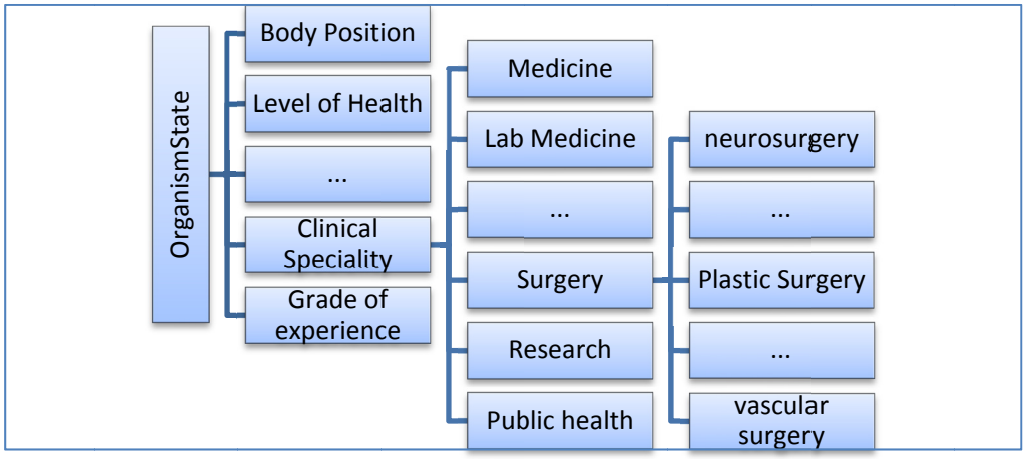


Figure 2.3: Example of GALEN on plastic surgery

The output of GALEN comprised 4 layers, as in the "plastic surgery" example shown in Figure 2.3:

- The high level ontology to present major categories (OrganismState for "plastic surgery")
- The Common Reference Model to present sub categories (Clinical Speciality for "plastic surgery")
- Detailed extensions to describe sub categories (Surgery for "plastic surgery")
- Composite extensions to provide relationships between category members (plastic surgery)

It was observed that GALEN provided insufficient terminology in the medical area, the terms were so concise that even thesaurus of the main concepts were excluded. For example, "cosmetic surgery" was not a concept in the GALEN terminology. Although there seemed to be clearly defined layers in the GALEN ontology, a key statement by the GALEN team was that clear boundaries do not exist between these levels, because there were no restricted logical criteria to separate them. However, GALEN still applied such a hierarchy structure due to its advantage in automated systems.

GALEN's emphasis on unclear boundaries highlighted that there should be much richer internal relationship in the hierarchy structure. GALEN recognised the importance of rich internal relationships. This requires a network of relationships which not only strictly specify theoretically axiomatic ties vertically from a concept to its instances, but also to horizontally link concepts from practical perspectives.

However GALEN itself did not provide such rich internal relationship. For example, in the GALEN ontology, "orthopaedic surgery" only directly links to its upper category "surgery", the same gap as was left in the TMG ontology.

The SNOMED CT project attempted to provide much richer internal relationships in its medical ontology.

2.1.3 SNOMED CT

SNOMED CT itself was founded as early as 1965 as SNOP - Systematized Nomenclature of Pathology. To enrich the internal relationships, SNOMED CT (IHTSDO, 2009) spent more than ten years ("SNOMED CT" was rebranded and officially launched in 1999) consulting with thousands of medical professionals from six countries, and identified nearly one and half million relationships between its concepts. These relationships organised the concepts from different perspectives and understood them in multiple ways, such as through various descriptions or connections with other concepts. This "relationship sea" greatly improved the number of internal links, and even provided some links to other domains. The great number of internal relationships changed its structure to a network form containing: Concepts and Concept Descriptions (the terms or names assigned to a concept).

This simple concept–description structure reinforced the vague boundary proposal from the GALEN project, by filling gaps between concepts with shared descriptive terms. In this structure, concepts can mutually define each other, and this addresses the shortage of internal links experienced by TMG and GALEN.

For instance, SNOMED CT allows a child node to have multiple parents to enrich the vertical structure, as well as horizontal connections among concepts in the same level to link the same level concepts. If such a structure were applied to the TMG's example on cosmetic surgery (Figure 2.4: original TMG relationships were illustrated in blue), it could turn the hierarchy into a network like structure (Figure 2.4: added vertical

relationships are highlighted in red; and added horizontal relationships are highlighted in green).

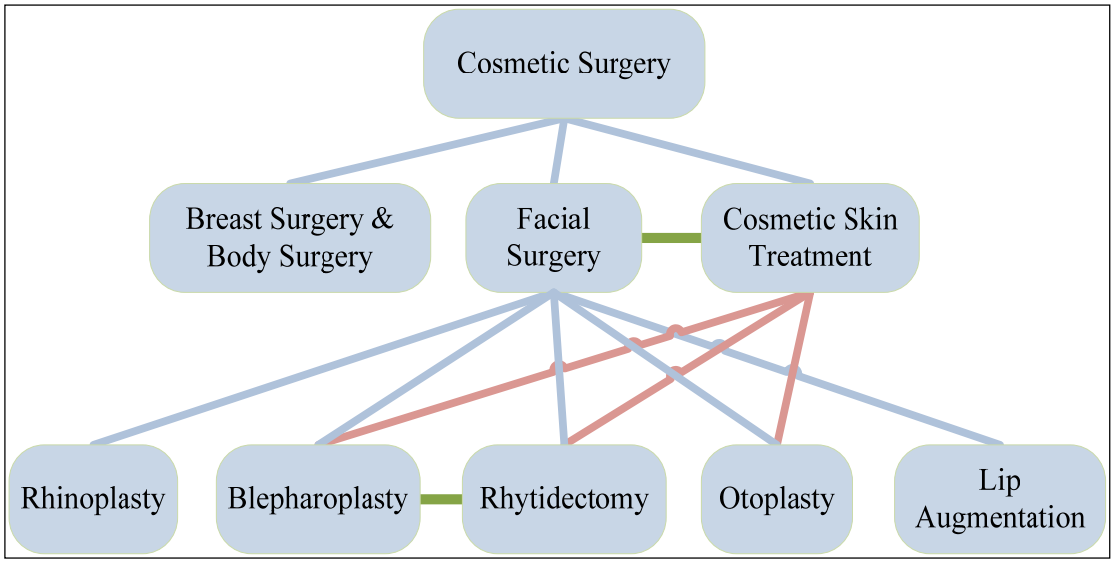


Figure 2.4: Illustration of SNOMED structure with example of TMG cosmetic surgery

In theory, SNOMED CT could be the optimum ontology that can be reused to provide ontology for applications such as in TMG's case. However, in practice SNOMED CT was difficult to reuse for the following reasons:

1. Size

With 1.5 million links, SNOMED CT has captured more than 311,000 unique concepts and more than 800,000 descriptions. It has become one of the largest medical ontology that organisations can reuse. However, Heiner and Michel argued that it might have grown too large to be effectively used or maintained (Heiner and Michel, 2004): applying such a complex ontology required highly trained experts;

and traversing 1.5 million relationships for reasoning reduced computable efficiency and thus it may need high performance IT infrastructure.

Considering the current facilitators in the medical tourism industry are mainly SMEs like TMG, it is unlikely that such organisations would directly apply the whole SNOMED CT ontology. Customising SNOMED CT into smaller scale may be an alternative approach, which is also recommended by IHTSDO for smaller organisation application.

2. Customisation

SNOMED CT does not directly provide any "light" version, although it claimed to be scalable and flexible. It relies on the individual organisation (who tries to reuse SNOMED) to develop its own "appropriate" subsets/segments. There have been developments in automated ontology segmentation methods that could limit the need for large ontology. For example, a subset of ontology can be created by mining all linked concepts and relationships to a given concept or relationship representing the subset (Noy and Musen, 2001); or a segment could be highlighted by only exploring concepts and relationships "atomically" describing a given concept, without considering the descriptions' further reach (Seidenberg and Rector, 2006).

These methods proposed the traversal of all related ontology structure for a target ontology. When a large ontology like SNOMED CT is the target, computational efficiency can be a barrier for organisations or individual practitioners in forming

subsets for their particular needs (Patrick et al., 2008). Additionally, these methods are intended to extract some "relevant parts" of ontology from the perspective of a given domain, but they do not necessarily provide the "matching" subset/segment from the application environment perspective. To reform a subset or segment, the practical needs in the real-life application may also play an important part in providing application oriented concepts collection and relationships among concepts. In fact, matching ontological structure between domain viewpoint and application viewpoint has been emphasised as a fundamental problem (Jarrar and Meersman, 2009) while developing ontology for practical applications.

Furthermore, SNOMED suffered drawbacks such as an unclear identification from different healthcare aspects and was out of alignment with other widely accepted ontology claimed some critics (Schulz et al., 2009). Reusing SNOMED for TMG's business application would require reconfiguration (or reconstruction of a subset), but reconfiguration would be unrealistic due to the effort required to customise SNOMED CT and align it with practical usage in the TMG scenario (and further alignment to its suppliers' ontology). Therefore, it may be difficult for TMG to directly apply the SNOMED ontology in practice.

3. Non-specialist terminology

SNOMED CT was designed to facilitate diagnosis for professionals, not for improving natural language information categorisation. Although the number of concepts and relationships are greatly increased, when compared to other medical

domain ontology, they were still not sufficient to cover broader general terms and relationships. For example, “Otoplasty” neither linked to “Cosmetic surgery”, nor to “ear pinning”. In fact, “ear pinning” is not even included in SNOMED ontology.

SNOMED CT appeared to have coverage of the understanding of a large group of healthcare experts, but it is questionable whether it accurately represents the terminology of non-specialists. Schulz and colleagues highlighted that SNOMED CT had the weakness of an insufficient number of connections to non-specialist terms, as well as limited ambiguous relationships towards them (Schulz et al., 2009). Although ambiguous relationships toward non-professional terms may encourage fuzzy matching to the professional terms, the limited coverage of the relationship to the non-expert terms still undermined SNOMED CT’s ability to meet this research’s requirement on broad conceptual coverage and bridging gaps between specialist and non-specialist terms.

4. Maturity

Although it has taken doctors and nurses from six countries more than ten years to reach the current stage of progress (by the end of 2010), SNOMED was still a working draft and abstract model. SNOMED CT seems immature compared with those ontology that have been established and adopted in more practical applications (such as SIC (National_Statistics, 2008) or UNSPSC (Granada_Research, 2001)for business categorisation).

Research in the Unified Medical Language System – UMLS (UMLS, 2010) project believed that directly collecting information from first hand sources such as healthcare professionals was the cause of issues discussed above. Therefore, they preferred to reuse published sources or existing ontology as source to build UMLS.

2.1.4 UMLS

UMLS (UMLS, 2010) was established by the United States National Library of Medicine as a database and applied as a large biomedical terminology. It expanded its vocabulary by considering an integration of several large sources of medical terminology including SNOMED CT, MeSH[2], etc. These sources were called "Metathesaurus" - a larger scale mapping of over 100 source vocabularies, which supplied concepts and their relationships for UMLS. Then "The Semantic Network", made up of semantic types (categories) and semantic relationships (relationships between categories), incorporated multiple sources by connecting their concepts based on the relationship specified in the semantic network. Additionally, the "SPECIALIST" lexicon provided extra natural language lexical information (UMLS, 2010).

UMLS brought linguistic factors into the ontology by reusing existing sources to cover the general public side terminology, and further linking them to specialist

[2] Medical Subject Headings (MeSH®) MeSH is the National Library of Medicine's controlled vocabulary thesaurus. It consists of sets of terms naming descriptors in a hierarchical structure that permits searching at various levels of specificity.

medical lexicons through the semantic network. This research identified that relationships from linguistic sources could be used for ontology building to collect non-professional terminology.

Although reusing existing sources certainly accelerates the ontology building process resulting in lower cost, such sources are updated relatively slowly (if at all!) than direct up-to-date information collection from humans such as that by SNOMED CT. For example, an emerging breast enhancement medicine "b2up" was included, and linked to the relevant category in the TMG ontology a few months after its appearance in the market (August 2009) due to popular patient enquiries. It was not included in any of the medical ontology described by December 2010.

A survey of TMG's healthcare providers (appendix 2.1) also revealed the need to obtain up-to-date treatment information: most of these healthcare organisations in the survey were involved in medical tourism, and they needed to react quickly to new treatments being made available elsewhere. New services are appearing in the market frequently, and it is not easy for these organisations to capture such changes and update their service menu in a complete or timely manner.

2.1.5 Medical ontology Summary

The review of TMG's needs and the other medical ontology derived highlighted the practical service issues in applying them. The discussion revealed some target characteristics that this research should consider:

1. Richness of internal relationship

Medical domain ontology have already proposed rich internal relationship to enhance reasoning among concepts. Ontology targeting a wide application area should incorporate a networked structure such as SNOMED CT's "relationship sea". Domain specific relationships should be linked from different perspectives (such as application oriented) to form such a "relationship sea".

2. Use of semantic relationship:

Where linguistic relationships (via semantic relationships) have been brought into ontology (UMLS and TMG), ontology have shown an ability to provide relationships between specialist and non-specialist terms. When generating and applying ontology for multi-disciplinary work, linking between specialist domains will require a rich non-specialist language interface. This is a key goal of this research.

3. Lightweight

Ontology may be described as "lightweight" if they have relatively flexible definition on concepts and their relationships. If some axioms and constraints were added to a lightweight ontology, to strictly restrain the concepts and relationships definition, the restriction added ontology would be treated as a "heavyweight" ontology (Uschold and Gruninger, 1996). For example, ontology may have a relationship between "red" and "apple" via "red $\xrightarrow{is\ related\ to}$ apple", which is a

"broad" relationship without heavy axiomisation. If a second ontology has been tightened to a combination of "color $\xrightarrow{is\ a\ property\ of}$ apple" and "red $\xrightarrow{is\ a\ kind\ of}$ colour", through which concepts and properties (or concept descriptions) have been rigorously defined, the first ontology can be treated as a lightweight ontology.

The medical ontology discussed are expected to be a bridge between non-professionals and domain experts. They allows concepts to describe each other (not strictly prohibiting mixed use of concepts and concepts descriptions), and try to cover the non-professional area with semantic relationships (not forcing rigorous axiomatically relationship). These "lightweight" characteristics indicate that it is important to accept relatively flexible concepts and relationships' definitions, while richer internal relationships and broader concept coverage are required. (Further discussion on ontology weight specification is provided in Appendix 2.2)

4. Source reuse

Medical domain ontology have demonstrated that reusing existing sources could help reduce the reliance on domain experts, and speed up the ontology building process. Although there are concerns that require addressing, such as incorporation of the latest developments in the domain(s), as illustrated in UMLS, source reuse may be an option for quickly generating a specific ontology.

5. Up-to-date information

TMG's case showed that the ability to respond to rapid changes in the related subject areas is an important function of ontology. Existing ontology appear to have failed to address this important area due to the difficulties in collecting, structuring and inserting new information: it is too expensive to incorporate direct information updates from domain experts, and difficult to find other indirect sources that provide this capability.

Besides these issues, medical ontology do not readily address multi-disciplinary usage due largely to the way they are derived. In parts of the engineering sector, ontology have been more widely accepted, are more mature in their development, and are maintained by accredited international bodies. The next section explores engineering ontology via a case study of a Collaborative Engineering Marketplace, the West Midlands Collaborative Commerce Marketplace (WMCCM). Engineering ontology emerged soon after the emergence of online business applications, and have expanded rapidly since (Fensel et al., 2000, Berners-Lee, 2000).

2.2 ENGINEERING SECTOR CASE STUDY

The increasing need for information exchange and communication stimulated the generation of ontology (van Heijst et al., 1997, Mizoguchi et al., 1995), and engineering was among the earliest sectors to benefit. Ontology in this sector are considered to be more mature than medical sector ontology. Many ontology have been

built by organisations in the engineering sector, often in the form of industrial classifications to allow information exchange among organisations.

The experience with WMCCM and its ontology engineering uncovered issues in reusing such sources. WMCCM utilises its ontology to automatically match tender opportunities with company competency, by categorising tender and company information against its ontology covering the engineering sector and associated activities (section 1.2). When WMCCM started building its ontology, it followed a mixed approach: lower levels were derived from actual company information; upper levels from standard classifications such as SIC.

2.2.1 UNSPSC and SIC

The United Nations Standard Products and Services Code (UNSPSC) was designed as an upper level ontology to facilitate e-Business for quicker and more accurate procurement, marketing and sales. It classifies products and services in multi-sectors into a five level taxonomy. An example for welding classification is provided below and illustrated in Figure 2.5:

- Segment: 55 segments (Industrial Production and Manufacturing Services)
- Family: 419 families (Machining and processing services)
- Class: 2551 classes (Welding and brazing and soldering services)
- Commodity: 217189 commodities (Welding services)

- Business Function: The function performed by the actual organization in support of the business

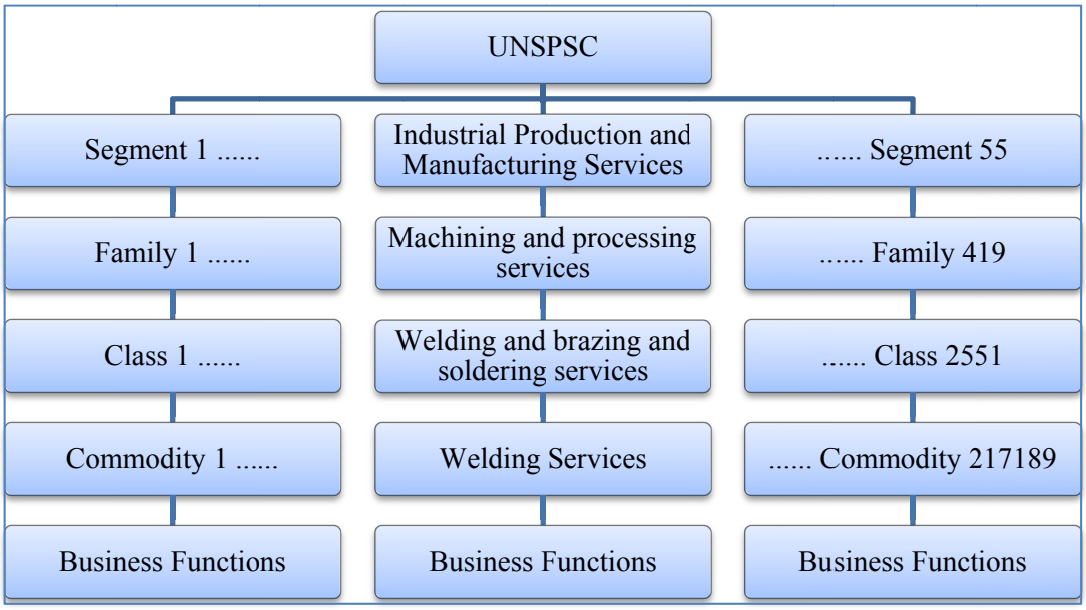


Figure 2.5: Illustration of UNSPSC structure with "welding" example

UNSPSC was designed for high level guidance, and it does not appear to be practical at the regional and country level. Different countries have thus proposed their own industry standards according to their own requirements (Fairchild and Vuyst, 2002).

In Europe, UNSPSC was modified to develop the regional business classification "Nomenclature statistique des Activités économiques dans la Communauté Européenne" (NACE) – common statistical classification of economic activities in Europe. Most European countries have made further modifications to the regional standards.

For example, the United Kingdom Standard Industrial Classification of Economic Activities (UK SIC) is the standard industrial classification widely accepted in the UK. It is used to categorise businesses in accordance with the scope of their economic activity (National_Statistics, 2010). UK SIC was initially designed in collaboration with NACE, and modified it to suit UK industries. The UK SIC contains five levels.

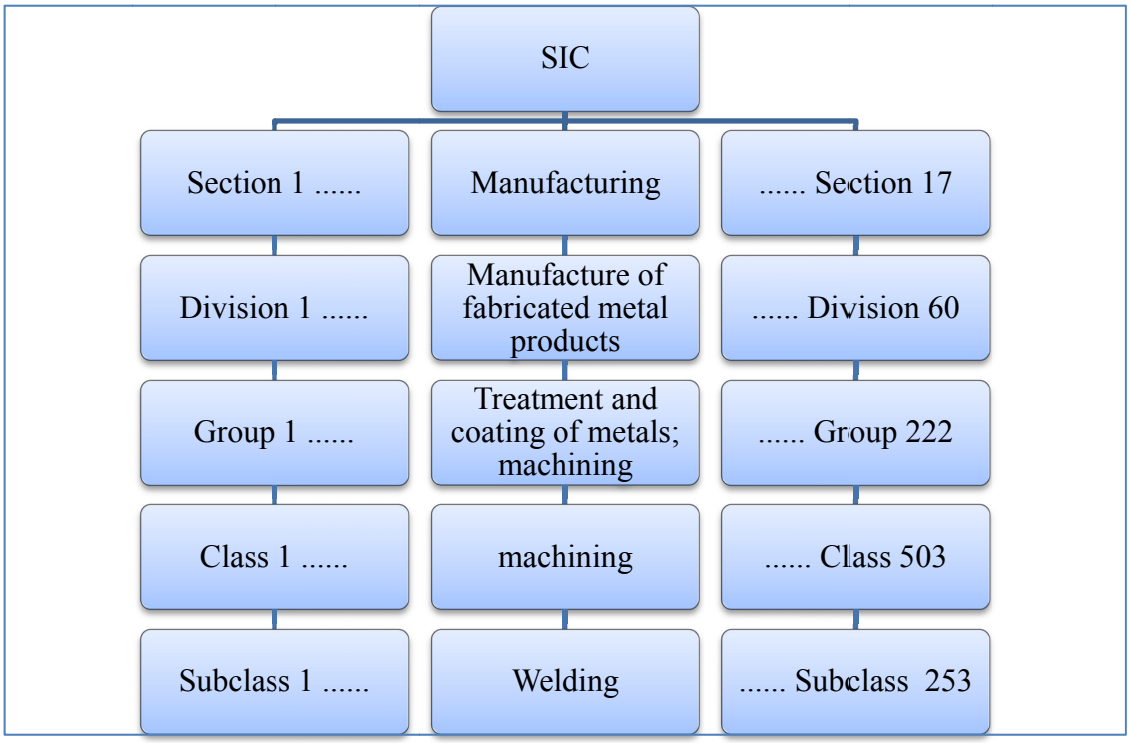


Figure 2.6: Illustration of SIC structure with "welding" example

An example of the welding classification is provided and illustrated in Figure 2.6:

- Sections: 17 sections (Manufacturing)
- Divisions: 60 divisions (Manufacture of fabricated metal products, except machinery and equipment)

- Groups: 222 groups (Treatment and coating of metals; machining)
- Classes: 503 classes (machining)
- Subclasses: 253 subclasses (welding)

The UNSPSC and UK SIC differ significantly in structural levels and the members within them. These assigned levels do not directly align with each other regarding to the concept at the level and the instances in the levels. They have been customised extensively based on the nature of their requirement (business oriented, region specified, or even organisational required). These ontology have tried to reuse existing knowledge in their fields but in practice they have been separated in response to specific needs. It may be that these "needs" are more political in nature, restricting outside competition through a lack of "interoperability".

Although fundamentally they were supposed to represent the same knowledge and its structure, they appeared to be derived from different viewpoints. The viewpoint oriented structure means they require considerable modification to communicate with each other. This scenario illustrates that while ontology have reused such sources, they still require considerable consultancy from domain experts to clarify the relationships between such sources. These sources are classification systems, in which the original design allows gaps between classes (concepts) to form clear boundaries (Jacob, 2004). These gaps may omit some areas, so that full coverage of the industry is not available.

For example, UNSPSC showed a lack of broad cover of the classes, especially with regard to the actual products and services, and insufficient relationships to demonstrate inheritance and commonality among classes (Corcho and Gómez-Pérez, 2001). Corcho and colleagues also pointed out that the condensed classes produced by experts did not have enough attributive descriptions around concepts. In other words, there were too few words to cover a much larger generic keywords variation in natural language information. Finally, classes (concepts) proposed by such sources tended to stay at a higher level compared with the company/user proposed classes for WMCCM. The high level classes were found not to be specific or detailed enough to differentiate between the competences proposed by companies.

These issues suggest that directly summarising ontology from existing sources (a single top-down procedure) may not satisfy WMCCM's practical requirement for broad coverage and rich internal relationship. Therefore, WMCCM followed a mixture of top down derivation and bottom up synthesis (collecting terms and relationships from actual ontology users).

2.2.2 WMCCM

WMCCM combined the UK SIC and first-hand data collected from engineering industry (aligned SIC classes with capability in practice) to produce a WMCCM business categorisation (WMCCM, 2010) based on the following structure (an example of welding classification follows and is illustrated in Figure 2.7):

- Tier 1: 24 Industrial processes (machining)
- Tier 2: 209 sub processes (welding)
- Tier 3: 629 Business Function \ descriptions (arc welding)

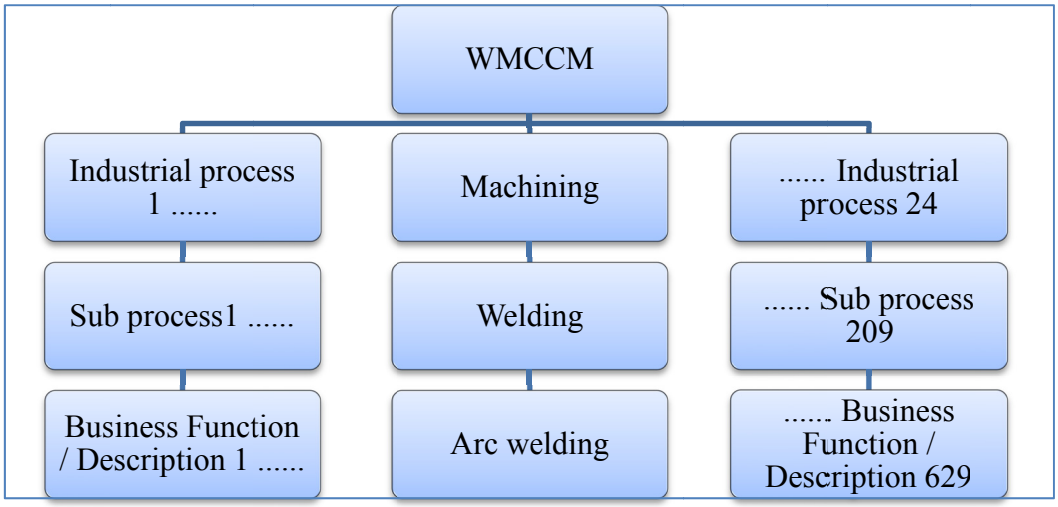


Figure 2.7: Illustration of WMCCM structure with "welding" example

This structure focused on what "capabilities" the engineering businesses may have rather than what "products" they can make; the change of focus also led to different "capability" hierarchies from SIC's product/service hierarchies. Additionally, WMCCM also considered other factors such as computational efficiency and ontology presentation for both professional and non-expert users.

As a result, the WMCCM ontology differs significantly from other engineering ontology in both content and structure (comparing with UNSPSC and SIC). WMCCM's viewpoint (in terms of how it was designed to be used) of processes in the engineering industry certainly played a part in constructing such ontology. Another strong influence was the terms and their structure from participating

companies based on their real needs. The differences result from the fact that ontology are affected by the target application and need to be orientated towards the target applications.

The customisation undertaken did not fully satisfy the WMCCM's tender matching process. For instance, the following tender was looking for capability in ICT system design and development, but was classified as a tender requiring "Service and Supply Chain Management".

> *"UK-Bristol: equipment management system. The Medical Equipment Management Organisation (MEMO) is a semi-commercial organisation supplying medical equipment management and maintenance services to a variety of public and private sector healthcare organisations. …… The current system needs replacement as it no longer meets all our demands and is no longer supported. One of the key requirements is the ability to migrate current data from the current system into any replacement system. Interfaces will be required to other hospital systems. Following is a list of the functions any new system must include. Management (inventory) job management; Call logging of breakdowns & repairs; Prioritisation of jobs; Monitoring the status of jobs; Printing or e-mailing of jobs sheets; Planned maintenance; Scheduling (planning); Task sheets (tick lists); Service & maintenance history, capable of searching back to 20+ years of data; Contract management; Financial records; Service level agreements / customer contracts; Customer billing / links to finance systems; Customisable reports / Statistical analysis; Customisable engineer notifications/alerts e.g. jobs about to breach SLA's; Security, Password control / different levels or profiles, Encrypted passwords; Compatible with business continuity requirements; Audit trail; Industry specific software package."*

The existing WMCCM ontology captured "Service and Supply Chain management"; however, it failed to interpret the actual requirement: software design / ICT maintenance & support. Although the ontology had some ICT processes as concepts, these concepts lacked fuzziness in the descriptive terms as the classification was relatively strictly defined. The lack of fuzziness denied fuzzy reasoning towards the same set of descriptive keywords but around different concepts. As a result, keywords such as "management", "inventory", "monitoring", "engineer" and "maintenance" led the system to deduce "supply chain management and service" as the concept domain.

It is arguable that the linkage within concepts could be regarded as "cross domain" at the micro level (such as the alternative descriptions in the ICT class and concept descriptions in SNOMED). However, at the overview level – taking ontology as a whole – example ontology still lack fuzziness at the edges. This may be because the core sources and the domain experts defined domain concepts as a typical classification task. In typical classification, an entity can only be explained by one series of reasoning rules and allocated to one class.

A faceted classification system may be a solution, as it allows the assignment of multiple classifications to an object, enabling the classifications to be ordered in multiple ways, rather than in a single, pre-determined, taxonomic order. A faceted classification system provides its concepts with multiple "facets", which is borrowed from Library Science to express different "aspects of meanings" (Ranganathan and

Gopinath, 1967). For example, a concept "boring" may have a facet of "emotional statement" that means boredom, and it may have another facet of engineering process that refers to a specific type of hole drilling process. As a result, a faceted system can assign both facets of "emotional statement" and "engineering process" to the concept "boring".

Faceted classification can create network based systems that are flexible in nature, and this may overcome the issue for which the commonly adopted hierarchical ontology structure could be too rigid. However, in the same way that building domain ontology requires extensive input from domain experts, building multiple "faceted" ontology (could be understood as cross domain) may require even more consultation from experts in different domains (Giunchiglia et al., 2009). Furthermore, as a type of classification system, faceted classification still requires a "clearly defined" and "mutually exclusive" description of its concepts (Specia and Motta, 2007). Thus, a faceted system may still leave gaps between concepts, which can act as stops to reasoning processes.

Experts tend to use specialist domain language to accurately classify terms. This desire to be "pedantic" and precise makes building ontology in the necessary level of fuzziness through the use of non-specialist terms conceptually difficult for domain experts. The source ontology (UNSPSC and SIC) used by WMCCM also lacked the necessary level of fuzziness/redundancy to be able to be applied to human oriented systems. The reuse of existing ontology only provides the necessary structure and

description of domain knowledge, but lack relationships to terms that not strictly bounded by the core domain terms. For example, ICT relevant terms such as e-mailing, SLA (Service Level Agreement), password control, searching, data, and software package were normally not defined as concepts in ICT processes, but are semantically linked with many ICT processes.

Vander Wal claimed that online tags may provide wider linkages to more relevant terms from end users. These tags can be entered by online users in a free form, and a collection of such tags may form a "folksonomy" for obtaining related terms (Vander Wal, 2007). However, folksonomies can be chaotic due to the "free form" nature of user tagging, as any tag and any relationship from any perspective can freely enter the structure (Specia and Motta, 2007). Additionally, this chaos of disordered concepts and relationships create a barrier for organising domain focused terms into subsets (Xu et al., 2006), thus it is difficult to utilise folksonomies to serve domain oriented ontology.

Linkages to more relevant terms could be more reliably provided by linguistic connections from natural language: the connections should neither be strict "concept – description" or "concept – sub-concept" relationships within a domain defined by formal ontology, nor completely free formed random links proposed by folksonomies. Therefore, in a way similar to medical ontology, the WMCCM ontology also required linguistic support to enrich the relationships between core members and members at the edges.

The WMCCM approach tackled these issues by using non-experts as intermediaries in the process of information collection from the domain experts, but this approach is very costly. Over 75% of the cost of developing the WMCCM system was the costs of employing these non-specialists to competence profile 400 plus companies, and through this activity refine the WMCCM ontology. UMLS firstly introduced source reuse to reduce the knowledge collection effort. The same approach could be applied to reuse linguistic sources to increase the coverage of natural language relationships between concepts.

Furthermore, reusing exiting relationships (especially linguistic relationship) may provide measurable weight on the relationships. This could enable conversion from traditional descriptive relationship (normally logical relationship which can be understood as "linked" or "not linked") to numeric relationship, a measure of how "much" two terms are "linked" (Tho et al., 2006). Numeric relationships may be the key to enable fuzzy logic among concepts (Lau, 2007, Zhai et al., 2008), so that a normal ontology may gain fuzzy matching ability to become a fuzzy ontology. It seemed that rich internal relationships from existing linguistic source may bring extra benefits towards fuzzy reasoning among concepts.

2.2.3 Engineering ontology summary

The discussion on engineering ontology uncovered similar issues to those identified in medical ontology. To be effective in ICT systems used by non specialist, there is a

requirement for rich internal relationships, source reuse and an ability to incorporate the latest domain knowledge:

1. Relationship reuse

When reusing sources, the result should inherit the original relationships from the sources' perspectives (such as UMLS' semantic network). However, such process can be expensive. Identifying target relationships and reusing them in the new ontology may require intensive effort from domain experts and the relationship may not fully represent the application environment. Thus, it is important to ensure that the reuse sources provide the required concepts and the relationships between them.

2. Fuzziness around the concept

The WMCCM case highlighted the requirement for fuzziness around concepts. This fuzziness could be gained from increasing the semantic relationships with non-expert terminology. The "relationship sea" with rich internal relationships among concepts needs to be expanded in order to contain a network of both expert and non-expert terms for multi-disciplinary usage. The conversion of the descriptive relationship to numeric relationship may play an important role in the fuzzy matching mechanism.

3. Application orientation

Engineering ontology are structured and populated to fit their special needs. Thus the way they are intended to be used determines how they are formed. Application

orientation is also emphasized in the Developing Ontology-Grounded Methods and Applications (DOGMA) approach (Jarrar and Meersman, 2009), where the ontology structure is designed as "double articulation" – a domain specific articulation and an application specific articulation.

The practical requirement of the ontology application environment also drives the engineering ontology discussed to stretch the traditional ontology boundaries in terms of representation and weight specification (see appendix 2.2). For example, some engineering ontology accommodate concepts from multiple domains to break the domain focus limitation; they mix terms from the conceptual abstract level with the practical instance level to break concept representation coverage limitations; they manipulated "weight" of the ontology (constraints from internal structure and relationships) to modify relevant terms according to the application environment.

This implies that ontology that is developed for practical purposes may contain relatively fewer constraints on the relationships and vague boundaries of the representing realms. Application orientation may also require a tree type hierarchy output (as SIC, UNSPSPC and WMCCM adopted) to achieve user friendly interfaces and computational efficiency (Giunchiglia et al., 2009).

4. Linguistic connection between terms

It is believed that expert condensed domain vocabulary could explicitly specify a subject area. However, more details also need more, less-explicit assumptions. Such

assumptions could be derived from linguistic connections between terms (Jarrar and Meersman, 2002). The engineering and medical ontology review in this research also highlighted the need for linguistic connections between terms. Linguistic ontology concentrate more on explaining the relationships between terms rather than modelling a particular domain. The intention of linguistic ontology was to capture natural language concepts and instances, in conjunction with their bounding within grammar units to facilitate building other ontology(Gómez-Pérez et al., 2004).

2.3 LINGUISTIC ONTOLOGY

Several linguistic ontology have been built to provide natural language terms and their relationships, such as WordNet, Microkosmos, and SENSUS.

2.3.1 WordNet

One of the biggest projects, WordNet implemented at Princeton University, is a large lexical database. It is structured based on the cognitive synonyms of English terms (aka synsets in WordNet, each of which represents a lexical concept) rather than their formation (Miller et al., 1990, Miller, 1995).

WordNet collected natural term relationships by focusing on more complex lexical relationships. WordNet's lexical relationship revealed human natural languages relationship in a narrow sense without full semantic relationships. For example, "blue" in WordNet had a relationship to "parties whose uniform or badge is blue", but it is

hard to link "the Conservative party in the UK" with blue, as there was no direct lexical relationship among them.

WordNet also enriched the hierarchy structure by "weaving a net of lexical relations", in a similar way to SNOMED CT's relationship sea. Such rich internal structure provided sufficient information for WordNet to absorb new information and turn the incoming new knowledge into part of the ontology. However, this structure was established from a lexical viewpoint with no domain specific assistance, so extracting part of the ontology for a domain was complicated. It may require traversal of all possible domain keywords and re-calculation of the relationships to specify domain related terms and general terms. Such issues also exist elsewhere, for example UMLS without domain specific support is mainly referenced as a meta-thesaurus rather than a practical ontology.

Improvements in the relationships between lexical domains have been tested by the Microkosmos ontology.

2.3.2 The Mikrokosmos ontology

Mikrokosmos (Mahesh, 1996) paid more attention to the clarification of the relationship between the lexical meaning of terms in language, particularly with respect to their environment and their natural meaning in language neutral representations. In other words, the distinctness between professional terminology and general terminology, and also their different roles in the same ontology, are

highlighted in Mikrokosmos. This change bridged the gap between linguistic ontology and domain ontology, so that natural languages relationships can be applied to domain ontology for adapting multi domains or translating terms between specialists and non-specialists. This approach was also applied in the SENSUS project.

2.3.3 SENSUS

The Natural Language Group at ISI[3] developed SENSUS with a focus on natural language (Swartout et al., 1997). SENSUS was developed to improve information retrieval and machine translation by exhibiting the deliberation and inference of deep semantic relationships between words.

The top level structure of SENSUS was created by taking the essential branches from ontology such as Mikrokosmos and WordNet. It retained the relationships with these ontology. As a result, the mixture of semantic networks provided opportunities to achieve a connection from SENSUS to other sources. This provided a foundation for SENSUS to include cross domain terms.

One of the distinctive characteristics of SENSUS was the method by which semantic relationships were identified. It was based on the assumption that the definition or description of a certain word should contain a small but highly relevant set of

[3] The Information Sciences Institute (ISI) in the University of Southern California

prepositional terms, which enable connections between this word and other words. For example, "sadness" may be described by a set of terms: depressed, blue, gloomy, melancholy and sombre. Such connections reflect broader semantic relationships: a particular concept represented by a certain word may be defined by a number of propositional words in relation to it. This showed that ontology could utilise a broader relationship than lexical relationship – a co-occurrence relationship – between words that were provided by documents or domains.

Another distinctive characteristic of SENSUS is that it upgraded the simple hierarchy structure for ontology to a network type structure (similar to the network structure proposed by SNOMED and UMLS). A network type structure hugely increased the internal relationships so that higher reasoning mechanisms can be applied to ontology. Additionally, a network type structure could be perceived from different viewpoints (facets classifications), this would make the ontology much more flexible in communication with other domains, or to be more easily structured into different formats according to application needs.

However, both Mikroksomos and SENSUS reuse other lexicon linguistic ontology such as WordNet, and thus, they may encounter similar reuse issues as discussed in section 2.2.

2.3.4 Linguistic ontology summary

The use of semantic relationships in linguistic ontology, particularly the "co-occurrence" relationship proposed by SENSUS, meets the relationship goals of this research in two ways:

- The design purpose of translation and cross domain coverage means that it can bridge terms between different ontology user groups.
- The semantic relationship from words co-occurrence may break the limitations of lexical and grammatical relationship, so that more sophisticated relationships may be created. This research could take advantage of such co-occurrence relationships to group semantically related terms from a differing viewpoint. (Discussion in section 3.1).

With the support of linguistic ontology, the semantic relationships required in medical or engineering ontology can be generated to enrich internal structure towards non-expert terms. This could provide fuzziness around domain (and hence ontology) edges where the non-expert terms lie, and enable enhanced multi-disciplinary communication.

The case studies have highlighted some desirable goals that an effective ontology for ICT system applications should have. The discussion so far has focused on example ontology, but a goal of this research is the "quick" generation of a reliable ontology. Therefore, an investigation on ontology building methods is required.

2.4 METHODOLOGIES FOR ONTOLOGY ENGINEERING

Ontology engineering methodologies have been employed in many projects such as SNOMED CT, UMLS, SENSUS. Several key ontology engineering methodologies has been created.

2.4.1 Cyc methodology

The Cyc methodology was one of the earliest attempts to formalise ontology engineering. It was applied to build the Cyc Knowledge Base (Lenat and Guha, 1989), which is one of the top level ontology that SENSUS refers to. It was constituted in a similar manner to SNOMED CT - by manually adding over a million pieces of consensus knowledge statements.

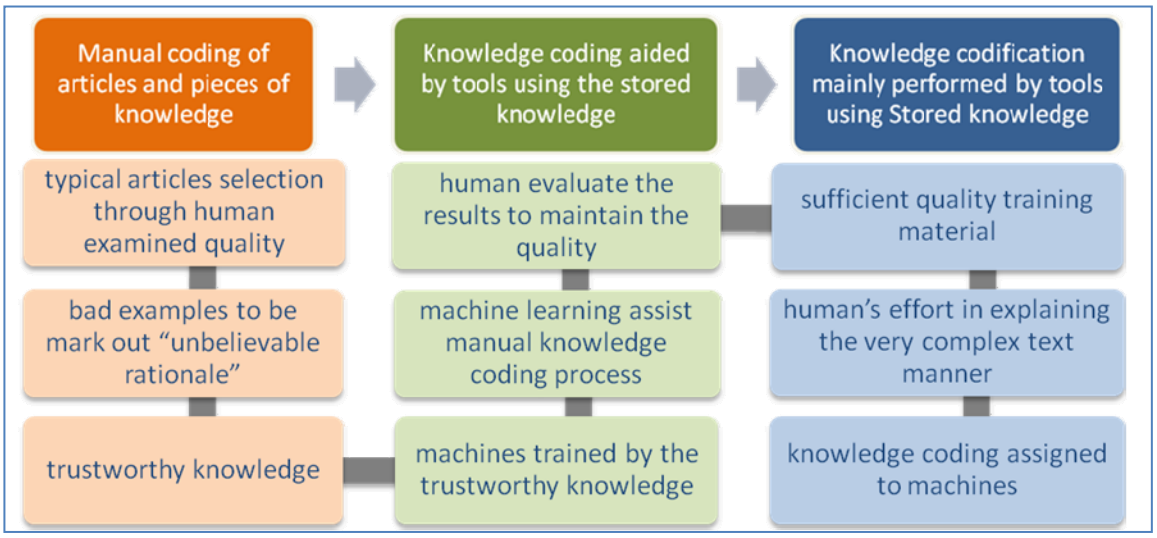


Figure 2.8: Approach to developing the Cyc ontology

Figure 2.8 illustrates the overall steps taken when developing the Cyc Knowledge Base. This linear process can be viewed as a knowledge base building process:

1. Manual coding for articles and pieces of knowledge, and
2. An ontological analysis process (knowledge coding), and an
3. Ontology presentation process (knowledge codification).

Domain experts were the starting point for building the knowledge base. Most of the knowledge in the system would be based on the opinions of a group of experts. However this may not be sufficient to cover all the different perspectives in the field and the common vocabulary of non-professionals. Domain experts were also needed in all of the later stages, resulting in a costly way of building such ontology.

However, Cyc highlighted the possibility of utilising automatic tools to assist such a knowledge acquisition process, if such tools were capable of producing the expected knowledge codifications (training may be required). This provided an option to reduce the cost of building the corpus (vocabulary and binary relationships required) of ontology.

Cyc suggested nine "to-dos" for ontology engineers but did not give advice on detailed techniques about "how" to do them. Without the technique details (how to do the tasks), it requires highly skilful ontology builders to execute the whole process. From a management viewpoint, Cyc proposed a linear development process without breaking the tasks into stages; this blurred the start and end point of each process.

Further optimisation was made by Grüninger and Fox in building the TOVE ontology (Grüninger and Fox, 1995). They introduced a methodology to break processes into

stages, and allocate expert contributions to the relevant stages. It also specified a binary relationship in the methodology to guide the ontology building.

2.4.2 TOVE methodology

TOVE's approach proposed a methodology with detailed techniques at each stage. However, the technique details limited the methodology into its own application environment. For instance, using "first order logic" to specify the terms and relationships led to its inapplicability for developing ontology, which requires other types of binary relationship, i.e. semantic relationship. Although this relationship could be altered, it was bounded to TOVE's development environment, and any alterations might require much greater consideration so as to modify the remaining part of the methodology, for use in other projects.

TOVE (Figure 2.9) suggested pre-development stages to conduct requirement analysis and application environment study. Pre-development stages provided a cautious entrance to avoid errors in a quick start. However, without evaluating the ontology output, TOVE proposed a specified exit solution to end the process. A "dead end" shut the processes down once the development was finished, but in practice, there would be a continuous improvement. The ability to accommodate further changes (such as latest information) is highly desirable for an ontology engineering method.

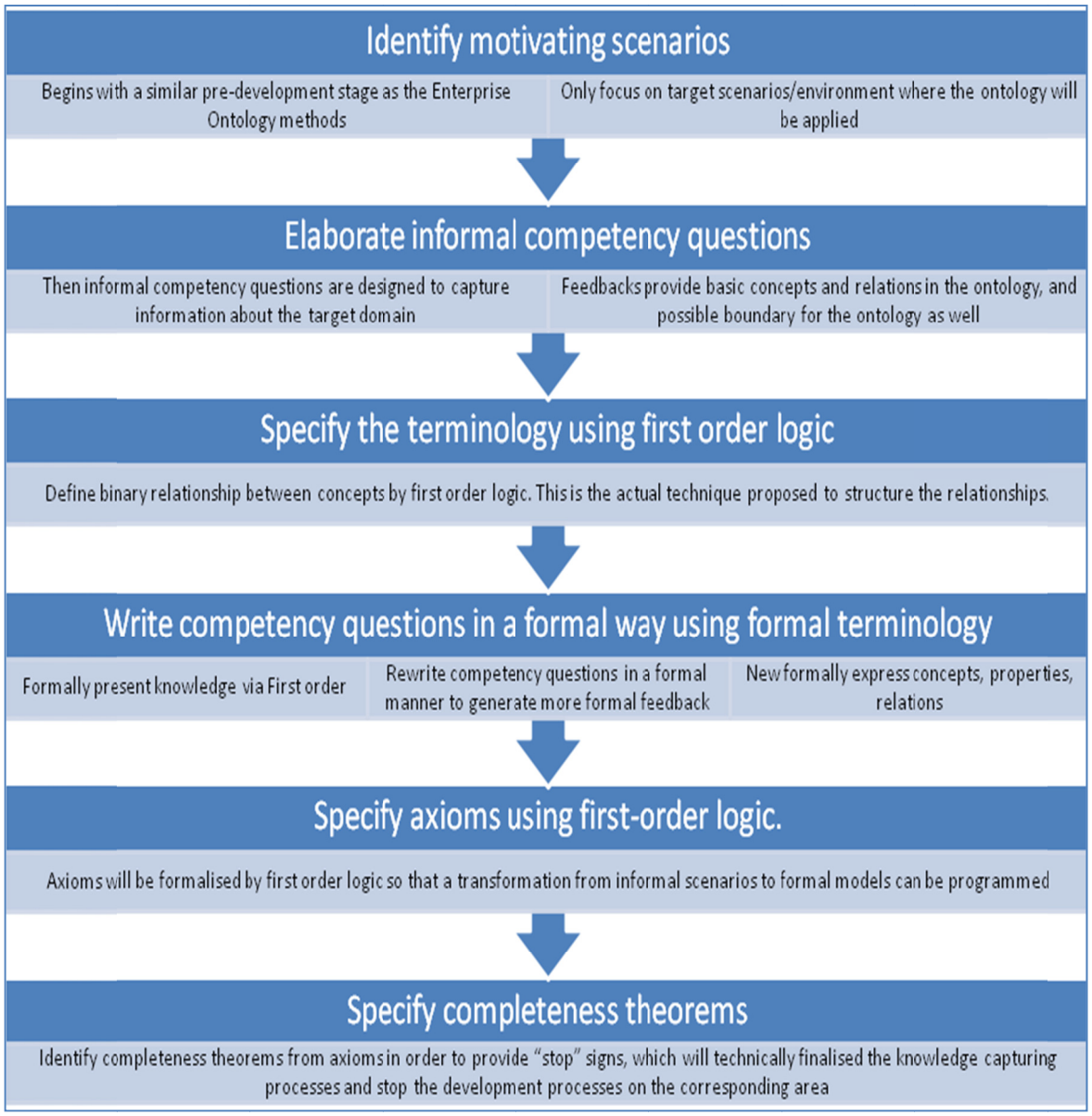


Figure 2.9: Approach to developing TOVE Ontology

This particular issue was structurally created by a linear process, and it required a change to develop a continuously improvable model. The "On-To-Knowledge" project (Staab, 2001) proposed a circular process to replace the linear approach.

2.4.3 On-To-Knowledge and KACTUS methodology

On-To-Knowledge's circular process (Figure 2.10) ensured that ontology can be continuously improved and updated as needed.

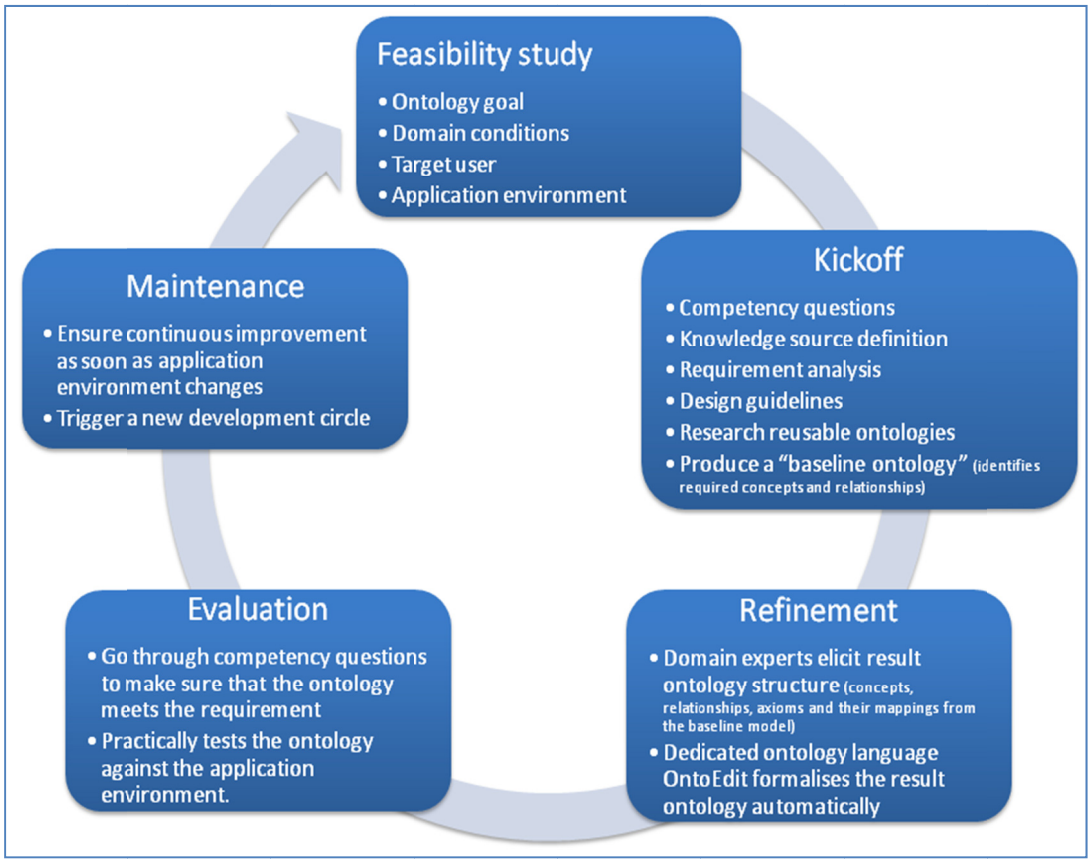


Figure 2.10: Approach to developing "On-To-Knowledge" Ontology

This methodology kept TOVE's pre-development stage; it also added connections to ontology evaluation and maintenance stages. These stages enabled errors in the ontology output, or even in the process, to be corrected in a systematic manner.

Additionally, On-To-Knowledge showed that ontological analysis and output generation can be automated. The OntoEdit tool (Sure et al., 2002) was utilised to

create ontology output automatically from the corpus constructed. With Cyc's proposal on automatic corpus construction, this implied that the whole ontology development process could be automated. This highlighted that researchers should try to automate the techniques involved in the ontology building processes to assist the development in order to minimise the requirement for domain experts.

Although not clearly stated, the circular process of On-To-Knowledge demonstrated a certain level of ontology re-use by improving the developed ontology from the kick-off stage again. This may be an approach to meet the "quick" and "economic" requirement for a good ontology. This reusability approach was further demonstrated in the KACTUS methodology.

The KACTUS approach (Schreiber et al., 1995) originally captures concepts and relationships from its application, and then uses them to represent the domain knowledge. When further development took place, this approach retrieved new knowledge in the same domain and under similar conditions. New knowledge benefited the ontology, since it provided more information for reuse, and offered more evidence for refining the existing version of the ontology. Many researchers realised the value of such a benefit (Bernaras et al., 1996, Fernández-López et al., 1997, Fernández-López et al., 1999, Gómez-Pérez, 1998). They integrated a formalised methodology with ontology reuse methods. One representatives of such an approach is METHONTOLOGY.

2.4.4 METHONTOLOGY methodology

METHONTOLOGY (Fernández-López et al., 1997) has proposed full development circles (pre-development, development and post-development) and technique details (supporting activities), and it was able to generate ontology from scratch, reusing existing ontology and re-engineering them (Figure 2.11).

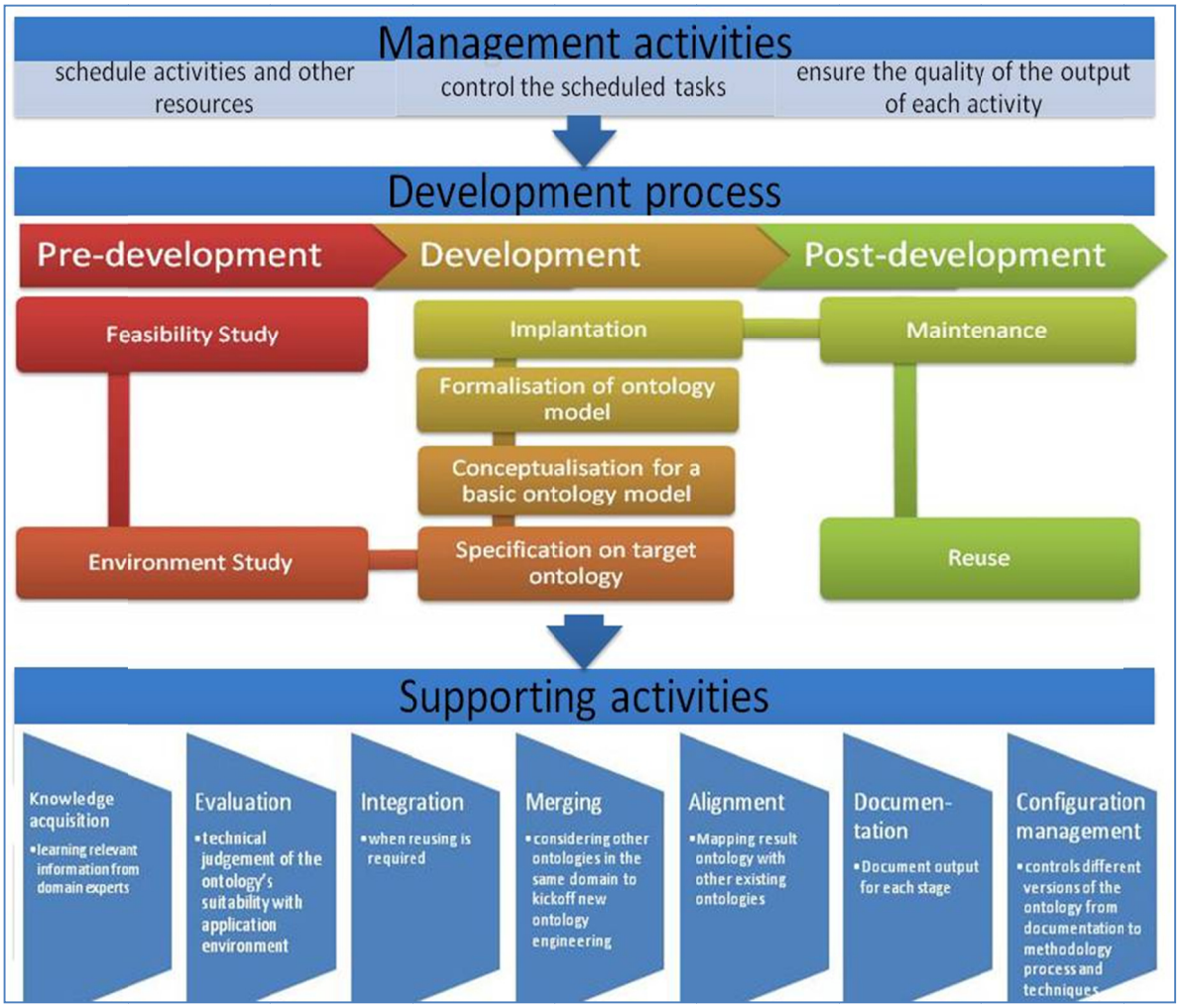


Figure 2.11: Approach to developing METHONTOLOGY ontology

Up to this approach, methodologies all kept Cyc's essential development stages: to construct the ontology corpus, to analyse ontological structure and to present output.

However, the METHONTOLOGY approach had such a detailed design that there were many more processes to execute. Such an approach cannot respond to practical changes as "quickly" as some applications require. The ontology required by the case study applications required ontology engineering methodologies to rapidly prototype the ontology as quickly as possible. This "rapid prototype" not only models the domain(s) required, but should also accept and incorporate new information.

Unfortunately METHONTOLOGY did not appear to have the flexibility to rapidly respond to changes within the domain. In contrast the SENSUS methodology appears to be able to quickly capture emerging information. This methodology differs significantly from the other examples.

2.4.5 SENSUS Methodology

SENSUS (Swartout et al., 1997) constructs ontology for a domain from the foundation of a large knowledge base, or ideally, a previous large ontology. However, it does not engage in a traditional reusing or re-engineering process. It identifies key domain specific terms, a.k.a. seeding words, and then links them to the large ontology. Afterwards, the terms irrelevant to the new ontology can be pruned from the large source ontology. The following processes should be undertaken in the SENSUS approach (Figure 2.12):

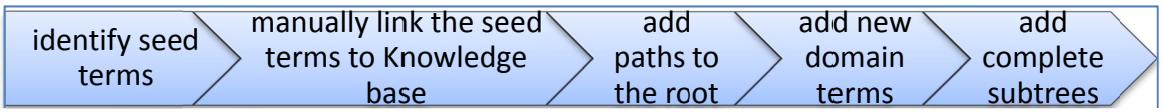


Figure 2.12: Approach to developing a SENSUS ontology

1. Identify seed terms (key domain terms)

   Keywords were obtained from domain experts, these keywords were treated as seeding words for the new ontology. SENSUS does not seem to state a pre-development stage (or application environment study), however the idea of using seeding words indicates that pre-development tasks have been conducted to extract keywords from experts.

2. Link manually the seed terms to SENSUS

   The selected domain keywords were linked to the existing SENSUS knowledge base or ontology structure. This process linked the seeding terms (reflecting the target subject areas) to a larger ontology source based on semantic relationship. This would extract a greater number of terms related to the seeding term, therefore constructing the ontology corpus.

3. Add paths to the root

   The route from the seeding words to the root of SENSUS (core concept of the target ontology) was identified, and the concepts and relationships were highlighted along this path. There might be more than one path from a seeding word to the root, and ideally the majority of the important routes need to be analysed. This process analysed the corpus constructed, and produced the structure of the ontology.

4. Add new domain terms

   Up to this stage, if there were still terms that should be included within the domain but had not yet appeared, domain experts again had to manually identify

them. Steps 2 and 3 should be repeated to capture additional concepts and relations.

5. Add complete subtrees

   It is interesting to note that some nodes might have a large number of paths through them in the new tree generated by the first four steps. This places them at the top of a new sub-tree. In such a case, if many of the nodes in this new sub-tree are relevant to the resulting ontology, it is likely that the whole sub-tree will be relevant to it. Therefore, this whole subtree should be added into the ontology.

This approach contains unique characterises that provide advantages over the other approaches discussed:

1. It is an obvious improvement that SENSUS no longer requires constant input from domain experts, and instead, only needs the initial seeding terms and their relationships to the knowledge base. This mechanism reduces costs in the information collection process from domain experts, and potentially accelerates this process by retrieving additional information from the different "routes to the root" generated.
2. A further distinguishing characteristic of SENSUS was its unconventional approach to mixing corpus construction with ontological analysis. Methodology examples from Cyc to METHONTOLOGY all collected terms and their relationship first; then ontological analysis was applied to clarify the internal

structure. However, SENSUS integrated such processes and focused on the semantic relationship between terms during the whole process, in order to ensure that the terms collected were semantically connected to the seeding terms. This characteristic reinforced that the output ontology was formed based on semantic relationships. Moreover, it maintained trustworthy terms and relationships to supplement the possible reliability loss because of its lesser reliance on domain experts.

3. An advantage of the SENSUS approach was that the development of different ontology shared the same knowledge bases and their internal links. Swartout has pointed out that the main advantage of SENSUS was that the massive coverage of the SENSUS ontology becomes a "hinge" that marries the terminology and the organization of other ontology developed that are based on it (Swartout et al., 1997). Therefore, SENSUS's source sharing is suitable for obtaining terms from non-experts, as well as for multiple domains.
4. Extracting related terms from the same sources by different seeding words is similar to perceiving the same knowledge from different perspectives. This in theory could result in fuzziness around any given concept. Thus the SENSUS ontology construction method may be capable of building cross domain ontology. This sets an exemplar for this research in establishing multi-disciplinary ontology.

The SENSUS methodology seems superior to the others in the ways discussed. However, it is difficult to reuse SENSUS directly, as there is insufficient detail on

the techniques suggested. In addition, SENSUS did not propose any post-development stage, a development life cycle and project management mechanism (detailed discussion in section 5.3). Therefore, this research used the SENSUS approach as a foundation approach and developed techniques to formulate a new methodology that met the needs for faster, more economical, reliable, multi-domain as set out in chapter one.

2.5 RESEARCH OBJECTIVES SPECIFICATION

Combining the research question and suggestions from the ontology field review, the assumption for specified research goal can be formed as:

> *"Can an ontology engineering methodology quickly, reliably and economically generate a multi-disciplinary ontology that can provide the breadth and depth of coverage required for automated systems"*

Clearly, such an overall assumption contains many stage of development, so it is better to break it down, to focus on one stage at a time. Hence, the overall assumption has been separated to target objective specifications according to the ontology development stages.

1. First of all, the author proposes a new approach that can produce a specified corpus covering the required domain(s). To build such a corpus, the new methodology should be able to extract semantically connected domain keywords from given source(s). This process should only need a few seeding words and

their relationships from domain experts, and automated processes could retrieve more related terms from the seeding terms and relationships.

2. This solution should supply sufficient description of the target domain(s) to form a corpus. At least, it should collect a larger amount of domain specialist terms than TMG and WMCCM. Moreover, it should produce associated semantically related terms from non-experts, or even from other domains.
3. After corpus construction, ontological analysis clarifies the ontological structure from the corpus and produces the ontology output. This requires rich internal relationship within the corpus constructed. The proposed solution should form a network type ontology structure. Ideally, relationship among concepts should be rich enough to be presented by numeric figures (weight specified) for clearer indication to aid fuzzy matching. Additionally, the structure needs to be able to be observed from different viewpoints, so it can provide the hierarchical output required by different ontology applications such as WMCCM and TMG.

Specification on research objectives proposed a detailed assumption of the characteristics of a new ontology building solution. Now, a practical stage of ontology engineering is required to test whether a new solution can meet the assumptions.

CHAPTER 3: ONTOLOGY ENGINEERING

The investigation into ontology engineering methods in the previous chapter revealed that a SENSUS like approach could theoretically provide speed in ontology generation with less reliance on domain experts. These are highly desirable traits, but can these be implemented practically? This chapter investigates techniques that can make this possible.

3.1 ONTOLOGY BUILDING PROCESS CONFIGURATION

The SENSUS methodology recommended that the ontology building should be initialised by linking specified keywords to the target source (Figure 3.1).

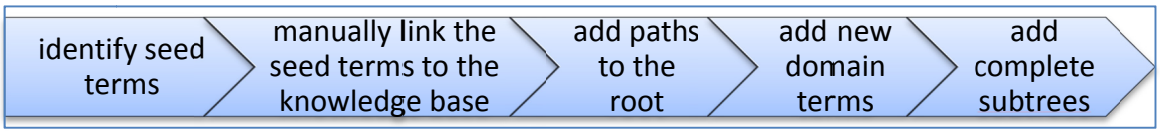


Figure 3.1: Approach to developing SENSUS ontology

3.1.1 Source selection

The source is where the domain(s) knowledge is contained. There are two main types of source that could be used. :

1. Directly collected expert and user data: first hand data; or
2. Directly reused or extracted data: data sources which contain words with linguistic connections or domain oriented semantic relationship regarding to

their content. For such a type of data source, the author identified five types of resources:

a. Thesaurus or dictionaries, representing a dictionary type

b. WordNet, representing general lexical ontology or databases

c. Industry/Government codes, representing domain ontology e.g. SIC

d. Ontology search engines, such as OntoSearch, OntoSelect and Swoogle, which represent searchable ontology databases that index lists or directories of ontology

e. Search Engine Index: Popular search engines focus on the full-text indexing of online, natural language documents. The author suggests the use of semantic relationship mined from search engine indexed pages to create the ontology corpus. This is a relatively novel idea, which has not been used to build new ontology as far as the author is aware. Current use of search engine indices in ontology engineering mainly concentrates on refining current ontology towards including fuzzy logic (Lau, 2007) and new knowledge acquisition (Agirre et al., 2000, Qi et al., 2009).

This research's requirement for less reliance on domain experts, broader coverage of concepts and richer internal relationship directs the source selection towards using ontology search engines or the search engine index: first hand data (type 1) usually requires significant input from domain expert; dictionaries (type 2.a) and lexicon ontology (type 2.b) do not usually provide enough domain focused relationships (discussed in section 2.3); and domain ontology (type 2.c) do not cover sufficient concepts and relationships (discussed in section 2.1 and 2.2).

Ontology search engines (type 2.d) seem capable of providing terms and relationships from different viewpoints as they retrieve information from a collection of existing ontology. However, testing the search results showed the following disadvantages.

- As with general search engines, the search results from ontology search engines are provided as links to content, which are ontology that have the query keywords. The result can be recognised as a faceted system, and therefore they face the issues that a faceted system may encounter.
- They do not necessary provide broad term coverage and have sufficient relationships (as discussed in section 2.2); and
- Identifying the relevant sources (facets) for a target ontology may require significant manual contribution from domain experts.
- With the sources selected, the internal relationships and concepts in them may differ or even be in conflict, thus a labour intensive ontology merging or alignment process may be necessary (such as UMLS' source network discussed in section 2.1).
- These search engine results are limited by the rigor within their concepts and relationships. The sources may not provide the semantic relationships required by this research.
- The wide fuzzy terms and relationship suggested by users may not be contained by these sources. For example, a combination of "boring" and "welding" (from

SIC code) does not produce any result from them (search conducted in Dec 2010 in OntoSearch, OntoSelect and Swoogle).

Thus a general search engine index, which crawls all types of web pages on the Internet, may better suit the need of this research for a broad coverage with rich relationships.

Increasingly the Internet is becoming the ultimate source of information in new or rapidly changing fields, and with its current rate of growth it could become the ultimate resource in most subject areas (Falagas et al., 2008). Information available on the Internet is retrieved by search engines and stored into their indices. In this work, the author chose to mine terms and their relationships from search engine indices. There are weaknesses in search engine indices - the content is often not verified sufficiently, and may well just disappear at a later date. The decision to use a search engine index was supported by a multi-criteria comparison across the sources above (details can be found in appendix 3.1).

There are many popular search engines available across the Internet, such as Google, Yahoo, and Bing. Among these search engines, Google has been widely regarded as the market leader search engine with the largest indexed content and popularity. Early as 2005, a study measured search engines results based on 438,141 queries in 75 different languages claimed that Google had the largest index size, and it also had the highest coverage of all search engines indices (Gulli and Signorini, 2005).

Kunder then tested this (Kunder, 2010) on a daily basis through using 50 selected queries (the queries varied everyday based on random selection from more than a million stored representative queries). It was found that Google still contained the largest index size in 2010 (around 19 billion pages), and the index was changing much more steadily compared with Yahoo and Bing. The steady change implies that Google's indexing mechanism may have functioned more effectively than the other two major search engines. Kunder's methodology found that the Yahoo index size overtook Google in 2009, but then shrunk dramatically towards the end of 2010 (less than 5 billion in the last quarter of 2010). It appears that Google has the greatest coverage of internet content among these search engines, and it is more reliable in terms of index size changes.

Additionally, Google is also able to produce domain focused terms with semantic relationships. Traditionally, search engines such as Google, Yahoo, and MSN provide their domain focused terms mainly via two channels: their directories (business classification) or search keywords selection tools. The directories are similar to the industrial categorisation discussed (such as SIC), which may suffer from insufficient coverage and limited business oriented relationships. As compensation for their directories' limited relationships, search engines also provide keywords selection tools which list relevant search terms based on previous searches made by internet users (similar to the folksonomies discussed in section 2.2.2). A Redmond report states that Google has 65 percent search query share in the U.S. plus over 85 percent

in Europe. Bing and Yahoo together have had roughly 30 percent in U.S. and only 10 percent in Europe (REDMOND, 2008). Thus Google has more opportunity to analyse the input from internet users to provide relevant keyword search.

However, relying on user input information is similar to absorbing users' tags to form folksonomies, and thus also inherits the drawbacks identified for folksonomies: accepting any relationship from any perspective undermines the domain focus ability. Uniquely, Google provides a method – Google Sets (Tong and Dean, 2008) - to generate "on-topic" terms based on given examples. This new method seems to provide an opportunity to generate domain related terms with wider but not chaotic relationships (discussion on Google Sets in section 3.1.2-3.1.4).

The Google search index was chosen as the key source, due to its wide and up to date coverage of all types of information, and its ability to provide domain focused information from different perspectives.

3.1.2 Seeding words Selection

Seeding words for this research were produced from both ontology builders and domain experts (Figure 3.2). It was expected ontology builders could contribute from application specification of terms, and the domain experts may strengthen the terms' domain representativeness in general.

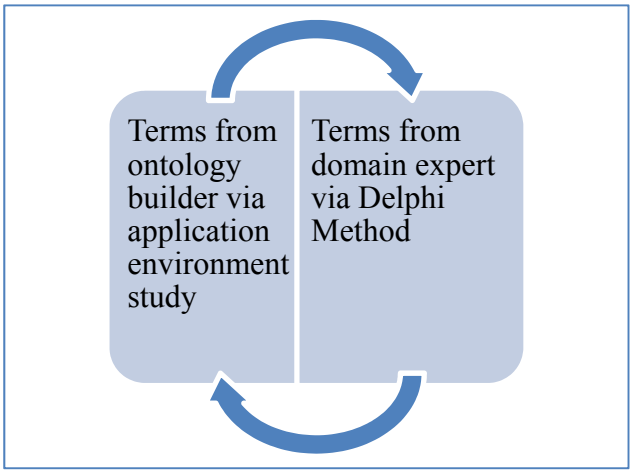


Figure 3.2: Illustration of seeding words selection

1. Seeding words from ontology builders via application environment study. In this research, studies on exiting ontology and their application environment (such as WMCCM and TMG) were used as a method to extract representative terms in corresponding domains. However, this choice may be subjective, due to the awareness of the application (Benbasat et al., 1987). Thus studies on other ontology (in the fields but not used in the same application environment with TMG and WMCCM) were also conducted to extract the pattern of terms from different sources/perspectives. Additionally, ontology builders are allowed to make modifications to the experts' proposed seeding terms, in case those terms do not reflect the correct application environment.
2. Besides extended ontology case studies, domain experts were also asked to provide their domain ontology seeding words or optimise ontology builder's proposed seeding words. Domain experts' suggested seeding words can be linked with the Google index to extract more domain related terms (concepts on the

routes from the seeding words to the knowledge base). By evaluating the newly extracted terms, domain experts may modify the seeding words in order to gain alternative domain related terms that may fit in better with the application environment.

Collecting seeding words in a subject area from domain experts based on their knowledge on the domain is also known as "Delphi Method" (Linstone & Turoff, 2002). Delphi method collects the opinions of different individuals, in order to increase the opportunity of picking objective seeding words and minimize subjective bias from directly study on the application environment.

A combination of application specified and domain representative seeding words can be produced via a consideration of both ontology builders and domain experts' proposal. The awareness of expected relationships between the nominated terms (in this research – semantic relationship) was also raised during the seeding words' selection process.

3.1.3 Semantic Relationship

SENSUS has suggested that seeding words and the source could be connected by semantic relationship. In other words, any seeding word should be connected to its semantically associated terms in the source. These terms are group(s) of concepts representing similar domain concepts to the seeding words. Techniques which provide grouped domain concepts are:

- Categorisation: “a method provides groups of entities whose members are, in some way, similar to each other”. It concentrates on “concept formation and coverage” and allows overlapping (Jacob, 2004)
- Classification (including taxonomy): “a method involves the orderly and systematic assignment of each entity to one class within a system”. It highlights “only one class and no overlapping” (Jacob, 2004), and emphasizes “delimiting and distinguish” (Mayr, 1982)

Categorisation better meets the research purpose, as allowing overlap can create keywords groups to maximize coverage over target subject areas. Focusing on overlapping coverage allows fuzzy concepts which link the terminology in the concept to other concepts in the domain or to other domains and also importantly to the non-specialist language in a domain. Classification and Taxonomy type relationships would leave gaps among concepts, since they highlighted uniqueness, no tolerance of overlapping, and delimitation.

Within categorisation techniques, a method called “Word Clustering” directly utilises “co-appearing in content” forming the semantic relationship between terms. Word clustering processes sets of words into categories if they are “semantically similar words”. Two different types of word similarity have been used in word clustering:

- Semantic similarity: two words that are paradigmatically similar (thesaurus), and substitutable in a particular context. For example, “I ate sausages for breakfast”, the word sausages can be substituted by “bacon” with little change to the

meaning and structure of the sentence, and therefore these two words can be identified as being semantically similar;

- Semantic relatedness: two words that often occur simultaneously in a text. For instance, fire and burn are semantically related, since they often appear together within the same context (ILC-CNR, 2007).

This research focuses on semantic relatedness rather than semantic similarity. This is because keywords representing the same concept are more likely to co-occur in sentences, but are not necessarily substitutable with each other. The emphasis on semantic relatedness means that thesaurus dictionaries and WordNet may not provide sufficient wider relationships with semantic relatedness between terms. Hence they are not suitable as a source knowledge base.

In addition, from a coverage perspective, semantically related relationships may also contain semantically similar relationships: substitutable words may co-appear in the same content (such as "burn" and "ruin"); but co-appearing words may not be substitutable (such as "burn" and "fire"). Word clustering using semantic relatedness may provide the desired binary relationships between terms for ontology.

The Google Sets tool could link the seeding words to the Google index via semantic relationships, since it is a word clustering tool which extracts semantically associated words from the Google index. However, it is not clear which type of word clustering is applied by Google Sets. In order to validate whether the output terms of Google Sets are semantically related or similar, an experiment was created. Ten words were

paired up with their synonyms (according to Oxford Synonyms Dictionary) as seeding words, and then fed to Google Sets. The ten words with their synonym in brackets are: big (*large*), strong (*powerful*), humanity (*the human race*), bet (*gamble*), hit (*strike*), study (*learn*), chat (*talk*), hard (*difficult*), near (*close*) and rich (*wealthy*).

Eight out of ten test sets generated purely semantically related keywords (no synonyms found). The other two test sets both found one synonym (hit found *strike*; hard found *difficult*) within their predictions. It appears that the majority of Google Sets' results are generated by semantic relatedness word clustering. Further test on the two predicted synonyms (*strike* and *difficult* were fed into Google Sets) did not bring back their original seeding words (hit and hard), and later experiment conducted on Google Sets (in section 3.1.4) also did not produce substitutable words to the seeding words. Synonyms not being able to predict each other indicated that Google Sets is based upon semantic relatedness word clustering.

The minority of the results (10%) which seemed to be based on semantic similarity are debatable. Because the two types of similarity are not mutually exclusive they do have overlaps. For example, "*hit*" and "*strike*" are synonyms but they can also be semantically related which is why these two words show in the Google Sets results. Thus Google Sets is a semantic relatedness based tool which can link seeding words to the Google index knowledgebase.

Google Sets (Figure 3.3) has several parameters that can be altered through the Google Sets settings, and the effects of varying these on the semantically related

words generated were not yet clear. This required a study of the Google Sets parameters so that they could be configured to provide the best results.

Figure 3.3: The Google Sets platform (Google, 2009)

3.1.4 Experiment Environment Configuration

Google Sets (Figure 3.3) has two options to control the volume of results: large sets and small sets, and accepts a maximum of five terms as seeding words.

3.1.4.1 Set Size configuration

Literally, "Large Set" generates more terms than "Small Set", which makes a "large set" selection more suitable for expanding the coverage of semantic similarity terms. However, if "Small Set" provides different terms and relationships regularly rather

than repeated subsets of terms from a large set, both small sets and large sets should be considered.

Table 3.1: The feedback from small set and large set (Google, 2009)

Small Sets Predictions	**Large Sets Predictions**		
cosmetic surgery	cosmetic surgery	abdominoplasty	cosmetic surgeons
breast enlargement	breast enlargement	blepharoplasty	eyelid lift
contact lenses		body	eyelid surgery
cosmetic breast surgery	cosmetic breast surgery	body contouring	face lift
cosmetic eye surgery	cosmetic eye surgery	body sculpting	Facelift
cosmetic plastic surgery	cosmetic plastic surgery	botox	facial plastic surgery
facial cosmetic surgery	facial cosmetic surgery	breast augmentation	facial rejuvenation
health care		breast enhancement	laser hair removal
laser eye surgery	laser eye surgery	breast implants	laser resurfacing
laser surgery	laser surgery	breast lift	nose surgery
Liposuction	liposuction	breast reduction	Plastic
patient education	patient education	breast surgery	plastic surgeon
tummy tucks	tummy tuck	certified plastic surgeon	plastic surgery
Vitamins		chemical peels	Rhinoplasty
weight loss surgery	weight loss surgery	cosmetic dentistry	ultrasonic liposuction
		collagen	tummy tucks
		cosmetic laser surgery	cosmetic surgeon
		Surgery	

Experiments were conducted in both engineering and medical domains to test whether "small sets" generates relations that do not exist in large sets. For instance, "*Cosmetic Surgery*" was chosen as a seeding word for the medical domain. A large

set configuration brought back 47 predictions (column 2, 3 and 4 in Table 3.1 with column 2 highlighting matching predictions with small set predictions), while the small set brought back 15 (column 1 in Table 3.1). Among those predictions, 12 of the terms (highlighted in green in Table 3.1) appeared in both the large set and the small set, and were semantically related to "*cosmetic surgery*". It was also observed that three terms (highlighted in red in Table 3.1) appeared to be unique terms brought back by the small set setting.

Table 3.2: A Comparison between the Results from Large and Small Sets

Seeding Word	large Set Predicts	Small Set Predicts	Predicts in Small Set Only	Small set only predicts proportion
breast enlargement	48	15	3	6%
cosmetic breast surgery	7	5	0	0%
cosmetic eye surgery	7	5	0	0%
cosmetic plastic surgery	7	5	0	0%
facial cosmetic surgery	7	5	0	0%
laser eye surgery	48	15	2	4%
laser surgery	44	12	3	7%
Liposuction	50	15	2	4%
patient education	14	8	0	0%
tummy tuck	49	15	0	0%
weight loss surgery	47	15	2	4%
cosmetic surgery	49	15	3	6%

The 12 terms appearing in both results were further fed as keywords back into Google Sets to compare their feedback differences for different volume settings. It was expected that Google Sets would produce "small set" as subset of "large set" for

these seeding words, and small proportion of "small sets only" terms would appear. The result (Table 3.2, details can be found in appendix 3.3) indicates the following characteristics of the volume settings:

- Generally, a large set can generate many more predictions than a small set. It covered more than 90% predicts in the shown example. This indicates that a large set prediction could almost cover the corresponding small set prediction. Under such circumstances, the experiment may only require large set configuration.
- A small set was not a definite subset of a large set, although there were six of the twelve large sets results fully included their corresponding small set results. Small sets and large sets seemed to elicit semantic related words slightly differently.
- A small set may sometimes provide "small set only" predictions (terms that do not exist in the large set for the same seeding terms), such as "contact lenses", "healthcare" and "vitamin" for "cosmetic surgery", but the proportion of the predictions was very small. This small proportion of "small set only" prediction may not be strong enough to yield any statistical differences on the prediction if the experiment was run more than once or started with more seeding words.

These characteristics showed that the large set had advantages in coverage, since it included more than 90% predicts that occurred in small set on average. The large set

generated more fuzziness, but was still constrained by "semantic relatedness". This research chose the large set setting to conduct further experiments.

Apart from the set size options, Google Sets also required configuration of the input terms. Google Sets accepts a maximum of five terms as seeding words, and each of these can be phrases, rather than absolute single words. Optimal configuration of seeding words required an understanding of the following:

- Number of seeding words: what differences in the results will be seen by changing the number of seeding words; is there an "ideal" number of seeding words?
- Domain sensitivity of seeding words: how will the selected seeding words affect the result in terms of domain focus? Is there a method to "lead" the result to be domain focused?
- Order of seeding words: how do the result change according to changes in the sequence of seeding words? Is there a certain order in which to feed them?

3.1.4.2 Seeding words number configuration

The number of seeding words may affect the number of predictions (quantity) and the domain focus level (quality).

3.1.4.2.1 Impact on Quantity of predictions

The WMCCM manufacturing ontology was chosen as a seeding words source to test the impact on the number of predictions. Five terms in ten categories were collected via the combination of ontology builders and domain experts. They were then run through the large set configuration five times (single keywords, two keywords, three keywords, four keywords and five keywords) to test the predictions.

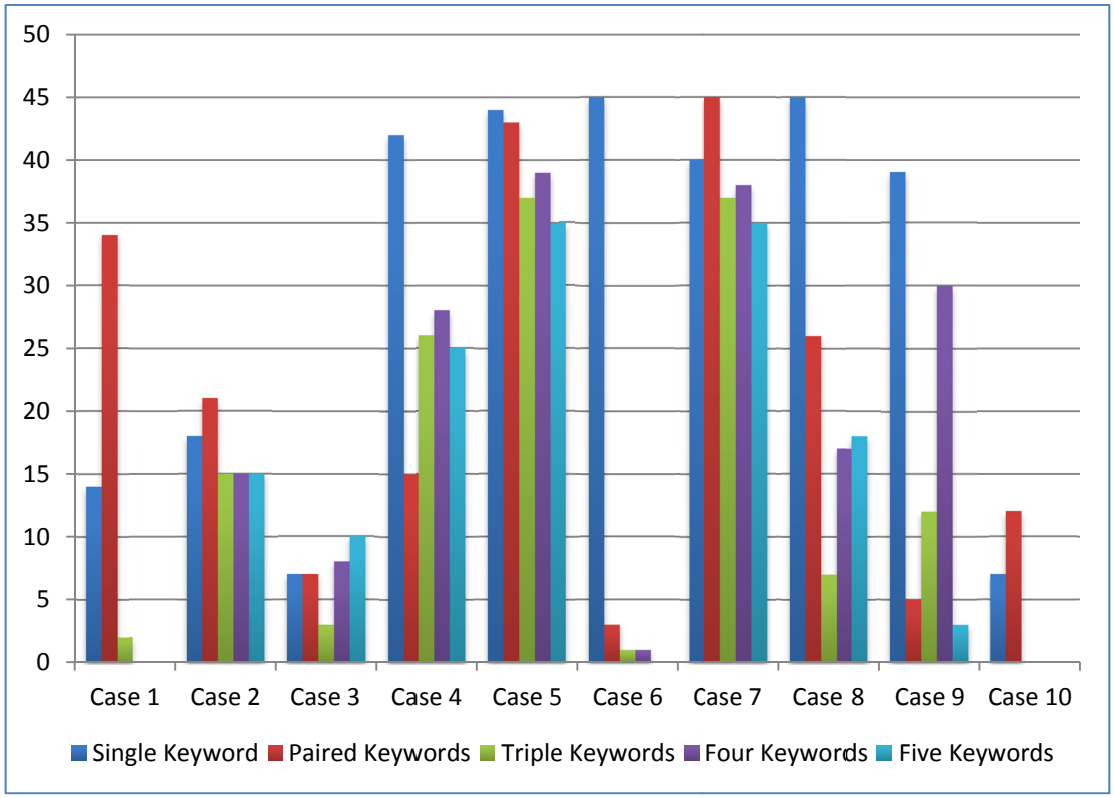


Figure 3.4: Relationship between the number of predictions and the number of inputs

Shown in Figure 3.4 (Bar height demonstrated the number of predictions generated as Y-axis marked), the experimental result revealed the following findings:

- Most cases suggested that the number of predictions was inversely proportional to the number of input. The more seeding words provided, the more constraints were put onto semantically related predictions. As a result, there were fewer predictions which met such restricted requirement.
- A large decrease in the number of predictions was observed when more than two seeding words were involved, particularly in cases 1, 6, and 10. Each seeding word seemed to add new constraints to the search area that the Google Sets result list had to meet. The reduction in predictions suggested that the more constraints (seeding words) Google Sets had, the fewer predictions could be provided.
- Using one seeding word or two seeding words settings produced many more predictions in the test cases than did the 3-5 seeding words settings.

It seems that single seeding word and paired seeding words are the optimum setting in order to produce a greater number of semantic related term predictions. However, a larger number of predictions does not necessarily denote a greater focus on the expected subject areas. Thus, further experiments were conducted to reveal the impact on the prediction quality of the number of seeding words from the domain focus perspective.

3.1.4.2.2 Impact on Quality of predictions

Single seeding and paired seeding words were used to examine the "quality" of the predictions. "Quality" in this experiment meant the level of domain focus, and also

the ability to generate clear statistical data on the importance of different predicted terms.

Using seeding words to obtain semantically related predictions was the same as seeking out contacts from "seeding" persons in a network. For instance:

- If an individual A shared its contacts, these contacts could well include family members, friends, colleagues from work, and classmates from school, etc. Hence, the "predictions" (contacts) were spread in all aspects of the "Seed" (the person who shares the contacts) social network.
- If the same individual A and another person B from A's contacts were to share mutual contacts, these contacts would probably come from a certain sub social network. If A and B were colleagues at work, their mutual contacts would be more likely to be other people who work with them. In such cases, a limited number of common contacts within a sub network of colleagues (compared with A's network including colleagues, classmates and families) were produced because of the network constraint B provided – a commonality (people from work) with A (Carrington et al., 2005).

Therefore, in terms of using seeding words to obtain their shared "contacts" (predictions), it is expected that smaller number of terms in a focused domain would be produced, when paired seeding words were used to constrain their representing domain. Single seeding words may produce terms with less restricted relationships in domains. Paired seeding words were expected to perform better than single seeding

words in producing domain focused predictions (demonstrated in Figure 3.5). However, this expectation required validation as Google Sets had not been tested for domain focus.

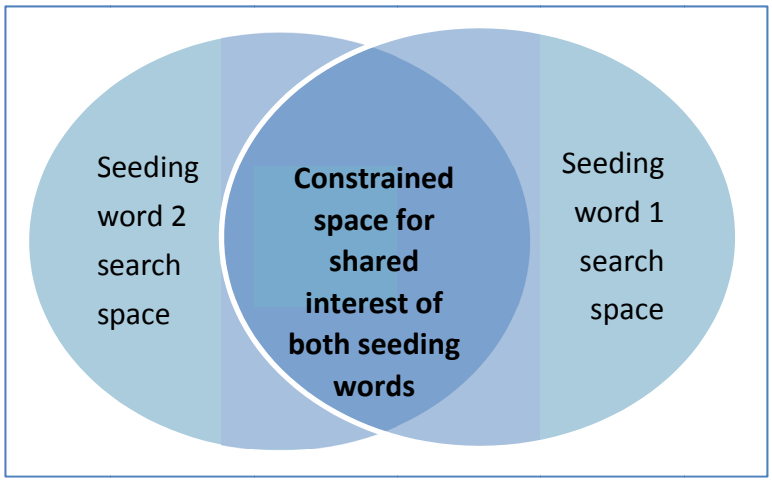


Figure 3.5: Illustration of paired seeding words search space

The purpose of the "Impact on prediction quality" experiment was to validate if paired seeding words would improve the domain focus compared with a single seeding word. Thus a generalised conclusion could be drawn towards a preferred seeding words number for Google Sets configuration. Three medical terms were selected as seeding terms from general terms to domain specific terms:

- Dental treatment: a high level medical concept, it is also likely to appear in general discussion outside of medical domain.
- Acute Glaucoma: a healthcare related term but not a popular general term, may appear in general discussions
- MRSA: a type of bacteria which only appears in a medical related context

Google Sets was used to generate semantically related terms. The resulting terms were too few to represent the corresponding domain or to yield any statistically relevant results. In order to generate more relevant keywords for the domain, the original resulting terms were applied as seeding words again to obtain yet more predicted terms. After this second round experiment, there was greater coverage of the domain, but there were still insufficient concepts and relationships to yield any statistical trend. Therefore the terms from the second round were used as seeding words to generate third level predictions.

Such an approach is common in social studies and statistics, especially within social network analysis (Salganik, 2003). "Snowball Sampling" as it is known to use the same approach to generate a large collection of related data to construct complex social network (Carrington et al., 2005). It is capable of producing a statistically meaningful distribution from unclear network structure. In addition, there are associated social network analysis techniques to discover more facts about such a network.

"Snowball sampling" the seeding words not only generated a network of terms, but also yielded frequency of occurrence data for these terms (statistical data was created). It was found that some of the predictions appeared quite often, and some of them appearing only once. To simplify the analysis, those predictions that appeared only once in the result were defined as "long tail predictions", with the remaining terms being defined as core and relational predictions.

The same seeding words were also used to start off the paired seeding words experiment: The first round experiment was the same as the single word experiment in terms of obtaining predictions. From the second round onwards, instead of feeding only one word as a seeding word, predictions from the previous round were paired in all cases, to feed into Google Sets. For instance, if "a" were the seeding word, and generated "b" and "c" as the first round predictions, the second round experiment would take three pairs of keywords as the feeding words these being – a and b, a and c, b and c.

Table 3.3: Results from the Single word experiment

Seeding words	Total number of prediction terms	Number of unique predictions	Average nomination per term	Number of long tail predictions	Proportion of long tail
dental treatment	1088	549	1.98	364	66.30%
acute glaucoma	941	381	2.47	243	63.78%
MRSA	1450	501	2.89	288	57.49%

Table 3.4: Results from the Paired words Experiment

Seeding words	Total number of prediction terms	Number of unique predictions	Average nomination per term	Number of long tail predictions	Proportion of long tail
dental treatment	584	87	6.71	7	8.05%
acute glaucoma	760	156	4.87	51	32.69%
MRSA	504	163	3.09	84	51.53%

Table 3.3 and 3.4 (primary data and analysis in appendix 3.3) showed the results after three rounds of experiments with single seeding word and paired seeding words. In these tables:

- "Total number of prediction terms" was the total of predictions regardless of predicted term repetition. It was also the number of binary relationships found in the experiments, since every prediction was found via a semantic relatedness relationship from its seeding term(s).
- The "Number of unique predictions" treated repetitive predictions as the same term.
- The "Average nomination per term" was calculated as the quotient of "Total number of prediction terms" and "Number of unique predictions". It is the average nominations any prediction had from other members in the network. The more nominations any member has, the more they may be recognised as a representative of this network (Gjoka et al., 2010). Thus, the average nomination number measures the domain focus level.
- The "number of long tail predictions" was the sum of those terms that appeared only once in the network. Being recognised by only one fellow member, a long tail prediction was regarded as a non-representative (outlier) term of the domain. Hence, the more long tail members a network had, the less representative (focused) the network was.

- The “Proportion of long tail members" could thus be understood as the level of chaos that reduced the domain focus. A lower proportion of long tail members in a network implies a better focused network. Long tail members however could also be regarded as the “strings” that link a domain with another domain, and which maybe key when trying to build multi domain ontology.

Compared to the single word configuration, the result for the paired seeding words showed some differences:

1. Paired seeding words had the advantage of producing a more focused domain terms network from both average nomination and long tail proportion perspectives.
    - By pairing up the seeding words, the dental treatment average nomination tripled from 1.98 to 6.71; the acute glaucoma domain focus doubled to 4.87 from 2.47; the MRSA domain focused increased from 2.89 to 3.09. The paired seeding words managed to raise the focus level for all test cases regardless of the expected natural focus of the seeding words.
    - In addition, pairing up seeding words reduced the number of long tail members in all experiment groups. In particular the “Dental Treatment” had only 8% long tail members after applying the paired seeding words (it had 66% long tail members in the single word configuration). Fewer long tail members indicate that the paired seeding words configurations generated a better

network than the single seeding word in terms of a better focus on valid members.

2. It seems that paired seeding words particularly benefitted the domain concentration for the less focused domain, in this case the "dental treatment" network. The "dental treatment" network was expected to be the least precise network due to the popularity of the phrase in a wider context. The single word experiments validated the fact that the "dental treatment" network was the least precise among the three – with 1.98 average nomination and 66.30% long tail predictions. A further experiment based on paired seeding words pushed the average nomination to 6.71 which was the highest among the three, with only 8.05% long tail members. Observation on the actual content of the predictions revealed that single seeding word brought back many non-treatment but wider medically related terms, such as "maternity", "specialist outpatient care", "parental accommodation" and "pre existing conditions". When these terms are further applied as single seeding words, they yielded to their most related domains and produced any-treatment related predictions. On the other hand, if they were paired with treatment terms such as "dental treatment", these constrained the resulting terms, so that the paired seeding words produced commonly related terms that were treatment related.
3. The paired seeding words improved domain concentration for the more naturally focused domains in a more limited way. In the experiment, the MRSA focus level only increased from 2.89 to 3.09, and the proportion of long tail members

reduced by just 6%. This is possibly because MRSA would naturally only lead to an increasingly focused domain.

An extreme case of the "Cosmetic breast surgery" network could provide more insight for such a scenario. "Cosmetic breast surgery" produced "cosmetic eye surgery", "cosmetic plastic surgery", "facial cosmetic surgery" and "cosmetic surgery". In this network, except for "cosmetic surgery", the remaining four seeding words only produced these same five predictions in both single seeding word and paired seeding words settings. This showed that these four terms naturally predicted to a restricted domain (inner circle in Figure 3.6), and such domain focus would remain almost unchanged regardless of the seeding words number configuration. Therefore, a naturally highly focused domain may not experience as much domain focus improvement as a less focused domain through pairing seeding words.

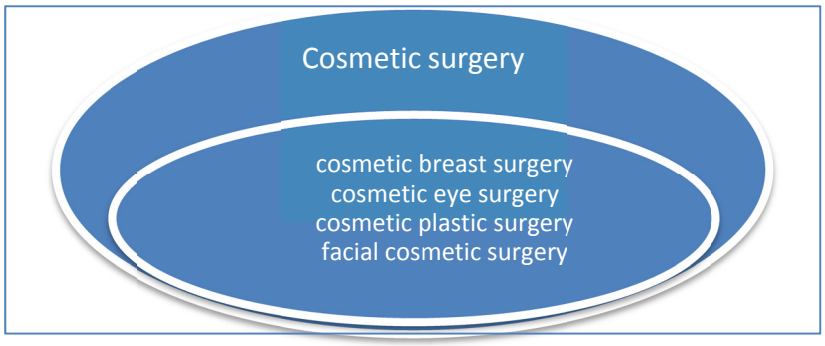


Figure 3.6: Cosmetic breast surgery network domain focus

In addition, both single seeding word and paired seeding words experiments showed that they could yield to a statistical trend on the distribution of the predictions (appendix 3.3) based on at least three rounds of operation. As Figure 3.7

demonstrated, conceptually, one round of experiment may only show a collection of terms without specifying their differences in terms of their appearance, relationships between each other and positions in the network. With two rounds of experiments, some terms may repetitively appear in the collection which may highlight their importance. The third round experiment acts as an "amplifier" to clarify the differences between different terms. It may be unnecessary for some domains if two rounds experiments could already form a clear difference between terms.

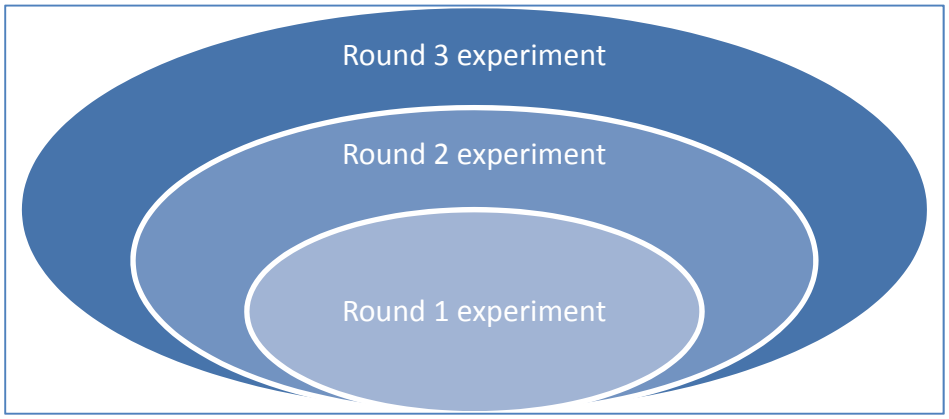


Figure 3.7: Illustration of Multiple rounds of experiment

Applying three rounds of experiment on both single seeding words and paired seeding words also confirmed that paired seeding words configurations provided clear trends on the differences on terms appearance, relationships and positioning in the network. Instead, the single seeding word experiment was only able to produce relatively flat variations. Clearer differences were much more valuable in further ontological analysis than the flat distribution. This also suggested that paired seeding words produced better result from statistical analysis perspective.

These results suggest that paired seeding words are superior to single seeding words in terms of the quality of domain focus and in the amount of data generated for further analysis. Therefore, paired seeding words are the seeding words' configuration for further experiments.

Having chosen to use paired seeding words, the obvious further question is, does the order of the paired seeding words make a difference?

3.1.4.3 Order configuration of seeding words

In order to test the seeding words order sensitivity, the same pair of seeding words had to be entered as seeding words in a switched positions. It was expected that the order of the seeding words should have little impact on the terms generated.

Table 3.5: Seeding words of the order experiment

Seeding words domain	Order 1	Order 2
Medical	Cardiac Surgery & Cosmetic Surgery	Cosmetic Surgery & Cardiac Surgery
Engineering	Welding & Engraving	Engraving & Welding
IT	Web design & Domain Name	Domain name & Web design
Fruit	Apple & banana	Banana & apple
Colour	Green & red	Red & green
Premier League Clubs	Leeds United & Manchester United	Manchester United & Leeds United
Luxury Brand	Armani & Versace	Versace & Armani
UK Politicians	Gordon Brown & Tony Blair	Tony Blair & Gordon Grown

For this experiment, in order to validate the generalisability of the order configuration, a wide range of paired seeding words (from domains besides medical

and engineering) were selected to test the order configuration of the seeding words. For instance, eight pairs of seeding words (Table 3.5) were fed into Google Sets in both orders (order 1 and order 2) to discover whether a difference in predictions would appear because of the order input difference.

All the test groups arrived at the same result with there being identical predictions for the paired keywords in either order. Such a result indicated that Google Sets was order insensitive. This experiment also validated that Google Sets was word clustering "semantic relatedness" based because the order of the words should not and did not affect the co-occurrence of words.

However, could one pair of seeding words always produce predictions in the target domain? Is it necessary to include more pairs of seeding words? More configurations test were conducted to find a starting number of paired seeding words

3.1.4.4 Starting number of seeding words pairs

A further concern suggested by number of seeding words was how to avoid seeding words with high potential for misleading the search direction. From section 3.1.4.2, it was clear that the term predictions from Google Sets were sensitive to the domain focus level of the seeding words. The selected seeding terms should be highly representative among the natural language domain terms, and also more remotely related to other misleading meanings from the target domain.

To maintain a satisfactory level of prediction, expert consultancy may be seen as a requirement, so that seeding words selection can be steered towards a better focus level. However, relying heavily on domain experts was one of the key disadvantages within traditional ontology engineering methodologies, and a major drawback that this research sought to overcome. Even where experts are heavily involved, predictions can still be vulnerable, since experts may make mistakes or have a particular viewpoint on the subject area. Hence, the domain sensitivity configuration of the seeding words aimed to reduce the contribution required from experts and to tolerate faulty seeding words, if possible.

Increasing the number of starting seeding words may lead to a better fault tolerance ability to accommodate human error in the choice of seeding words. Thus it is desirable to identify the minimum number of required seeding word pairs in order to enable fault tolerance. This is based on the assumption that one pair of seeding words might not represent the target domain as expected, but mislead the predictions to other directions:

- One pair of seeding words: If the selection of seeding words was wrong or they were relevant to multiple domains, it would cause a failure to generate domain predictions. Figure 3.8 showed that pairing the seeding words C and F may lead to two different domains (D concentrated domain and G concentrated domain). Such misleading was found in practice: for example in the context of “fruit”, an experiment on “apple” and “blackberry” did not bring back more types of fruit,

but produced mobile phone related keywords. Therefore, one pair of seeding words could quickly steer the search into unintended domains.

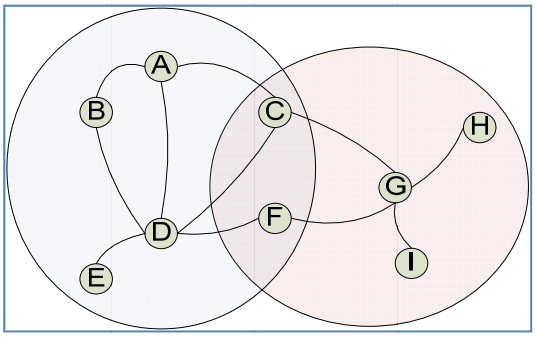


Figure 3.8: Misleading paired seeding words

- Two pairs of seeding words: If one pair of seed words did not represent the target domain; another pair may still be able to generate domain focused predictions. However, two pairs of seed words may produce predictions around two subject areas. In an extreme case (Figure 3.9), if a pair did not produce any target domain prediction at all, the experiment may end up with two separate distributions of terms, with no overlap. In such a case, the resulting corpus of terms may not target any particular domain, and further expert guidance may be required.

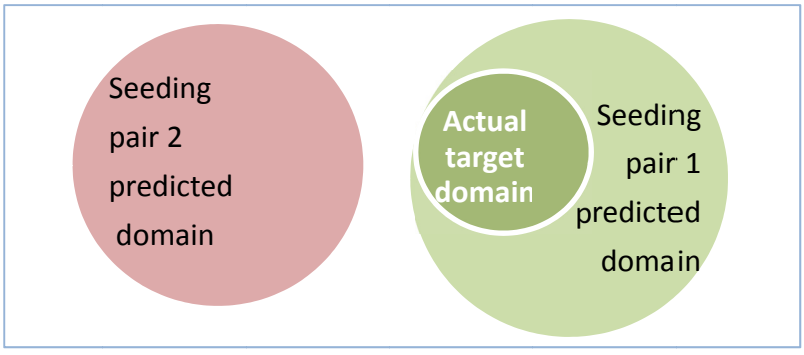


Figure 3.9: Complete prediction separation of two pairs of seeding words

- Three pairs of seeding words: Even if one pair of keywords represented a completely irrelevant domain against the target domain, two thirds of the

predictions would still lie in the same domain. Overlapping predictions could add up to a much higher appearance frequency than those misleading predictions (illustrated in Figure 3.10). The expected predictions would have a statistical advantage. In such a scenario, an automated generation can still be valuable, and extra expert consultancy may no longer be required.

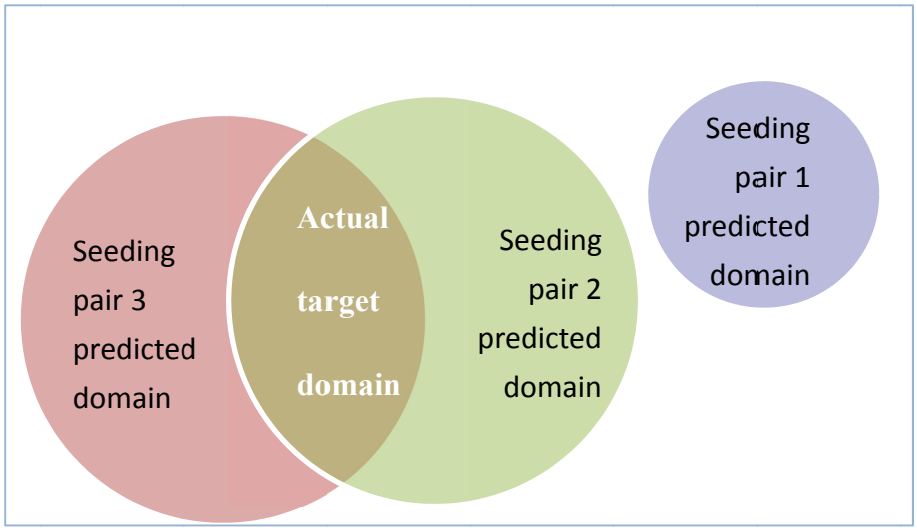


Figure 3.10: Illustration of one fault pair in three pair seeding terms scenario

The domain sensitivity experiments demonstrated that the expert consultancy may still be required, but this can be reduced to a minimum level. Choosing three pairs of seeding word, one pair can be allowed to be misleading. Even if the experts made a mistake in the suggested seeding words at the very early stages of the methodology and provided a faulty pair of seeding words, the system could tolerate this.

Choosing to start with three pairs of seeding words decided the final parameter in the Google Sets settings. These settings optimised the variables in the Google Sets settings for ontology generation. However manually conducting the experiment is

not really feasible with the huge number of terms that may be generated, so a software algorithm was devised to automate the ontology generation. The automation is discussed in section 3.2.

3.1.4.5 Experiment Configuration Summary

To conclude, the larger scale experiment should be conducted in the following manner:

- Begin with a minimum of three pairs of seeding words;
- "Large set" setting should be applied in the experiments;
- At least three rounds of experiments should be carried out, to reduce the consequences of an error in the initial keywords;
- During these experiments, seeding words can be fed in any order.

This configuration executes the first two processes in the new methodology. It also partially performs a third process: it found paths from the seeding words to the knowledge base by extracting the related terms from Google Sets.

However, there are no defined "root" members, since the knowledge base (Google index) did not provide any ontological structure for a given domain. Under such conditions, ontology could not be built and analysed around the target domain concepts and provide an ontological structure to represent the target domain. Therefore, this research has to construct the ontology corpus to identify domain representative concepts ("root" members).

3.2 CORPUS CONSTRUCTION

Corpus construction was enabled by conducting the snowball sampling on a larger scale, so that a large number of predictions in the given domain(s) could be captured to form the ontology corpus. Experiment configuration indicated that at least three rounds of snowballing sampling on three pairs of keywords are required in order to yield statistical meaningful data for further ontological structure analysis. This process is coded into mathematical algorithms to enable computer aided automated processing.

3.2.1 The 1st round experiment

$k_1\&k_2, k_3\&k_4, k_5\&k_6$ were predefined as three pairs of keywords, which are selected as keywords in a chosen domain/application M (where M is the concept/definition of the domain).

Function $f_{GS}(x, y)$ was the process of executing a snowball naming machine against given paired seeding keywords x and y. Set $S_{(x,y)}$ represented the collection of predicted keywords, from $k_1^{x,y}$ to $k_n^{x,y}$ (superscript "x,y" defines predictions' set, subscript "n" describes the number of predictions in the target set) which were brought back by function $f_{GS}(x, y)$. For example: $k_8^{3,4}$ referred to the eighth prediction generated by k_3 & k_4. This set contains $n_{3,4}$ predictions in total.

The first round experiments can be formulised as:

$$S_{(k_1,k_2)} = f_{GS}(k_1,k_2) = \left\{k_1^{1,2}, k_2^{1,2}, \cdots, k_{(n_{1,2}-1)}^{1,2}, k_{n_{1,2}}^{1,2}\right\}$$

$$S_{(k_3,k_4)} = f_{GS}(k_3,k_4) = \left\{k_1^{3,4}, k_2^{3,4}, \cdots, k_{(n_{3,4}-1)}^{3,4}, k_{n_{3,4}}^{3,4}\right\}$$

$$S_{(k_5,k_6)} = f_{GS}(k_5,k_6) = \left\{k_1^{5,6}, k_2^{5,6}, \cdots, k_{(n_{5,6}-1)}^{5,6}, k_{n_{5,6}}^{5,6}\right\}$$

This experiment accepted input keywords as seeding words, and then obtained predictions from Google Sets. (Detailed technique is described in Appendix 3.4)

3.2.2 The 2nd round experiment

The second round experiments collected predictions from the first round, paired them up with the original seeding words as new seeding pairs, and then fed them to Google Sets for obtaining more predictions. For example, the first round seeding word k_6 and its 8th prediction $k_8^{5,6}$ in $S_{(k_5,k_6)}$ would act as the seeding words to generate a new extended collection $S_{(k_6,k_8^{5,6})}$ of $n_{6,8,5,6}$ predictions from $k_1^{6,8,5,6}$ to $k_{n_{6,8,5,6}}^{6,8,5,6}$:

$$S_{(k_5,k_8^{5,6})} = f_{GS}\left(k_5,k_8^{5,6}\right) = \left\{k_1^{6,8,5,6}, k_2^{6,8,5,6}, \cdots, k_{(n_{6,8,5,6}-1)}^{6,8,5,6}, k_{n_{6,8,5,6}}^{6,8,5,6}\right\}$$

The same tasks were performed against every possible combination of any prediction and either of its seeding words. Thus, if $S_{(k_5,k_6)}$ had $n_{5,6}$ unique types of predictions, $2n_{5,6}$ types of seeding words were populated ($n_{5,6}$ predicts joint with both k_5 and k_6) for extended snowball sampling experiments.

So all possible extended collections for k_1 and k_2 can be formulated as:

$$S_{(k_1\ ,k_1^{1,2})} = f_{GS}\left(k_1, k_1^{1,2}\right) = \left\{k_1^{1,1,1,2}, k_2^{1,1,1,2}, \cdots, k_{(n_{1,1,1,2}-1)}^{1,1,1,2}, k_{n_{1,1,1,2}}^{1,1,1,2}\right\}$$

$$\vdots$$

$$n_{1,2}$$

$$\vdots$$

$$S_{(k_1\ ,k_{n_{1,2}}^{1,2})} = f_{GS}\left(k_1, k_{n_{1,2}}^{1,2}\right) = \left\{k_1^{1,(n_{1,2}),1,2}, k_2^{1,(n_{1,2}),1,2}, \cdots, k_{(n_{1,(n_{1,2}),1,2}-1)}^{1,(n_{1,2}),1,2}, k_{n_{1,(n_{1,2}),1,2}}^{1,(n_{1,2}),1,2}\right\}$$

and

$$S_{(k_2\ ,k_1^{1,2})} = f_{GS}\left(k_2, k_1^{1,2}\right) = \left\{k_1^{2,1,1,2}, k_2^{2,1,1,2}, \cdots, k_{(n_{2,1,1,2}-1)}^{2,1,1,2}, k_{n_{2,1,1,2}}^{2,1,1,2}\right\}$$

$$\vdots$$

$$n_{1,2}$$

$$\vdots$$

$$S_{(k_2\ ,k_{n_{1,2}}^{1,2})} = f_{GS}\left(k_2, k_{n_{1,2}}^{1,2}\right) = \left\{k_1^{2,(n_{1,2}),1,2}, k_2^{2,(n_{1,2}),1,2}, \cdots, k_{(n_{2,(n_{1,2}),1,2}-1)}^{2,(n_{1,2}),1,2}, k_{n_{1,(n_{1,2}),1,2}}^{2,(n_{1,2}),1,2}\right\}$$

The same formula will be applied to the rest of the first round predictions in the same manner. Following this, a similar process of feeding new seeding words to Google Sets will be performed, to obtain the final extended collections and record them.

3.2.3 The 3rd round experiment

All the unique predictions after second round experiment (from k_{p1} to k_{pn}) formed the final prediction Set S_p, and they were re-paired to be the seeding pairs of the third round experiments.

Table 3.6: The third round experiment formula collection

$S_{(k_{p1},k_{p2})} = f_{GS}(k_{p1},k_{p2}) = \left\{k_1^{p1,p2}, k_2^{p1,p2}, k_{(n_{p1,p2}-1)}^{p1,p2}, k_{n_{p1,p2}}^{p1,p2}\right\}$ $\vdots$ $S_{(k_{p1},k_{pn})} = f_{GS}(k_{p1},k_{pn}) = \left\{k_1^{p1,pn}, k_2^{p1,pn}, k_{(n_{p1,pn}-1)}^{p1,pn}, k_{n_{p1,pn}}^{p1,pn}\right\}$	}n-1
$S_{(k_{p2},k_{p3})} = f_{GS}(k_{p2},k_{p3}) = \left\{k_1^{p2,p3}, k_2^{p2,p3}, k_{(n_{p2,p3}-1)}^{p2,p3}, k_{n_{p2,p3}}^{p2,p3}\right\}$ $\vdots$ $S_{(k_{p2},k_{pn})} = f_{GS}(k_{p2},k_{pn}) = \left\{k_1^{p2,pn}, k_2^{p2,pn}, k_{(n_{p2,pn}-1)}^{p2,pn}, k_{n_{p2,pn}}^{p2,pn}\right\}$	}n-2
$\vdots$	$\vdots$
$S_{(k_{p(n-2)},k_{p(n-1)})} = f_{GS}(k_{p(n-2)},k_{p(n-1)}) = \left\{k_1^{p(n-2),p(n-1)}, k_2^{p(n-2),p(n-1)}, k_{(n_{p(n-2),p(n-1)}-1)}^{p(n-2),p(n-1)}, k_{n_{p(n-2),p(n-1)}}^{p(n-2),p(n-1)}\right\}$ $S_{(k_{p(n-2)},k_{pn})} = f_{GS}(k_{p(n-2)},k_{pn}) = \left\{k_1^{p(n-2),pn}, k_2^{p(n-2),pn}, k_{(n_{p(n-2),pn}-1)}^{p(n-2),pn}, k_{n_{p(n-2),pn}}^{p(n-2),pn}\right\}$	}2
$S_{(k_{p(n-1)},k_{pn})} = f_{GS}(k_{p(n-1)},k_{pn}) = \left\{k_1^{p(n-1),pn}, k_2^{p(n-1),pn}, k_{(n_{p(n-1),pn}-1)}^{p(n-1),pn}, k_{n_{p(n-1),pn}}^{p(n-1),pn}\right\}$	1

If there are n predictions addressed, the seeding word pairing possibility will be $(n-1)!$ types (note that it is not $(n-1)^2$ types since $f_{GS}(x,y)$ and $f_{GS}(y,x)$ are

the same), so eventually $(n-1)!$ collections were produced as table 3.6 shows, according to the formulas in table 3.6. In these formulas: any $k_{pn} \in S_p$.

Corpus construction stored the seeding words, their predictions and the semantically related relationships between them. These terms and their relationships formed a "domain keywords (concepts) network". This network is similar to many social networks and there are well established social network analysis methods which can be applied to the collected data to conduct ontological analysis.

3.3 ONTOLOGICAL ANALYSIS

Ontological analysis reflected the later stage in the development cycle: finding the "roots" – representatives of the network; clarifying links between new domain terms and "roots"; clustering subtrees and defining boundaries of subtrees and even the whole network. This analysis was essential to for ontology output with a hierarchical structure to enable easier application in ICT system, and to be able to form ontology output from different perspectives to suit different applications in different domains. The analysis started from deriving each keywords social position, namely their centrality in the network.

3.3.1 Centrality Analysis

After building up such a network, those members who had been "derived" (named by others) more times than others could be regarded as more representative of the group, or more "centrally" located, since such a centrality generalised representative

concepts (super connectors) from a groups of keywords (social network members) within the network (Katz, 1953).

In this research, centrality referred to the "Degree Centrality" of a social network analysis, it measured the direct connections a node had in a network (n connections to a node = n degree on a node). Centrality has been one of the most important and widely used methods in network analysis to identify the most important actors – at both individuals and groups levels.

- Individual Centrality: Identifying those individuals in the keywords network generated in the corpus construction stage would locate the "super connector", who can be the "shared property" other members had, and it could represent large portions of the corpus, or even the whole corpus in extreme cases. Keyword D in Figure 3.11 had five connections in the exemplar network; therefore, D was the most representative of the network. G also had a relatively higher representativeness with four direct connections, so G was also a representative of the network.

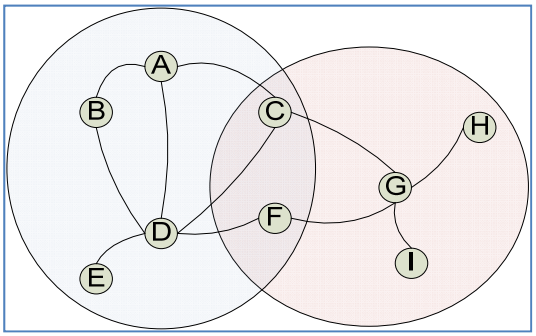


Figure 3.11: Example of Centrality Analysis

A very centralized network is dominated by one, or a few very central nodes. Hence, such super connectors are more general or more core conceptual than other members of the network. These super connectors may be treated as the root nodes of the corpus.

- Group Centrality: Identifying those groups who would map the clusters of the members within a network. Figure 3.11 shows that there are two groups that could be clustered: D centred and G centred. Such groups in the keywords network indicated conceptual clusters (group keywords). Their shared concepts (a super connector in the centre of the group) may be the leading concept of a "subtree" as SENSUS describes. Thus, group centrality may help to identify the subtrees which could be included in the ontology.

In practice, centrality counted the number of connections a target member had. While the target member was an individual in the network, the centrality equals the number of relations it had (edges in a graphical representation). It may be expanded to group centrality when the target members were a group of members in the network. In such cases, it took account of the number of connections in the group from members outside of the group.

3.3.1.1 Centrality Calculation

The corpus construction result described the experiment network in $(n-1)!$ sets of collections. To examine the centrality of a target member m in such data structure,

the calculation had to go through every collection to count the possible relations m has had with all the possible seeding words. Thus, the centrality algorithm had two steps:

1. Firstly, verifying the existence of m in every collection (Set S), under the conditions that Set S was not seeded by a pair of words including m itself. The existence of m in Set S was configured as $f_E(m,S)$ to generate a numeric value:

$$f_E(m,S) = \begin{cases} 1, m \in S \\ 0, m \notin S \end{cases} \mid f_{GS}(m,k) \neq S\Big\}$$

Where: $S = \forall S_{(k_{pi},k_{pj})} \mid 1 \leq i < j \leq n$

And: $m \in \{k_{p1}, k_{p2}, \cdots k_{pn}\}, k \in \{k_{p1}, k_{p2}, \cdots k_{pn}\}$, and $m \neq k$;

2. Then, the total connections of m in these sets were the aggregation of $f_E(m,S)$. It can be calculated as the centrality:

$$f_c(m) = \sum_{i,j} f_E\left(m, S_{(k_{pi},k_{pj,})}\right) \mid 1 \leq i < j \leq n$$

The centrality calculation $f_c(m)$ did produce distribution of the centrality changes. However, this distribution covered such a large area that it cannot be explained as one particular trend. Therefore, various alternative centrality measures can be applied to provide a better understanding, such as segmentation based on the centrality distribution, centrality increment, and relevant centrality increment

(Detailed techniques discussed in Appendix 3.5). Different methods provided multiple analytical methods to explain the centrality variation. With further support of closeness and betweenness analysis (discussion in section 3.3.2), the similar tri-sectional division was uncovered in both engineering and medical corpus (Figure 3.12).

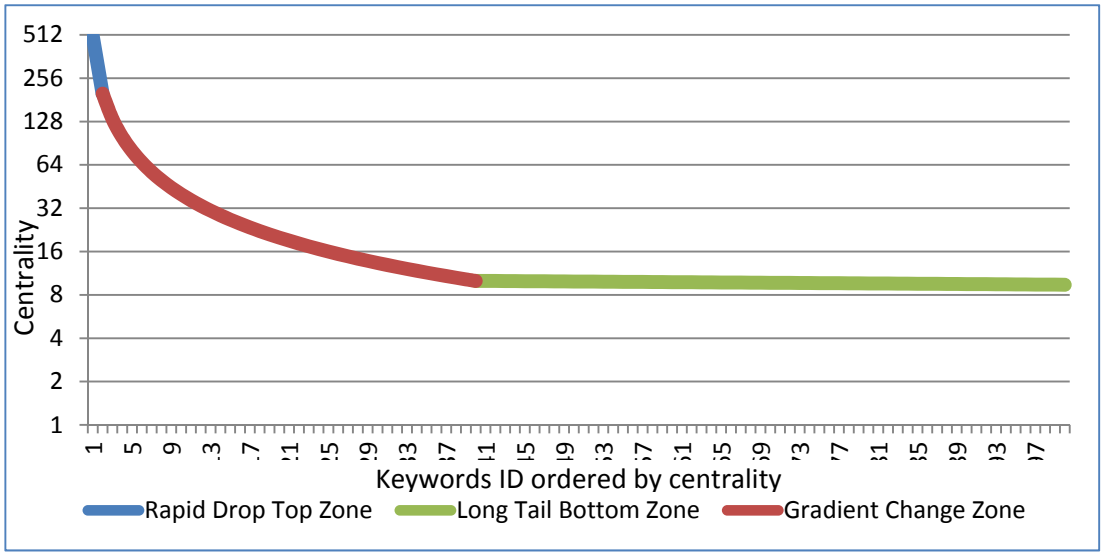


Figure 3.12: Tri-sectional trend of centrality distribution

This distribution is similar to a Poisson distribution in probability theory and statistics, which presents the probability of a number of events occurring in a fixed period of time if these events occur (Figure 3.13). In such type of distribution, three distinguish zones can be observed:

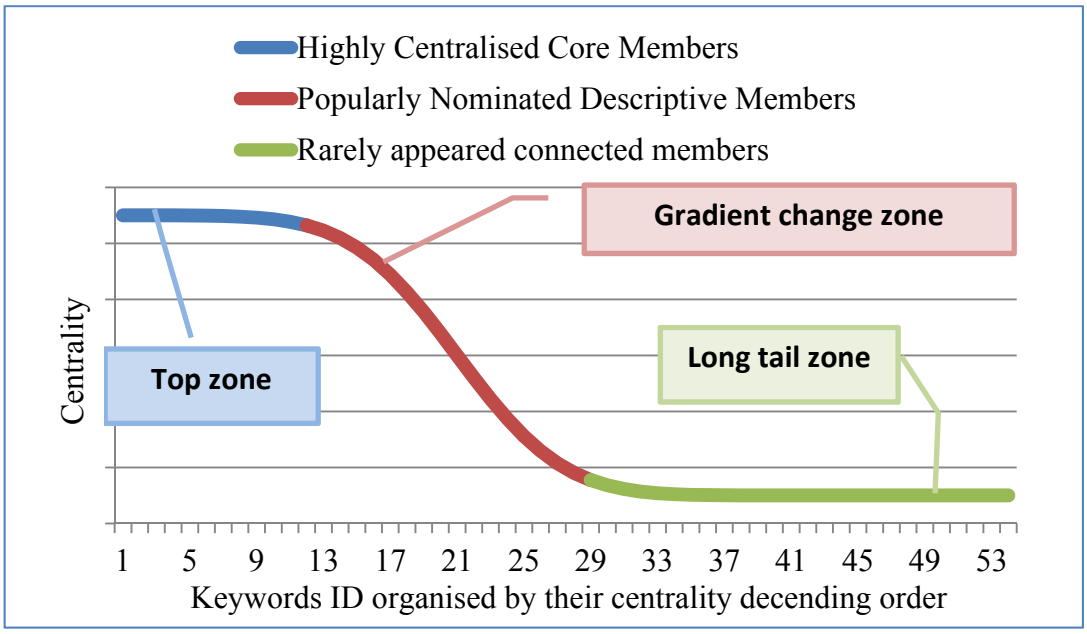


Figure 3.13: Illustration of Poisson distribution

1. A fully connected top zone with dozens of highly centralised members. The top zone was filled with highly centralised top level concepts, which were the most popular concepts "referred" (as a prediction) by the fellow concepts. These words were the keywords to define the domain. Mathematically, these keywords appeared much more often than the other members outside of the zone.
2. A partially connected gradient change zone with a few hundreds of highly (but not fully) connected members. Their centralities were much lower than the top zone, but most of them were connected to top zone members. The gradient change zone was composed of popular supporting concepts or instances that described the top zone members. They complemented the top level conceptual clusters. Terms in this level were closely related to the top level members but were not as important as the top level members in terms of domain representation. Further observation reveals that many of these members were phrases containing

words in the definition zone or their thesaurus, which made them more like descriptions than definitions.

3. A long tail zone made up by thousands of low centralised members, most of which had centrality of one (indicating they were mentioned only once in the whole network). The long tail zone contained keywords mentioned by the members in the top two levels. Members at this ground level did not necessarily describe the concept accurately, but they were connected to the top two levels more or less in a context which may or may not be associated with the target domain. This may benefit the research in boundary analysis and these terms may bridge relationships with other domains.

A trisection divided mechanism made the overall centrality understanding much easier as a means of representing the nature of domain concepts variation: different level of representatives of the domain had been created. From an ontology engineering viewpoint, this was a process of identifying the core concept and group members based on their conceptual relatedness.

Centrality analysis may the highly representative terms (top zone members) of the ontology structure, but was not able to "find the paths" from those original seeding words to the top zone members. Further analysis was required to reveal the nature of different concepts in terms of their relations with each other. Thus analysing binary relationship between members (or groups) became the core issue at the next stage of

processing. There were mainly two types of relationship considered: the "closeness" and the "betweenness".

3.3.2 Closeness and Betweenness

"Closeness" analysis helped to shape the conceptual clusters around the centralised concepts to provide a more comprehensive description of the concepts and clarify the relationships among them. "Betweenness" analysis was then implemented to assist in uncovering the overall structure of the network to identify the bridging elements that connect every member together in the domain network structure.

3.3.2.1 Closeness Analysis

"Closeness" analysis took clusters as observation objects to measure how close clustered concepts were to each other. Unlike centrality analysis, it counts the connections to a cluster from non-cluster members outside it. A cluster can also be an individual node, which made the closeness analysis also applicable to binary relationship analysis between individual members within a network. The sum of connections from a conceptual cluster to another could be treated as the relevant power between them. This relevant power can indicate the closeness between conceptual clusters. In addition, the sum of connections provided a numeric value, and it could be converted (a simple method is to use reciprocal) to a value from 0-1, which could represent the distance between conceptual clusters.

For example, Figure 3.14 showed that if the circled concept (including A, B and C) was the observing object, D would be a useful term to include, since it was very close (three connections) to the cluster. Therefore, closeness was useful for complementing conceptual clusters (or concepts), addressing instances, and clarifying sub concepts and their relations to the main concepts.

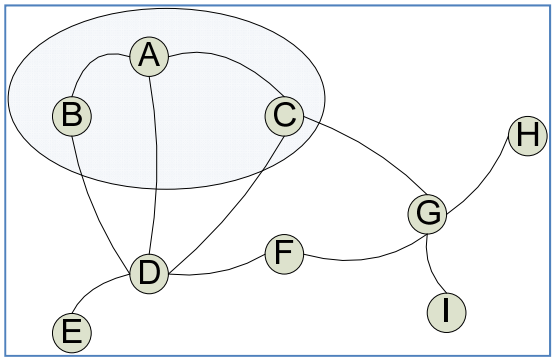


Figure 3.14: Example of closeness analysis

A key advantage of closeness analysis was that it had very flexible measurement techniques: many calculation methods have been proved to be valid, including a total count number from centrality analysis. This indicated that firstly, the total count data collected from the centrality analysis could be reused. Secondly, a different focal cluster only required a different aggregation of their individual closeness figure.

In this research, the closeness investigated how important a seeding word k was in predicting m, and in semantic relatedness terms, how much did seeding word k determine the appearance of prediction m in the domain. Centrality analysis defined $f_c(m)$ to track m appearances in all the prediction sets, regardless of their

seeding words. If seeding words were considered, for example a seeding word k, $f_c(m,k)$ can calculate m's appearances via a traversal of these sets, based on k.

$$f_c(m,k) = \sum_i f_E\left(m, S_{(k_{pi},k)}\right) \mid 1 \le i < j \le n$$

Then, the decisive power of seeding word k on predictions m could be presented as a closeness distance $f_d(m,k)$, the greater $f_d(m,k)$ was, the greater the decisive power k had to predict m.

$$f_d(m,k) = \frac{f_c(m,k)}{f_c(m)}$$

It was predictable that $\mathrm{f_d(m,k)}$ does not need to be identical to $\mathrm{f_d(k,m)}$, since both the numerator and denominator would probably be different. $f_c(m,k)$ depends on how important m "thought" that k was in its own sub network. As an analogy, in an actual social network people's definitions of "close friends" may vary. Additionally, there were many more members in the network in general to produce different sub networks for both of them, which led to a very different $f_c(m)$ and $f_c(k)$. For example, k and *m* may be friends and have 10 phone calls to each other every day. However, *k* may have 100 phone calls every day, only 10 of them with *m* , another 20 conversations were with *p*. Thus, from *k*'s perspective, *m* may not be the closest friend with one tenth of daily phone conversation, since *p* talked to *k* twice as much every day. On the *m* side, *m* may only has 12 phone calls a day. And with 90% conversation with *k*, *k* could be *m*'s closest friend.

The result of practical closeness analysis on the corpus confirmed that different seeding words had different decisive powers over the number of appearances of a target word. A quantified value helped to refine the zone definition from centrality analysis, as centrality analysis can only conduct zone specification from a structure perspective. Specified weight of binary relationships also uncovered that lower centralised keywords may be even more remotely connected than centrality analysis showed (centrality analysis treated all connections equally). Such distinctive variation was ideal in shaping the concept clusters in terms of addressing core concepts and boundary concepts.

Furthermore, it proved that f_d was directional, this provided the ability to observe and output the ontology from any required angle. This was especially important for multi-disciplinary ontology analysis.

Finally, low f_d flagged potential conceptual brokers, who could be the key to cross domain relationships. To further examine the brokers and understand the whole relationship mapping, "Betweenness" analysis was required.

3.3.2.2 Betweenness Analysis

"Betweenness" analysis identified those members whose importance may be ignored by centrality analysis and closeness analysis but who bridged the gaps between concept clusters.

For example, in Figure 3.15, G played the key role of connecting the D centred cluster to the H centred cluster, so G was the broker between the two concepts. If G was considered as a concept, it would have a high "betweenness" value. But it was not a highly centralised member from a centrality viewpoint, nor a decisive member for any conceptual clusters from closeness viewpoint.

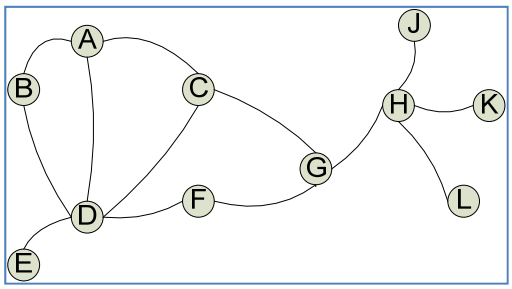


Figure 3.15: Example of "Betweenness" analysis

Betweenness analysis found those individuals or groups who have concurrent membership in overlapping clusters, so the relations between concepts became clearer. In this research, members with significant "betweenness" factors could be found via the following method:

1. Reference to the closeness addressed those members with a low closeness in the network; this meant that such concept clusters were semantically further than others. In this research, special attention was paid to those numbers that are remotely positioned in both directions. For instance, the traversal of f_d could address predictions m_1 and m_2, where

$$f_d(m_1, m_2) \rightarrow 0 \quad \text{and} \quad f_d(m_2, m_1) \rightarrow 0$$

Addressing this sort of relationship was the key to clarifying the conceptual clusters, especially when both m_1 and m_2 were highly centralised members. It provided numerical figures to draw boundary between m_1 and m_2.

2. But there may exist a prediction k which is decisive for both m_1 and m_2:

$$f_d(k, m_1) \rightarrow \max_{1<i<n} f_d(k, k_{pi}) \quad \text{and} \quad f_d(k, m_2) \rightarrow \max_{1<i<n} f_d(k, k_{pi})$$

Such k connected m_1 and m_2 from k's view point. The existence of such k showed that bridging concepts existed and could be located. It also indicated that the peripheral players of a network should not be omitted, since they may be the bridge to other networks which are not revealed in the target domain.

For example, table 3.7 showed that "folding" and "honing" in the generated engineering ontology were not particularly close to each other, but there was a member "tool grinding" which is tightly connected to both of them.

Table 3.7: Example of betweenness in the engineering experiment

Seeding Words (**k**)	Predict(**m**)	$\mathbf{f_c(m,k)}$	$\mathbf{f_c(m)}$	$\mathbf{f_d(m,k)}$
Folding	honing	3	2121	0.001414
Honing	folding	1	1131	0.000884
tool grinding	honing	83	2121	0.039132
tool grinding	folding	58	1131	0.051282

Being able to address members like "tool grinding" had a significant practical benefit as in a "folding" centred context, the target ontology may not even consider "honing" as an option, but if the context contains "tool grinding", then "honing" should be a

potential aspect to consider. The analysis revealed that this method of analysis was able to create well positioned "betweenness" measures between members.

3.3.2.3 Closeness and Betweenness Summary

The "Closeness" and "Betweenness" analysis techniques supplied ontology analysis with distance measurement between concepts and their internal members. With the understanding of position of the members (centrality) and their relationships (closeness and betweenness), the corpus now can function as ontology.

However, different ontology applications may require different representation, which may lead to different size of conceptual cluster and depth of relationships. For example, WMCCM may practically require a certain depth of relationships for the core concepts in order to increase the computational efficiency, or TMG may only require concepts within a certain distance to its existing concepts. Thus, a final stage of boundary analysis was conducted to shape the ontology output.

3.3.3 Boundary analysis

Boundary analysis prunes the ontology based on the findings from the previous analysis – centrality, closeness and betweenness: Core members in a network may be addressed via these types of analysis. The problem remaining was how to identify the boundary players of a network and how this research should treat them. This depended on the ability to identify who they are (peripheral player identification) and what further action should be taken towards them (network reach analysis).

3.3.3.1 Peripheral Players

Centrality analysis may help identify some peripheral players by spotting members who have very low centrality value, which implies that they are not strongly associated with the core concepts of the network. For example, in Figure 3.16, suppose G was the core concept, previous analysis would point out that E, J, K and L had low centrality ($f_c = 1$) and did not directly connect to the core, hence they could be treated as the boundary member.

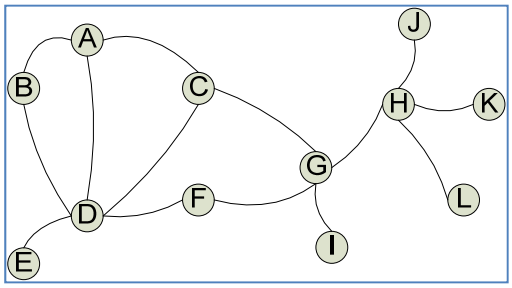


Figure 3.16: Example of Peripheral Players

From a "closeness" point of view, a low "closeness" figure to the top zone members could also define the outskirt members. Thus peripheral players could also be members directly linked with core members, if their relations were very weak.

Besides centrality and closeness analysis, network density could measure the influence of low centralised members to locate the boundary players for the whole network. In this research, network density calculation was defined as f_D, it measured the proportion of the captured relations in the network (N_t) out of the possible

maximum relations f_{Dm} (if there were N members in the network, the maximum binary relations in this network was $N(N-1)$).

$$f_D = \frac{2N_t}{N(N-1)}$$

The analysis of both corpus revealed that long tail zone member heavily dragged network density down. For example, engineering keywords network could have had 36% density but actually it had only 0.25%, because 90% were long tail member who only contributed less than 40% of the connections. This indicated that the long tail members should be defined as peripheral players.

Social network analysis research normally treated the peripheral (boundary) member of a network as unimportant data or "noise", and it was essential to draw a line between core data and peripheral data. However, it could be dangerous to ignore such noise in ontology engineering, since it might represent new knowledge in the domain. In particular, peripheral players in one domain could be well-positioned, new relating terms from a different domain, but the current relationships between the target domain and the peripheral players were not strong enough to fully explore the extended network of the new domain. Since all members were semantically related terms extracted from the Google search engine index, emerging new concepts in the domains identified by the seeding words may exist as boundary members in their early stages. Thus, they were the key factors in bringing in new concepts and relationships in terms of multi-domain connections.

3.3.3.2 Network Reach

Conflicts between being the peripheral players and potential entrants from another domain made it difficult to decide how to treat the peripheral members. A network reach perspective could provide other viewpoints in helping to prune the ontology.

It is predicted that the shorter paths in the network were more important and practical. It is recommended that applicable research should focus on a network which only includes concepts within three steps from the target observing concepts (Friedkin and Johnsen, 1997, Burt, 2004). For example, in Figure 3.17, if A is the observing concept, the network range beyond B1, B2 and C1 may not worth investigating.

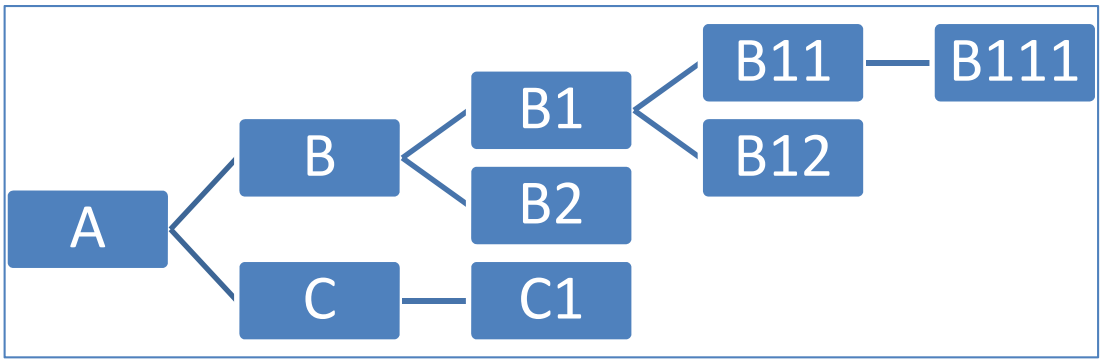


Figure 3.17: Network reach illustration

From domain focus perspective, wider network range may cover too large an area. The "six degrees of separation" concept revealed that “everyone is on average approximately six connections away from any other person on Earth” (Watts, 2004). This also applies to ontology network: terms that are three steps away from the core concepts may not be reasonably related to a domain, but only serve to expend the

network to an uncontrollable level. Furthermore, the two steps network reach proposal also suggested that three rounds of snowball sampling may be an optimum number, since the unique predictions from the third round would be three steps away from the core concepts.

In this research, the goal of the network reach control was to maintain only manageable contacts; in other words, to those closely connected members only. Thus, the members who were three steps away from the top zone steps may be beyond their reach.

Centrality analysis extracted major concepts as the top zone in the distribution curve (demonstrated by Figure 3.12 in section 3.3.1.1), then closeness analysis refined this zone. As a result, the top zone members in this research were fully connected, they reached a maximum density ($f_D = 1$), which indicated that all the possible relationships were revealed. Being able to form such a complete network, top zone members also strengthened their highly centralised position, therefore they were entitled to be at the core of the network.

Most members in the gradient change zone had at least one direct connection to some but not all of the top zone members. Only one step away from a number of core members kept gradient change zone members within the network reach. However, they were neither able to form a complete network among themselves, nor had full connections to the top zone members. As a result, there were occasions that members had network reach as far as five, although most of them could reach each

other in a maximum of three steps. In order to achieve a fast ontology pruning from any given focal, the three steps rule was applied to gradient change zone members to eliminate members who were too far away from the core concepts.

Long tail members could be considered as boundary players in the peripheral player analysis. However, a lot of long tail members had short network reach (less than three steps) from the top zone. Such members may need to be kept in the ontology structure to enable potential domain crossing. As a result, there could be too many of them for some ontology applications (e.g. WMCCM ontology does not required thousands of concepts for any category). Arguably, thresholds on closeness measure could limit the number of long tail concepts. The closeness measures were calculated to reflect absolute values from the whole network's perspective. In practice, the same closeness value may imply different level of closeness from individual's viewpoint, since every individual in a network may have its own preferences to define "closeness".

For example, "custom fabrication" had 0.0126 closeness towards "electroforming" (Table 3.8). From electroforming's viewpoint, it was not a "close friend". However, 0.0126 was higher than many other concepts' closest distance to their members including "custom fabrication". In this case, a threshold of 0.0126 would keep too many members in the electroforming conceptual cluster, but remove all connection to "custom fabrication" and many other conceptual clusters. Hence it is difficult to set a

threshold according to the "closeness value", as it may ignore individual variations on "closeness".

Table 3.8: Example of relevant distance in the engineering experiment

Seeding Words (k)	Predict(m)	$f_c(m,k)$	$f_c(m)$	$f_d(m,k)$	Distance
surface grinding	electroforming	122	1593	0.076585	1
Slotting	electroforming	121	1593	0.075957	1.008264
threading	electroforming	61	1593	0.038293	2
deburring	electroforming	33	1593	0.020716	3.69697
punching	electroforming	29	1593	0.018205	4.206897
surfacing	electroforming	24	1593	0.015066	5.083333
custom fabrication	electroforming	20	1593	0.012555	6.1
assembly	electroforming	16	1593	0.010044	7.625

Therefore, "relevant distance" was introduced to control the relationships around concepts. Relevant distance compared the distance from concepts k to concept m with the closest relationship m had $(\max f_d(k,m) / f_d(k,m))$. It converted absolute binary relationships to relevant relationship against given concepts. Relevant distance provided a different measurable distance from any concept's viewpoint. Threshold now could be applied to limit members around concepts.

By applying relevant distance, conceptual clusters (or concepts) could be pruned individually according their own relevant network reach perspective.

3.3.3.3 Boundary analysis Summary

Peripheral player identification and network reach limitation shape the final ontology output. A series of steps has been described to generate ontology. These steps can be

repeated, and enable the key steps and considerations identified as important by the analysis of other ontology building methods in chapter two.

3.4 SUMMARY

This approach has been derived initially on the basic form of the SENSUS methodology, but the detailed techniques differ. The approach has utilised semantic relationship to extract terms from the search engine index, and further applied social network analysis methods to conduct the ontological analysis. Thus the author proposes this approach as SEA: Semantic (S) relatedness oriented ontology engineering via retrieving information from the search Engine (E) index with assistance from social network analysis (A).

The SEA approach to ontology generation now requires evaluation by testing the resulting ontology for the two case study applications against other practical ontology developed for the corresponding applications.

CHAPTER 4: ONTOLOGY EVALUATION

The previous chapter described how a range of techniques and tools could be combined to create a new ontology engineering approach - SEA - to derive domain and intra-domain focused ontology. The SEA approach has benefits compared to other methods as discussed in chapters two and three. However, a key aspect that is difficult to evaluate, is how good the resulting ontology is. "Good" can only really be defined in terms of the how well the ontology meets the needs of its applied areas.

The case studies described in chapter one and two will be used to assess the two separate ontology derived using the SEA approach. These are the West Midlands Collaborative Commerce Marketplace (WMCCM), a 10,000+ company members system that has helped generate over £10 billion in revenue for its members in the past two years. The second is the Taj Medical group (TMG), a micro SME in the medical tourism area. Thus WMCCM required a broad ontology that covered the general engineering sectors, while TMG required one that covered the medical domain, largely but not exclusively focused on the "elective" treatments area.

The evaluation was conducted by integrating several evaluation proposals. An early method for ontology evaluation was proposed by Gómez-Pérez and adopted in METHONTOLOGY (Gómez-Pérez, 2001) as a formal evaluation stage. This method mainly evaluated ontology from a knowledge representation point of view: evaluating the conceptualisation processes and resulting concepts to examine the internal relationships and the ontology structure against the real-life knowledge

structure. Welty and Guarino specified more detailed methods for ontology representation evaluation, such as structural validation, content evaluation on concepts and their relationships' rigidity (Welty and Guarino, 2001). This evaluation measured a methodology based on the evaluation of its ontology output, an evaluation normally conducted by domain experts. The On-To-Knowledge project (Staab, 2001) alternatively recommended an evaluation by considering ontology builders and users' opinions: a verification should be conducted to evaluate whether ontology engineering process and the ontology derived met the ontology builder's original requirements; and a practical assessment should be carried out to evaluate whether the ontology developed satisfies users' expectation within the application environment. These evaluation methods emphasized different aspects: ontology structure, contents, ontology building requirements and practical usage.

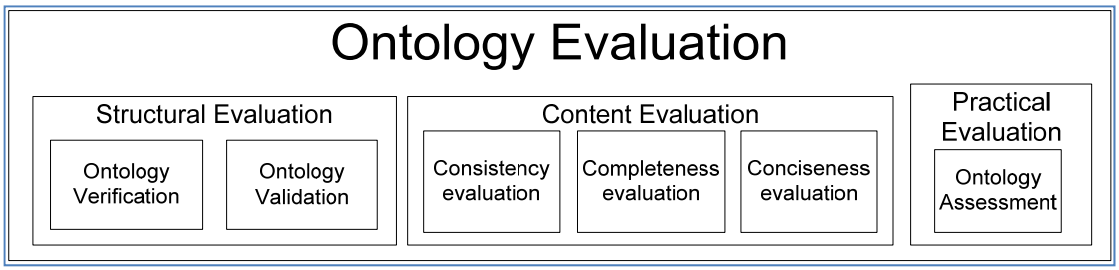


Figure 4.1: Ontology Evaluation

Individually, evaluation on each of these aspects may serve to justify ontology usability, but may neglect hidden issues that are more visible from other aspects. This research proposes a novel approach – SEA – and derives two new ontology, whose applicability and usability have not been formally verified. Therefore, it is

sensible to examine these resulting ontology from multiple angles. As shown in Figure 4.1, the evaluation stage in this research consisted of three activities: structural evaluation (including evaluation from the builders' perspective), content evaluation, and practical evaluation (evaluation from users' perspective).

4.1 STRUCTURAL EVALUATION

Structural evaluation starts the evaluation process by verifying and validating the constitution of the derived ontology and it includes:

- Ontology Verification: verifying whether the new ontology built via SEA was built according to the ontology requirement. In this research, this is focused on whether the new ontology contain a sufficient number of concepts and relationships, and whether these are "economically" and "quickly" built.
- Ontology Validation: validating whether the derived ontology structurally modelled the expected scenarios in the real world. In this research, it assesses the zone definitions of the new ontology.

4.1.1 Ontology Verification

Two separate ontology verification exercises were conducted, covering engineering (WMCCM) and medical (TMG) ontology to compare the production of the corpus and the development process.

4.1.1.1 Engineering Corpus

An initial three pairs of seeding words were obtained from WMCCM project team members using the Delphi method. These were three pair keywords from the machining section of the WMCCM categories: drilling & cutting, milling & sawing, and turning & grinding. From these, using the procedure described in chapter three, 10,803 unique terms with 266,176 relationships among them were automatically generated (detailed data available in appendix 3.5 page 2-30, raw data available in appendix 1.3.3). Previously WMCCM had used manual processes to collect only 862 unique concepts with 2,126 relationships from a combined contribution of both SIC and domain experts. The new ontology contained fifty times more terms, and more than a hundred times the internal relationships of the original WMCCM ontology.

Apart from generating more terms and relationships and thus hopefully a richer mapping of the domain, the ontology required less time to generate and the efforts of fewer domain experts. The automated construction process took less than 12 hours to generate the new ontology from the initial three pairs of seeding words: six hours to capture the terms and relationships; and five hours of ontological analysis to clarify the relationship and construct the structure.

The previous manual WMCCM method started by building the ontology from a pruned SIC structure and involved interviews with more than 300 expert user groups (engineering companies) to validate and expand it. This method took the ontology builder at least an hour to capture the client's industrial categorisations, so in total it

cost at least 300 hours for the corpus construction process. Indeed, these measures did not take into account the time and monetary cost of arranging meetings, getting to meetings, discussions and refinements. Nor did they cover the time spent on "conflict" resolution, combining the inputs of so many experts. There is no doubt that the new methodology is significantly more efficient in terms of time and cost. It also generated a richer corpus – more concepts and richer internal relationships.

4.1.1.2 Medical Corpus

The medical corpus construction followed the same procedure as in engineering domain. Three pairs of seeding words from common medical treatments were selected using the Delphi method with TMG staff, and from other studies of the medical tourism industry. The terms were: cardiac surgery & paediatric surgery, orthopaedic surgery & otolaryngology, plastic surgery & neurosurgery.

The corpus construction experiments brought back 1,269 unique terms with 347,994 relationships among them (detailed data available in Appendix 3.5 page 20-32, raw data available in appendix 1.2.3). On average, there were 275 relationships to a single term. The current TMG ontology could only generate 301 concepts, less than 900 relationships, and thus less than three relationships link to a single concept. This comparison highlights the fact that the new methodology had an advantage in generating a large number of concepts with rich internal relationships.

The internal relationship richness is comparable to some larger ontology such as UMLS and SNOMED CT in the medical domain. The UMLS meta-thesaurus which combined relationships from more than 100 sources only provided 13 relationships to a concept on average. Although the new methodology extracted fewer concepts (due to the focus of seeding words and only three rounds of snowball sampling) than SNOMED CT, it captured nearly half the number of relationships that SNOMED proposed. SNOMED CT extracted large number of concepts by consulting thousands of medical professionals, while the new corpus was achieved by starting with three pairs of seeding words in a fraction of the time.

Less than 12 hours were spent on automated corpus construction and ontological analysis in building this new medical domain ontology. SNOMED CT cost more than 10 years of time and required doctors and nurses from six countries to reach the current stage. The existing TMG ontology engineering took more than 600 hours of effort (one ontology builder's research time of 500 hours and 100 hours domain experts' contribution). SEA shows an advantage in building medical ontology more quickly and economically.

4.1.1.3 Other findings

SEA generated more terms and relationships with direction and weight. This gives the generated ontology broader coverage, richer internal connections and more specified binary relationship (instead of simple connections).

Taking the new engineering ontology as an example: compared to the limited connects in the top level categories in the current WMCCM ontology, observation on the new engineering ontology's top zone found that members were fully connected in effect. A top zone member already possessed high centrality, which reflected its representativeness for the domain. In fully connecting to all other similar members (nominated by other top zone members), a top zone member demonstrated stronger evidence to represent the domain since the recognition from other strong domain representatives. Thus, taking the whole group of top zone members as an object, full connection strengthened the top level terms as the core concepts in the whole corpus.

While moving downwards to the second level, the current WMCCM ontology members had very limited connection with each other, and only linked to their direct parent or child members. In the new engineering ontology, descriptive members in the gradient change zone still cross referenced to each other (though not fully) and even to most of top level members. This created a richer network structure. With this rich network structure established, conceptual clusters could be observed from various angles/viewpoints, this enables the new ontology to function (partially) as a faceted system (Giunchiglia et al., 2009). A rich internal structure also provides more paths to bridge gaps between concepts, to link core concepts (specialist terms) with peripheral players (non-specialist terms), and formed a foundation for fuzzy match.

The new methodology generates connections between different terms that are weight specified directional relationships (like vectors) based on the "closeness" value. Such

relationship expresses the binary relationship more richly than simple weightless connection. For example, table 4.1 demonstrates the relationship between several terms to the concept "electroforming" (raw data available in appendix 1.2.5 and 1.2.6). In the other ontology (SIC, WMCCM), such relationships do exist but only as a simple connection showing "terms are connected". With the SEA approach, such relationships are specified ($f_d(m,k)$) to show "how much" terms are connected. It is also noticeable that directional relationships reflected different mutual recognition between terms: the bottom two rows in table 4.1 showed that welding and electroforming "treat" each other differently from their own perceptive. The awareness of "how much" and the direction of the relationship improve the fuzzy matching ability by directly indicating the probability of the match.

Table 4.1: Weight Specified Relationship

Seeding Words (k)	Predict(m)	$f_c(m,k)$	$f_c(m)$	$f_d(m,k)$	Distance
surface grinding	electroforming	122	1593	0.076585	1
Slotting	electroforming	121	1593	0.075957	1.008264
Reaming	electroforming	107	1593	0.067169	1.140187
broaching	electroforming	82	1593	0.051475	1.487805
Gear cutting	electroforming	81	1593	0.050847	1.506173
tool grinding	electroforming	77	1593	0.048336	1.584416
centering	electroforming	76	1593	0.047709	1.605263
Gear shaping	electroforming	71	1593	0.04457	1.71831
custom fabrication	electroforming	20	1593	0.012555	6.1
thread rolling	electroforming	19	1593	0.011927	6.421053
Welding	electroforming	19	1593	0.011927	6.421053
electroforming	welding	86	3076	0.027958	5.354822

In summary, the new ontology in both engineering and medical domains appears to be an improvement not only in generating many more concepts along with richer

internal relationships, but also in reducing the reliance domain experts and the time required to build the ontology.

After an overall evaluation against the ontology building requirement, the next stage of the evaluation focused on the structure of newly built ontology.

4.1.2 Ontology Validation

Ontology validation examines whether the new ontology structurally modelled the relevant knowledge in practice. Both WMCCM and TMG modelled their corresponding domain knowledge (concepts and relationships) according to their domain representative and organised them into hierarchy structures. For example in the TMG categorisation, most highly representative (or abstractive) terms were proposed as high level concepts (such as cosmetic surgery); then instances of high level concepts (or descriptions of the concepts) were modelled as a second level concepts (such as Rhytidectomy – an instance of cosmetic surgery); additionally, further details of instances or alternative descriptions of instances were provided as a lower level (such as face lift – an more detailed rhytidectomy procedure and alternatively described in natural language). Such hierarchical model also exists elsewhere, such as WMCCM and SIC (discussed in section 2.2).

In this research, the validation focused on whether the new ontology could provide similar hierarchical structures, which emphasized and models the concepts representative level according to their centrality and relationship towards each other.

The new ontology can form three zones to reflect the real world scenario according to the concepts representativeness: top level definition zone, supporting description zone, and ground level connection zone (Figure 4.2).

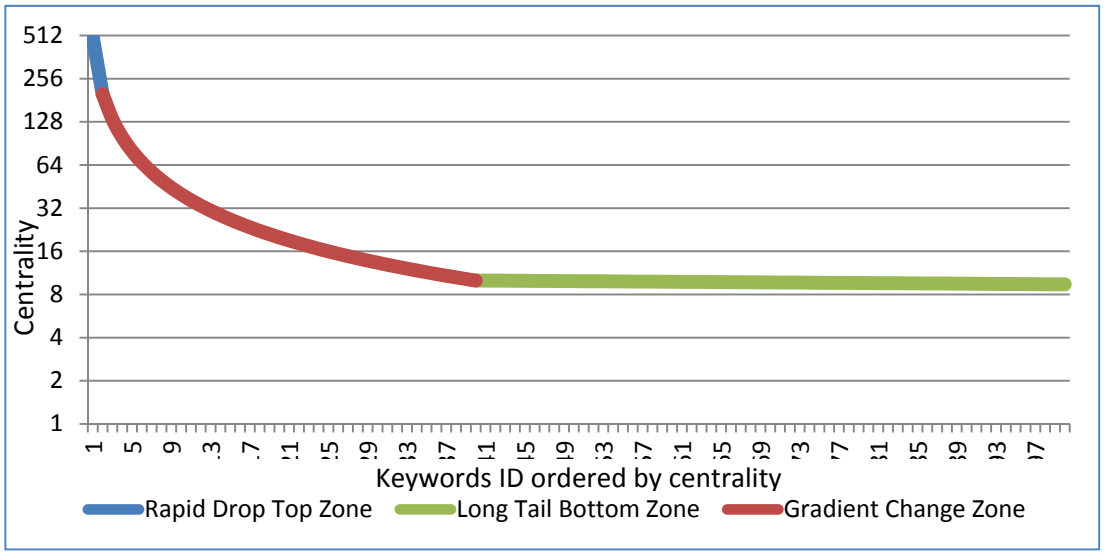


Figure 4.2: Illustration of three zones defined in this research

4.1.2.1 Top level Definition Zone

Both the engineering and medical ontology top zones were filled with highly centralised top level concepts that were the most popular concepts referenced by other fellow concepts. Members in the top zone mostly matched the top level concepts proposed by other ontology in the fields: 9/15 SIC top level terms and 16/22 WMCCM top level terms were found in the new engineering ontology; 9/11 TMG top level terms were found in the new medical ontology (the details are discussed in "content evaluation" in section 4.2). Structurally, these concepts appear

much more frequently than the other members outside of the zone. These concepts are the keywords that represent the engineering and medical subject areas.

Being at the centre of the network, top zone members also directly connected with each other. All the other domain terms who are not in the top zone can find routes to any other member term by connecting via the top zone members. Therefore, the top zone members play the "root" role (as SENSUS defined) to provide common nodes and paths for other members in the ontology.

4.1.2.2 Supporting level Description Zone

Below the top zone, there is the supporting description zone. This is populated by popular concepts or terms defining in more detail the concepts from the top zone. Observation of these concepts or terms revealed that many of them were phrases containing concepts or their thesaurus from the top definition zone. At this level, terms were inevitably connected to the relevant concepts at the top level but were not as important as them (lower centrality value). For example, "drilling" is a core concept in the new engineering ontology; its directly linked concepts "gun drilling" and "cross drilling" are description zone members.

Members in the description zone have at least one direct connection to a few but not all of the top zone members, and additionally they have limited connections with each other. Not being able to form a complete network is a distinguishing characteristic of the remainder of the network members. This pushes them away

from the centre of the network, so as to lower their representativeness of the network. The disconnection between description zone members also implies separation of their corresponding concepts (or conceptual clusters), thus borders could be drawn based on such disconnectivity.

Although not fully connected, these members can reach all top level members and most of the other descriptive members within three steps as required by network reach analysis. This short network reach indicates that the description zone was still tightly bound to the core concepts. They are relatively closer to the core concepts than the long tail members (non-expert terms) that occupy the ground level connection zone.

4.1.2.3 Ground level Connection Zone

The ground level connection zone contains "long tail" terms nominated by the terms in the two upper levels. Terms in the ground level did not necessarily describe the concepts accurately, but they were connected to the concepts or concepts descriptions to some extent in the domain context. Their focus on the main domain may be different from the upper level concepts, as they were mostly generic terms and generally did not always strictly associate with the target domains. For example, "food processing" was identified as a connection zone member in the new engineering ontology.

Practically, the connection zone does have a relationship with the main concepts. However, the frequency of appearance of the terms in this zone was the lowest in the three zones. Although the connection of its terms to the core concept was through semantic relatedness, the relationships to the core concepts were relatively further (according to closeness value). Many of its terms failed to reach the top zone terms within three steps of network reach, and this undermined their qualification of being "organic" members of the domain ontology. However, these third zone terms were valuable from other perspectives: in terms of structural clarification such members could be boundary players and from a cross domain viewpoint, they may be the brokers from the target domains to related domains.

These three zones could structurally represented core, associated and peripheral areas of domain knowledge. The detailed content within each zone needs to be also evaluated to ensure that the actual terms in each zone are as expected.

4.2 CONTENT EVALUATION

Content evaluation measured the content correctness of the resultant ontology. It combined different aspects from other formally proposed evaluation methods (Gómez-Pérez et al., 2004, Welty and Guarino, 2001) to assess the completeness, conciseness and consistency of the generated ontology.

4.2.1 Completeness

Completeness focuses on whether enough concepts of the ontology were generated to represent the target subject areas adequately. In generic terms, completeness evaluation shows how "good" the coverage of the domain ontology have.

It is almost impossible to prove the completeness of any ontology due to the appearance of emerging knowledge on a daily basis (Gómez-Pérez, 2001). Therefore, the completeness check is normally carried out as an incompleteness check. It is an attempt to scan the resulting ontology to detect whether there are any missing concepts, which either should be included in the ontology or can be inferred by other concepts. In this research, completeness evaluation compared the resulting ontology with the case study ontology to locate missing concepts. Due to the very large data set generated by SEA approach, only examples from the definition zones are shown in this discussion.

Table 4.2 shows the concepts in the top level definition zone of the "machining" ontology generated by this research in comparison with SIC and WMCCM (raw data available in appendix 1.2.4). Vertically, the first column included SIC concepts; the second column included WMCCM concepts; the third column included new derived engineering ontology concepts; and the fourth column included centrality value of the new derived ontology concepts, which indicates their location in different zones. Horizontally, the green background indicated concepts found in definition zone; the

blue background indicated concepts found in description zone; and red background indicated concepts found in the connection zone.

Table 4.2: Definition zone members in the engineering experiment

SIC	WMCCM	New Ontology matches	
boring	Boring	Boring	2408
broaching	Broaching	Broaching	2270
	CNC Laser Cutting	laser cutting	1206
	CNC Machining	cnc machining	1456
	CNC Milling	cnc milling	511
	CNC Turning	cnc turning	405
cutting	Cutting	Cutting	3012
	Drilling	Drilling	3432
eroding		eroding	64
	Fettling	Fettling	2
	Gear Cutting	gear cutting	1482
grinding	Grinding	Grinding	3128
	Hobbing	Hobbing	2305
	Manual Machining	Machining	3148
lapping		lapping	289
leveling		leveling	25
milling	Milling	Milling	3157
planning		planning	58
polishing		Polishing	2144
	Profiling	profiling	143
sawing	Sawing	Sawing	2824
	Splining	splining	37
sharpening		sharpening	92
splicing		splicing	2
	Tapping	Tapping	2879
	Thread Grinding	thread grinding	42
	Threading	Threading	2125
turning	Turning	Turning	2789
Welding	Welding	Welding	3330

SIC 2007 describes machining (first column in Table 4.2) as: “This class includes: cutting, boring, turning, milling, eroding, planning, lapping, broaching, levelling,

sawing, grinding, sharpening, polishing, welding, splicing etc. of metalwork pieces". Nine out of fifteen (9/15) concepts in the SIC were covered by the definition zone, with the remainder in the lower zones (4 found in description zone and 2 found in connection zone).

WMCCM proposed 22 concepts in the definition zone of machining (second column in Table 4.2): "Boring, Broaching, Laser Cutting, CNC Machining, Milling, Turning, Cutting, Drilling, Fettling, Gear Cutting, Grinding, Hobbing, Manual Machining, Milling, Profiling, Rotary transfer, Sawing, Splining, Tapping, Grinding, Threading, Turning and welding". With the new ontology 16 out of 22 these concepts were included in the definition zone, and another three of them have high centrality in the description zone, with the remainder staying in the connection zone.

The new ontology covers more areas (table 4.3) than both the SIC and WMCCM (raw data available in appendix 1.2.4). Such coverage indicates that the new ontology could represent the SIC and WMCCM concepts with more than 90% correct zone matching, but it also had a wider coverage.

Table 4.3: Extra definition zone members in the engineering experiment

New Ontology Extra Concepts (with centralities)							
Assembly	2765	Electroforming	1747	micro drilling	1643	Rolling	1263
Bending	2468	Engineering	919	Notching	1095	screw cutting	1741
Centering	1862	Fabrication	1490	Painting	2771	Shearing	2192
conventional turning	1852	Finishing	1511	plasma cutting	856	Slotting	1776
custom fabrication	1002	Folding	1169	Plating	1106	Stamping	1644
cylindrical grinding	1919	Forming	2331	powder coating	912	surface grinding	2077
Deburring	2344	gear shaping	1660	Punching	2685	Surfacing	1896
Design	912	heat treating	1216	Reaming	2080	tool grinding	1667

Table 4.4: Definition zone members in the medical experiment

TMG Concepts	New ontology concepts (with centrality)			
General Surgery (abdominal surgery)	general surgery	6280		
Cardiology (Cardiac Surgery)	Cardiology	7398		
comprehensive health checks				
cosmetic surgery	Cosmetic surgery	7108	plastic surgery	5099
Dentistry (Dental Care)	Dentistry	4465	Dental	2066
Neurology (neurosurgery)	Neurology	6481	Neurosurgery	4598
ophthalmology	Ophthalmology	7150		
orthopaedics (orthopaedic surgery)	orthopaedics	4582	Orthopaedic surgery	5388
otolaryngology (head and neck surgery)	Otolaryngology	4430		
Paediatrics (paediatric surgery)			Paediatric surgery	7119
vascular surgery	vascular surgery	2927		

The medical ontology also provided a much larger coverage. TMG proposed an informal categorisation with eleven top level concepts (first column in Table 4.4),

these being: orthopaedic, cosmetic, Cardiology, comprehensive health checks, abdominal surgery, Dentistry, vascular, otolaryngology, neurosurgery, paediatric and ophthalmology.

Ten of these eleven concepts were proposed by the new ontology as definition zone concepts (in Table 4.4, horizontal green background indicated concepts found in definition zone; blue background indicated alternative concepts found in definition zone; red background indicated concepts not found in the top zone. Vertically, the first column included TMG concepts; the second & fourth column included matching concepts in the new ontology in their formal terminology & thesaurus terminology; the third and fifth column included centralities of correspondent concepts).

Table 4.5: Extra top zone members in the medical experiment

New ontology extra concepts with centralities					
internal medicine	7049	Rheumatology	4321	Haematology	2750
Psychiatry	6853	Pulmonology	4287	Pharmacy	2650
Urology	6603	Endocrinology	4180	Geriatrics	2517
Anaesthesiology	5989	Nutrition	4087	infectious disease	2456
Dermatology	5922	Diabetes	3799	family medicine	2198
emergency medicine	5870	Gynaecology	3782	diagnostic radiology	2144
family practice	5566	Orthopaedics	3591	geriatric medicine	2054
Podiatry	4454	pain management	2898	Cancer	2034
Nephrology	4331	radiation oncology	2823	Orthodontics	2005

Moreover, four of the alternative phrases were also found in the definition zone. The only missing concepts - comprehensive health check was captured in the description

zone. In addition, the new medical ontology covered more areas by retrieving more top level concepts (Table 4.5).

Besides the top level concepts, TMG also proposed another 290 further concept descriptions. However, these descriptive terms did not all appear in the new ontology's lower zones. In total, more than 70% of the TMG concepts appeared in the definition zone and description zone. This result may be because these descriptions in TMG were specifically defined medical procedures with a higher relevance for medical tourism. While the new ontology were chasing connections from three pairs of high level seeding words, some detailed terms representing specific procedures were missed. Further experiment revealed that these detailed terms could have been linked to the ontology by involving more specific terms as seeding words.

The completeness evaluations in both domains showed that the new ontology not only provided almost full (more than 90%) coverage of core concepts defined in the case study ontology, but also achieved much broader coverage. The broader coverage might contain redundant data which were not as precisely linked as the others, and could undermine the accuracy of the concepts defined and their reasoned relationship. Therefore, further conciseness evaluation was carried out.

4.2.2 Conciseness

As opposed to completeness, conciseness evaluation focuses on pruning the redundant definitions to form ontology with only the "necessary" concepts. Ontology such as GALEN, UNSPSC and SIC disallow redundant concepts and relations to form a "minimum set". However, such little tolerance of redundancy limits the ontology' coverage and may leave gaps among concepts. Strictly following the conciseness requirement may undermine the purpose of this research. Human orientated ontology needs to be concise, with human judgement used to bridge gaps. Ontology for ICT systems are not so memory limited, but find it much harder to bridge gaps lacking access to common sense. In order to satisfy both conciseness requirement and a broad coverage, the ontology developed in this research allows a certain level of redundancy (particularly in the description and connection zones) as long as the redundancy can be measured and acknowledged.

The measurable redundancy issue was addressed by utilising directional weight specified relationship. If a certain conciseness was required, those members who had weaker relationships to top zone concepts than the set value could be pruned from the final output. For example, without weight limitation, all top zone members are connected with each other in the new engineering ontology, thus any concept should lead to the network of all top zone members. However, if weight were limited to $f_d \leq 1.3$, it would show that a "honing" centred network only contained the six most "relevant" concepts (Figure 4.3). (raw data available in appendix 1.2.6)

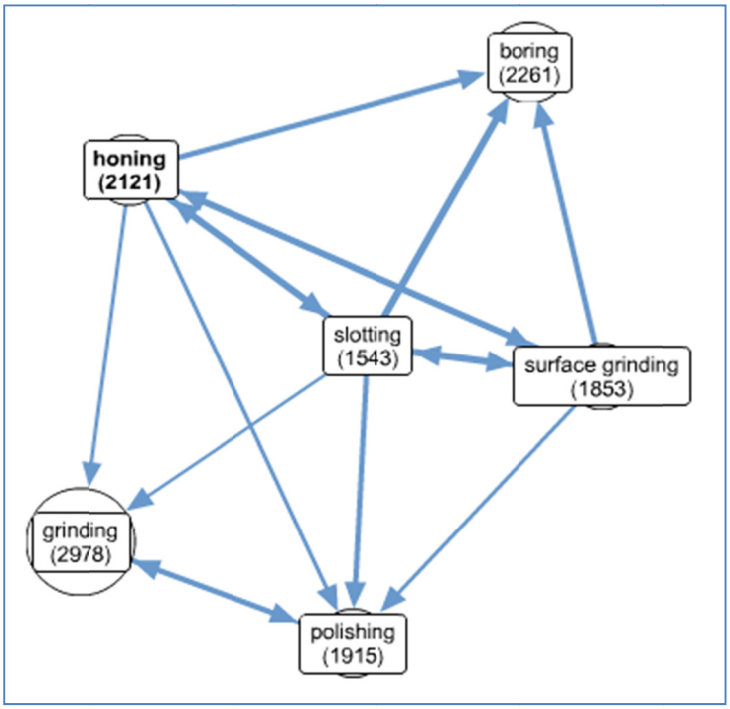


Figure 4.3: Illustration of conciseness alternation

Although the new ontology generated were not absolutely concise, their configurable output could be used to fulfil different conciseness requirement. In such a case, the ontology can keep the rich terms and relationships, and flexibly form a desirable ontology output according to the requirement for conciseness.

As well as helping achieve conciseness, the directional binary relationships were also important in maintaining the consistency in the ontology.

4.2.3 Consistency

Consistency evaluates whether the concepts and relations are defined consistently in the ontology. Consistency is proven if any concept in the ontology is only defined by its own particular relationships with others. In a consistent ontology, there should be

no contradictory knowledge that can be inferred from such relationships (Gómez-Pérez et al., 2004).

4.2.3.1 Concepts and relationships consistency

The new ontology generated had quantified the centralities for concepts, and specified weight and direction for the binary relationships. With the new methodology proposed any concept was defined by its centrality, inbound relationships from connected terms, outbound relationship to these terms and their centralities. Contradiction could only exist if two concepts shared exact values in all of the above four factors. In fact, concepts in the new derived ontology do not even share any two of these factors.

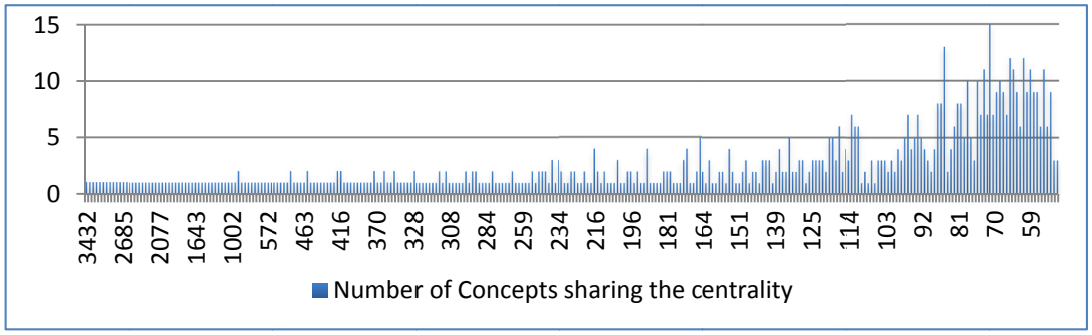


Figure 4.4: Centralities Frequency in the engineering top two zones

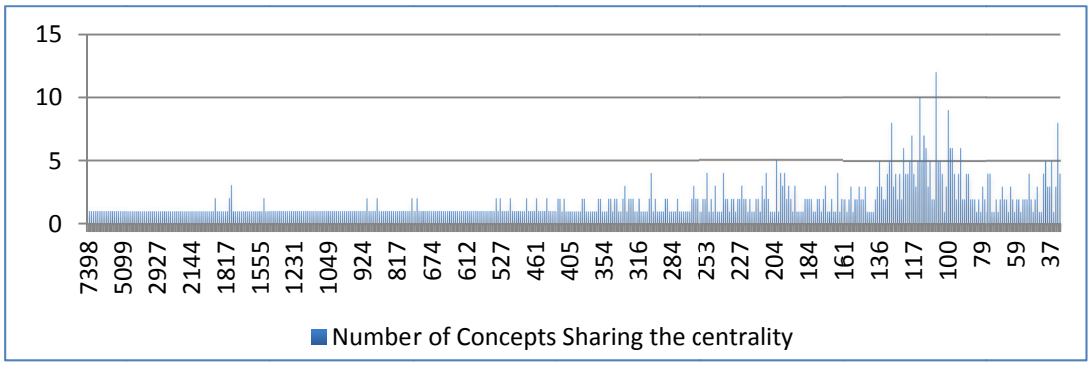


Figure 4.5: Centralities Frequency in the medical top two zones

Figure 4.4 and 4.5 show the number of concepts sharing the same centrality value in both the derived ontology. The X-axis marks the unique centrality values obtained in the ontology, and the Y-axis shows how many concepts are sharing this centrality. In the ideal circumstance, any centrality value should be possessed by only one concept (y-axis values = 1), so that centrality solely could consistently define different concepts. On the contrary, if centrality values are shared by many concepts (y-axis values > 1), it could fail the concepts consistency by only considering centrality.

It is clear that the majority of the concepts can be identified solely by its centrality (y-axis values = 1), especially those concepts have high centrality values (in top zone and description zone). Although there were members with the same centrality, their centrality origins (links from descriptive terms) and the relationship to those descriptive members varied significantly (raw data available in appendix 1). The quantified centrality and directional binary relationships eliminated any possible contradictions in concept definition in the derived ontology. Thus we can conclude the concepts and relationships are consistently defined.

Looking further into consistency through the building process: can the corpus structure and the zones be built consistently and defined in the same manner?

4.2.3.2 Corpus consistency and Seeding words fault tolerance

If the seeding words are misleading, this may lead to different corpus structure (section 3.2). Therefore, the research was designed to have three pairs of seeding

words to reasonably tolerate one pair of faulty seeding words. This design requirement was implemented in both the engineering and medical ontology building experiment in order to maintain the consistency of the corpus constructed.

In the engineering experiment, the original seeding words "turning" and "cutting" were not strictly specialised engineering words, as they are also popular generic terms describing other areas. These seeding words could mislead the corpus into other domains. However the final network centralised on the manufacturing domain even with these potentially misleading seed words.

In order to verify the fault tolerance of the design, extra experiments were conducted with "bad" pair(s) of seeding words in engineering ontology building. Low centralised terms such as "hobbing" and "brazing" were selected with the expectation that they would not be highly centralised concepts even though they were the seeding words. As a result, they only appeared at the bottom area of the description zone, and the distribution had the same form and the definition zone members stayed the same (raw data available in appendix 1.4, 1.5 and 1.6). This demonstrates resilience in the constructed ontology.

Similarly, the medical domain was tested with the description zone members "knee replacement" and "breast reduction", which were instances from top level concepts (they are also procedures nominated by TMG as a second tier description of major categories). These seeding words with two other top level pairs produced the same definition zone members (raw data available in appendix 1.7).

It seems that a pair of bad seeding words did not affect the network generated significantly. These seeding words were also tested by pairing with other seeding words, and this did not affect the network either.

Further fault tolerance experiments showed that the more concentrated a domain was, the more faults could be tolerated. For instance, the engineering ontology network was much looser than the medical network (discussion in Appendix 3,2), and thus required two pairs of "good" seeding words (from concepts zone and the top of description zone) to keep the domain on target. In the much tighter focused medical domain, one pair of definition zone members and any pair of description zone members would provide the same result. However, both networks required at least one pair of seeding words from the top zone. If both of the non-misleading keywords were from the description zone, the constructed corpus may appear ill defined with the focus on other domains that are not targeted.

Therefore, this approach could provide corpus consistency as long as domain experts can provide two pair of domain specified keywords in the Delphi method stage (keywords obtained from experts).

Consistency in concept definition and corpus building reflect that the domain focused terms and relationship can be extracted and constitute a network structure consistently. There is still uncertainty that networks in different domains may not fit the hierarchical structure according to the rules defined (tri-sectional zone definition).

4.2.3.3 Zone definition consistency

A zone definition consistency should also be examined as another aspect of structural consistency. Experiments in both medical and engineering domains demonstrated that zones could be defined based on consistent analytical methods – centrality analysis, closeness analysis and betweenness analysis. Through similar analysis, the new ontology network could establish a similar tri-sectional understanding, which structurally represented three types of concepts according to their representativeness of the domain. Each zone could be represented by a similar mathematical trendline according to their concepts centralities (detailed discussion in appendix 3.2). These trendlines followed the same distribution in both ontology building experiments.

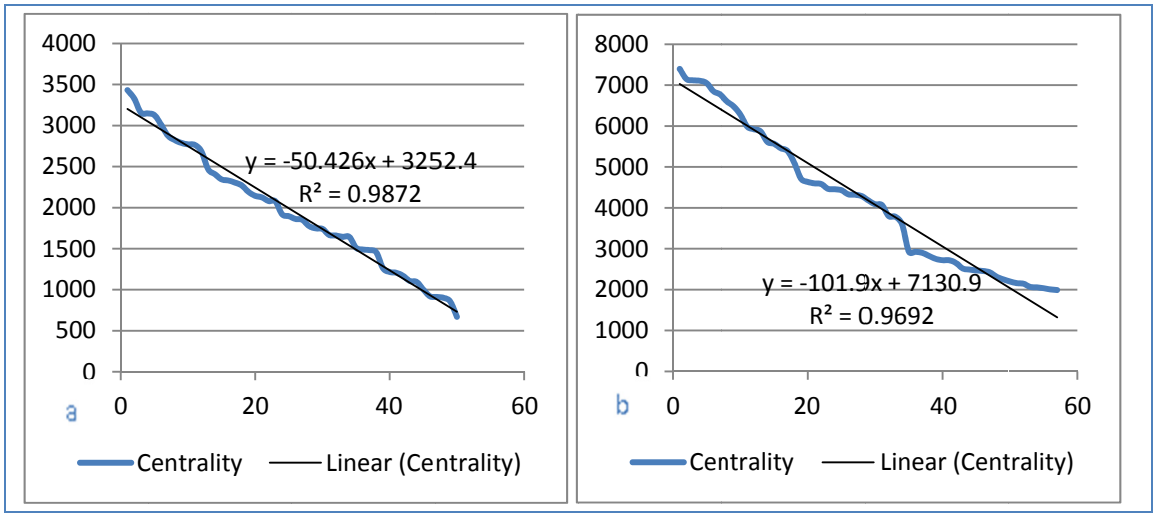


Figure 4.6: Centrality trend of the top zone members

For example, Figure 4.6 demonstrated the centrality distribution trend of engineering top zone members (Figure 4.6.a) and medical top zone members (Figure 4.6.b). Both of them, statistically speaking, strictly follow ($R^2 = 0.9872$ and $R^2 = 0.9692$) a

straight line regression ($y = -50.423x + 3252.4$ and $y = -101.9x + 7130.9$). The regeneration of the same regression suggests that SEA ontology top zone members follow a steady law in terms of their centralities.

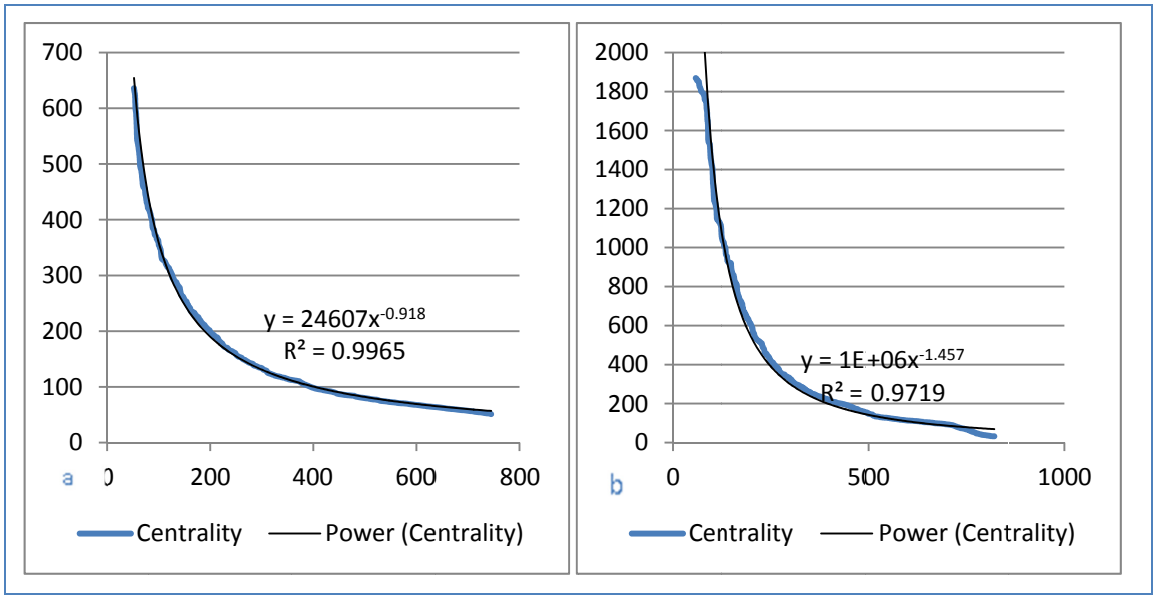


Figure 4.7: Centrality trend of gradient change zone members

After the top zone, both engineering (Figure 4.7.a) and medical (Figure 4.7.b) ontology' gradient zone members' centralities also strictly ($R^2 = 0.9965$ and $R^2 = 0.9719$) follow a similar type of power function trendline ($y = 24697x^{-0.918}$ and $y = 100000x^{-1.457}$).

As expected, long tail zone members of engineering (Figure 4.8.a) and medical (Figure 4.8.b) ontology also produced similar type of power function trendline ($y = 2E + 06x^{-1.571}$ and $y = 3E + 34x^{-11.29}$) for their centralities. It is observed that the long tail zone centrality distributions were much flatter than the gradient change zone, although both zones centrality increment curve followed power

function distribution. The flatness ($(\Delta y/\Delta x)$ highlights the characteristics of the long tail zones members: compairing with other zone members, they are extremely lowly centralised (less representative) members in the whole ontology network, hence their possible room for centrality change is limited.

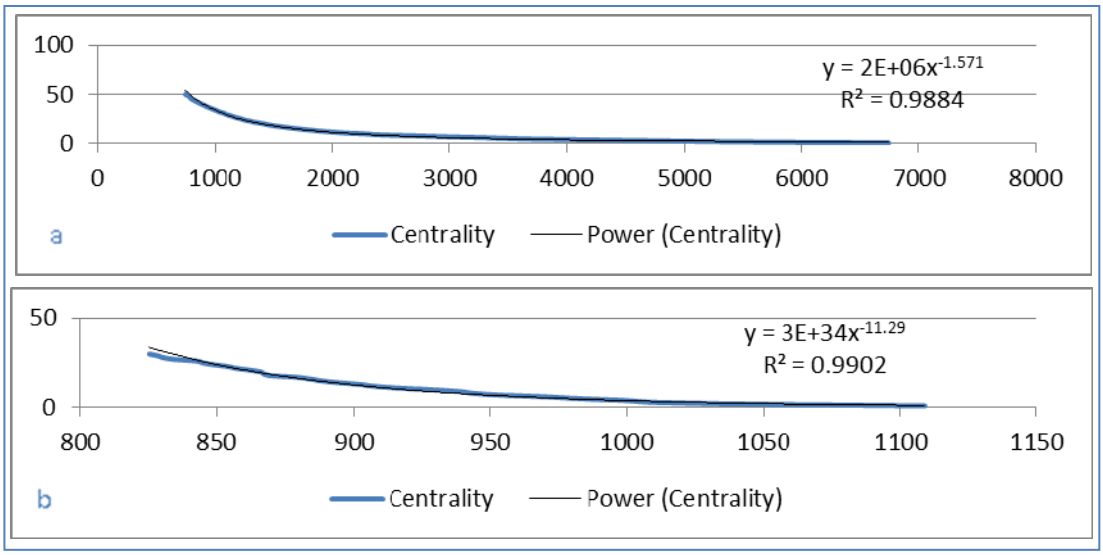


Figure 4.8: Centrality increment trend of long tail zone members

Beside centrality distribution, increment analysis on centralities ($f_{ci}(\Delta y/\Delta x)$) of different zone members also revealed the same tri-sectional zone definition. Figure 4.9 and 4.10 demonstrated the centrality increment differences among three zones of medical ontology and engineering ontology (increment distribution of top zones and gradient change zone in Figure 4.9.a, increment distribution of gradient change zone and long tail zone in Figure 4.9.b. Y axis represents centrality increment, while X axis represents keywords ID decreasingly ordered by centrality).

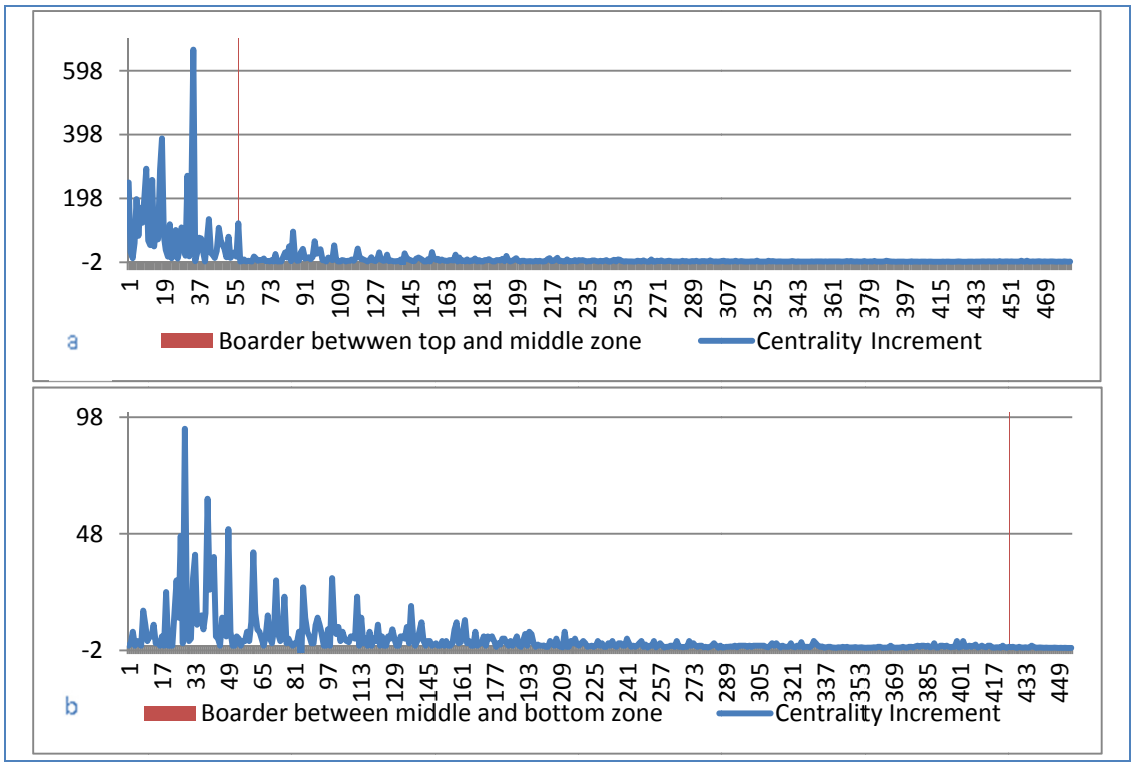


Figure 4.9: Centrality increment analysis for medical ontology

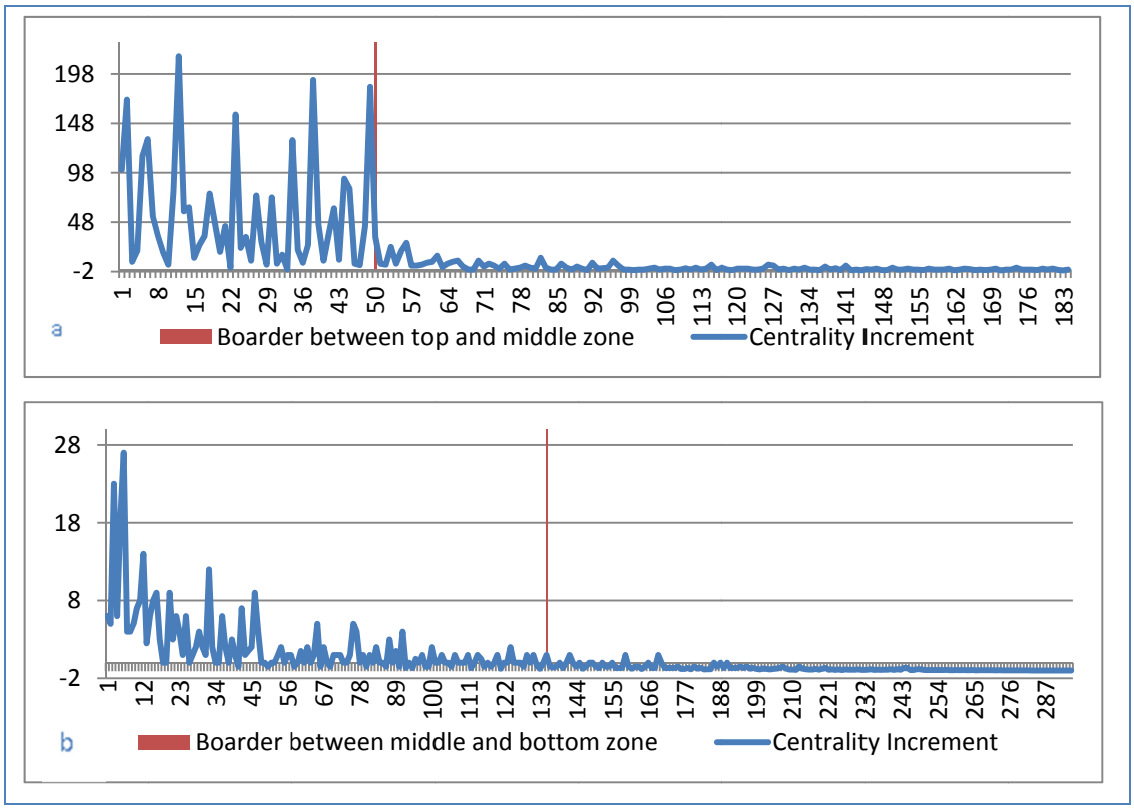


Figure 4.10: Centrality increment analysis for engineering ontology

Considering the plot of centrality increment as a series of non-periodic waves, it is observed that both ontology uncovered similar characteristics of their three zones: top zone members clearly showed much larger amplitudes than the gradient change zone members (Figure 4.9.a and 4.10.a); at the same time gradient zone members' increment vibrates much more than long tail zone; and the long tail zone members' increment hardly vibrated.

Similar distinguish increment changes of medical and engineering ontology further strengthened tri-sectional zone definitions (Figure 4.11), and reflected the consistency of proposed cutting-off points (details available in appendix 3.5).

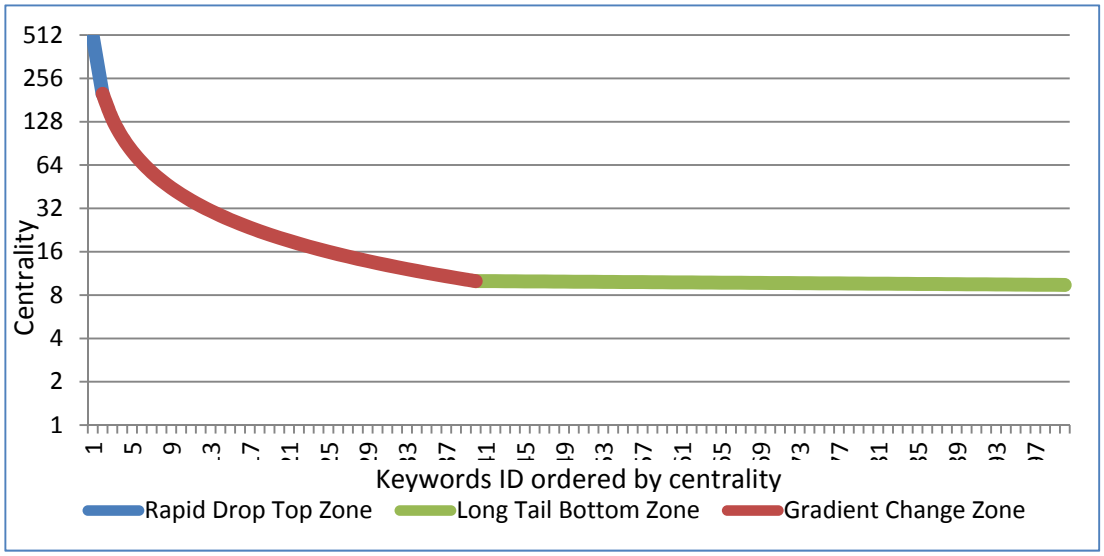


Figure 4.11: Tri-sectional trend of centrality distribution

Further experiments have been undertaken in other domains to verify whether their corpus would follow a similar tri-section division as Figure 4.11 illustrated. These experiments related to:

- Online security ontology for E-Security research in the International Digital Laboratory (raw data available in appendix 1.8, 1.9)
- An ICT ontology for the West Midlands ICT Clusters (raw data available in appendix 1.10)
- Business management ontology for MAS (Manufacturing Advisory Service) (raw data available in appendix 1.11)

All of these studies also formed a similar ontological structure with a tri-sectional division. It appears that the new approach can build a corpus in different domains with consistent structures and zones as defined in the engineering and medical examples.

In integrating the structural evaluation and content evaluation, the new ontology met the requirement of the research question - "quickly, reliably and economically generate ontology for a specific domain or domains, that can provide the breadth and depth of coverage required".

However, the research question posed also had a practical requirement – "for automated systems". The derived ontology need to be further evaluated in the application environment.

4.3 PRACTICAL EVALUATION

Structure and content are the evaluations usually performed by ontology builders or domain experts. If ontology are derived for a particular application, it would be more important that the ontology is evaluated by the actual users in the application

environment. The derived ontology for this research were built to solve practical problems in information categorisation for TMG and WMCCM.

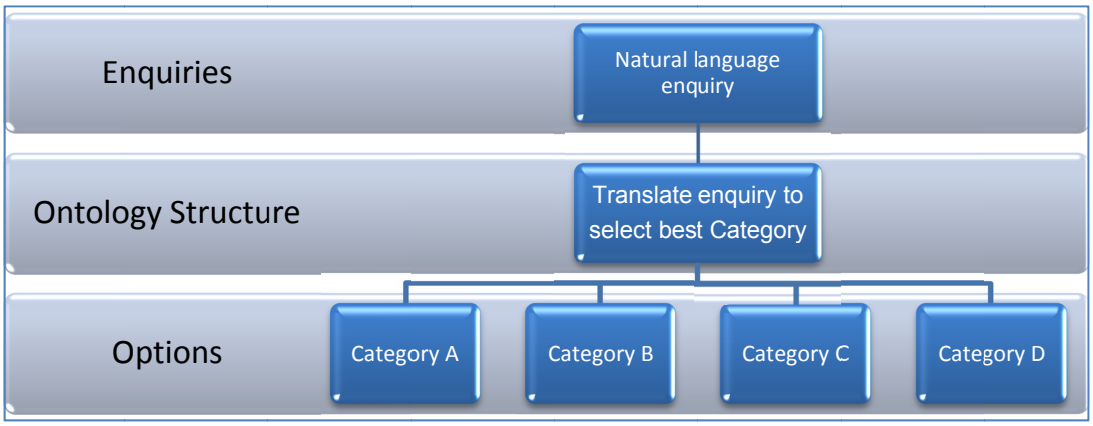


Figure 4.12: The ontology application environment

Thus, practical evaluation was carried out using an information system (Figure 4.12) to assess whether they would function as an information categorisation filter as desired. Natural language enquiries were fed tnto the system. As soon as the enquiry was received, the system used the ontology structure to analyse the enquiry and tried to tag the enquiry with ontology concepts. As a result, the tags indicated which category this enquiry should belong to.

The filtering process was reasoned by reverse application to the ontology structure as demonstrated by Figure 4.13. A natural language enquiry may have descriptions that contain terms stored within the ontology structure. These terms may cover different aspects of a domain (or even cover multi domains), and they may not be representative of the domain. The derived ontology can locate these terms in its structure (at the outskirt connection zone), and track the appropriate paths (via the

middle connection zone) to those highly representative concepts (in the core/top zone). Through such filtering mechanism, enquiries in natural language without clear indication of domain concepts could be tagged with representative concepts.

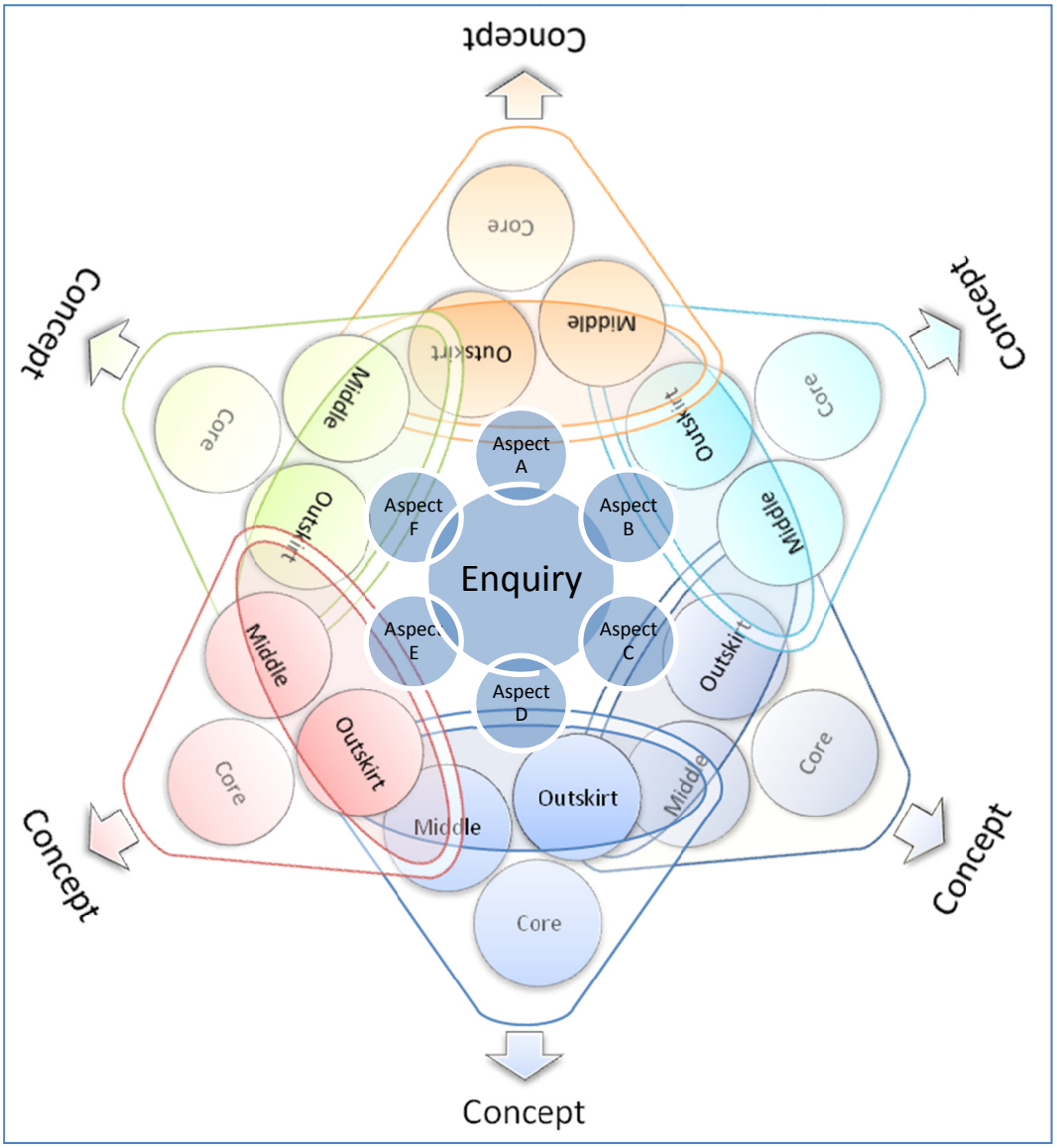


Figure 4.13: The filter mechanism

The derived ontology replaced the existing ontology (definition zone members replaced the top level categories, description zone members replaced 2nd level categories), and the system efficiency was assessed.

A period of five days monitoring was implemented for the original engineering ontology used by WMCCM and the engineering ontology developed by SEA. More than 1,000 pieces of information (engineering tenders) were processed through the system every day.

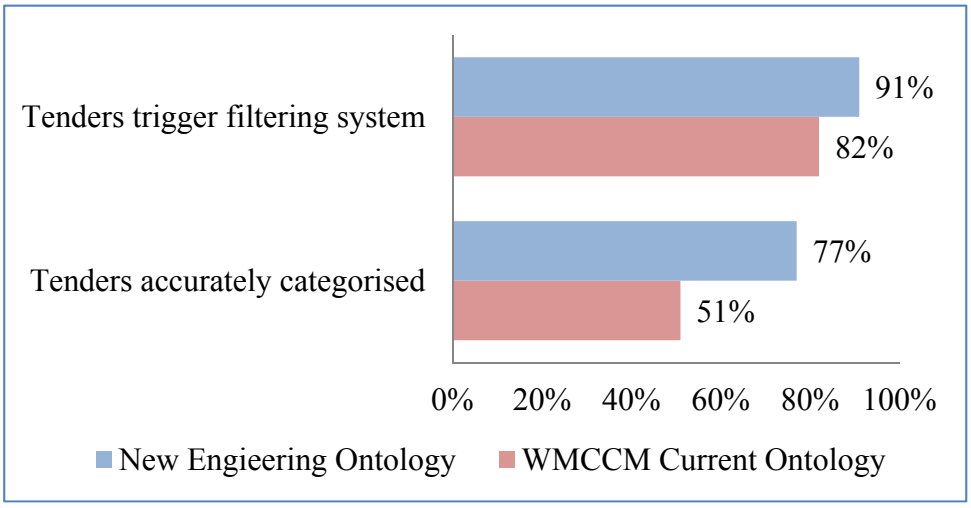


Figure 4.14: Illustration of practical evaluation on new engineering ontology

Figure 4.14 demonstrates that the filtering system has been improved by adopting the new engineering ontology:

- 82% of the information stimulated the filtering system via the existing WMCCM ontology. This percentage may seem high, but in practice this leads to a severe problem due to the lack of broader coverage. The WMCCM system imports around 100 tenders a day, 18% ignored tenders means that an experts' contribution is required to manually categorise more than 18 tenders per day.
- Among those filtered elements, only 50% of the information was correctly categorised due to insufficient internal relationships

+ The new ontology was trigged by more than 91% of the input information elements.

+ Among this categorised information, 77% had appropriate suggestions, although some of them may have extra "noisy" suggestions.

Twenty thousand potential customer enquiries from over a five year period to TMG was fed the system, both TMG's ontology and the new derived ontology separately processed the enquiries to test their practical efficiency.

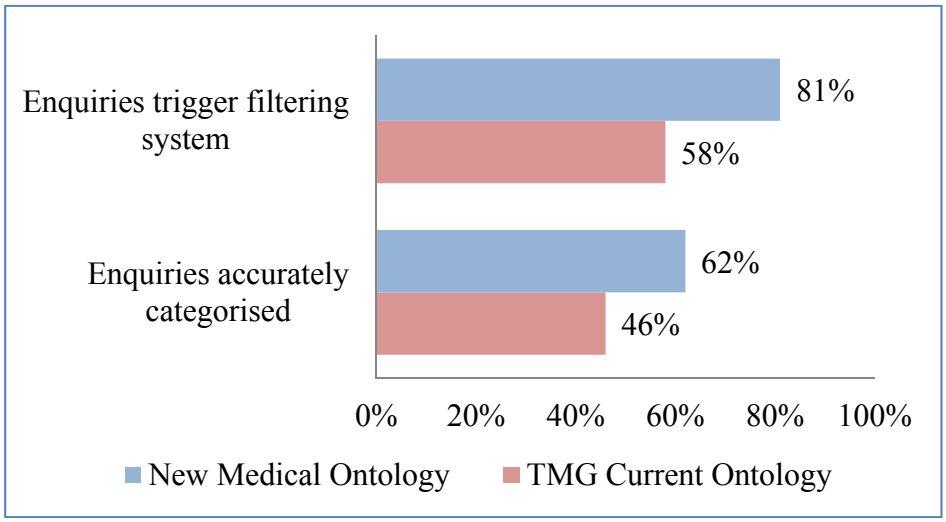


Figure 4.15: Illustration of practical evaluation on new medical ontology

Once more, the new ontology derived from SEA possessed clear advantages as Figure 4.15 demonstrates:

- Because the TMG's ontology had a limited breadth, it could only categorise less than 60% of the enquiries and correctly tag less than 50% of them.

+ The newly derived ontology was trigged by more than 80% of total enquiries, and more than 60% of the enquiries were correctly tagged.

However, directly implementing the new ontology may not meet practical requirement. The author, TMG and R. Huckman (Associate professor in Business Administration in Harvard Business School - HBS) studied TMG's customer enquires, and agreed that inheriting TMG's ontology top level concepts as their target categorisation. In this case, the new derived ontology was fitted to TMG's ontology by adding the new concepts and relationships as descriptions. Thus the original TMG top level concepts were turned into conceptual clusters with support from the newly derived medical ontology. These clusters benefited from the rich concepts and weight specified directional relationships produced by the new ontology. As a result, the combined ontology provided the same level of performance as solely implementing the new ontology, but categorised information was based on TMG's proposal. Similar integration was also tested on WMCCM ontology, and the combined ontology performance was at the same level as the new built ontology.

This practical evaluation proved that the new derived ontology can be fitted to the desired automated system and provided better categorisation results. More importantly, the new ontology could be fitted to an existing fixed ontology by adding the generated rich concepts and relationships as conceptual descriptions (Such descriptions only supplement additional terms and relationships without changing the ontological structure). This success implies that the new derived ontology could improve the performance of automated systems of appropriate architecture. They may be applied to multiple practical application environments.

4.4 SUMMARY

The ontology evaluation has sought to prove that SEA is capable of engineering stable and consistently structured ontology. These ontology were contextually complete and consistent, but they may not have been highly concise, because of the coverage focused nature of this research. Moreover, these ontology can function as desired in the real world application environment.

So far, SEA was primarily a prototyping of methods based on the integration of a key process from other methodologies. SEA will be more applicable to similar research if it is able to be generalised as a formal ontology engineering methodology.

CHAPTER 5: METHODOLOGY GENERALISATION AND FORMALISATION

The evolution of SEA derived medical and engineering ontology has revealed that the SEA is capable of generating new ontology in a fast, economic and reliable way. However, whether SEA could be extended to the wider context and used by other ontology builders as a methodology still requires further investigation and evaluation. This chapter explores how SEA could be generalised and formalised to a general ontology engineering methodology.

Yin (2009) has argued that a methodology should involve both specific cases and general facts, and could be derived from either deductive or inductive approaches. A deductive approach normally follows a conscious path from a generic fact to a specific case. Conversely, the inductive approach moves from a specific case to a general fact. It is more suitable to adopt the inductive process to generalise and formalise of SEA, since the SEA approach was derived from two specific cases – medical and engineering domains.

SEA was also a specific case from the ontology builders' perspectives: SEA has only been applied by the author. It is reasonable to question whether it could be used by other ontology builders.

Furthermore, SEA is a newly proposed ontology engineering method. It is, again, a "specific case" of ontology engineering methodologies, and requires further evaluation as to its completeness for being a formal methodology.

Therefore, to generalise the current approach to a formalised methodology, three aspects are considered:

1. Generalisability of the resulting ontology (a.k.a. external validity): Is it possible to derive ontology for other domains via this approach? And can the derived ontology be reused by other ontology or applications?
2. Methodology applicability to other researchers: Is it possible for other researchers to adopt the same approach for their ontology building project?
3. Formalisation towards methodological requirements (a.k.a. internal validity): Is it possible for the approach adopted to fulfil the existing ontology engineering methodological requirements?

5.1 GENERALISIBILITY ON ONTOLOGY ENGINEERING RESULTS

Generalisation of the research output is "concerned with the application of research results to cases or situations beyond those examined in the study" (Collis and Hussey, 2009), since the change of "cases or situation" could result in variations from those in the examined cases. A quantitative method may help to generalise or extend the

results obtained to other situations, as the result from quantitative data collection and analysis support the ascertaining of the magnitude of variations (Kumar, 2005).

Moreover, quantitative research methods may provide credible descriptions (characteristics) of how certain instances occur (Wellington and Szczerbiński, 2007), thus the characteristics of specific cases (derived medical and engineering ontology) may be used as the reference for the generalisability to other cases.

The generalisability of this research output focuses on whether the pattern observed in this research can be presented in wider use cases (Easterby-Smith et al., 2002). In both the medical and engineering subject areas examined, the same patterns have been found in terms of corpus constitution and ontological analysis. The pilot study in E-security, business management, and the ICT sectors also validated that such patterns exist in their domain areas. Although five cases may still not be sufficient enough to declare that such corpus patterns are applicable to any given subject area, they demonstrate that the same pattern should be generalisable because of its characteristics of: hospitability, flexibility and reusability.

5.1.1 Hospitability

Hospitability in this research means a methodology that can be applied to other domains. SEA exhibits hospitability in both its source coverage and the semantic relationships identified.

The use of the Google index as the sources should be equally applicable to other domains, as it provides a broad coverage of most domains (as discussed in section 3.2). The Google index also possesses advantages from the "latest information" perspective, since it is updated with "new knowledge" more frequently than many traditional sources, even compared to specialised online databases (Falagas et al., 2008). There is a worrying trend by information providers at the moment to close off their sites from Google indexing, but whether they can maintain such barriers is still open to question. Also, inevitably, there is a leakage of information from these sites into the indexed domain.

The use of semantic relatedness in producing related terms should be equally effective in other domains, since co-appearance of terms exists in any subject areas which have sufficient coverage of its content. Driven by semantic relatedness, the snowball sampling mechanism could collect large numbers of related terms in different domains to form a corpus with similar network structures to those derived for the medical and engineering domains demonstrated. These networks may suggest the same tri-sectional segmentation, even though their cut-off points vary according to their detailed internal relationship differences.

The corpus construction consistency is valuable, as ontology builders can reuse the same processing algorithms for different domains. It could further allow the same ontological analysis methods to thoroughly explore the internal structure and form

ontology outputs towards a commonly adopted hierarchy format (zone definition consistency).

However, there may be limitations in the domain hospitality due to the characteristics of the search engine index and semantic relatedness relationships within that domain. Retrieving semantic related terms generally available for a domain online reflect a general understanding of the domain. However, some ontology may require non-semantic related relationship from very particular perspectives. For example, a highly customised thesaurus ontology engineering ontology may be required by a company to reflect their special understanding on the part of the engineering domain. In this case, general understandings of the domain may not be suitable for the company's specific needs. Therefore, the method derived may not be hospitable for developing a highly customised ontology for an organisation's purely internal (non-general) perspective.

Based on the research undertaken, it seems reasonable to conclude that SEA approach should be capable of generating a full or partial ontology in a wide range of domains. Although there are difficulties in building highly customised ontology, the characteristic of hospitality can still be identified in the SEA approach.

5.1.2 Flexibility

SEA also provides a high degree of flexibility. Flexibility in this research refers to the ability to flexibly populate different concepts within a domain corpus,

search/view the ontology from differing viewpoints (i.e. from different application requirements) and join it to other ontology if necessary.

SEA's ontology supports flexible output: A highly connected network, especially the fully connected top zone, provides opportunities to centralise the ontology output on any given concept or conceptual cluster. Thus, different ontology outputs can be produced from each individual observing viewpoint (from the central concepts to the rest of the members). Moreover, the connectivity between concepts could be customised by limiting the relationship direction, weight or even relevant distance. Being able to emphasize given concepts aids the output in communicating/linking with other existing ontology systems. For example, the derived engineering ontology was centralised on the existing WMCCM concepts.

Secondly, SEA provides a flexible constitution for the resulting ontology: The whole ontology constitution and structure can be customised by changing the representative seeding words selected. The experiments conducted and described in chapter three sections 3.1-3.2 have shown that the domain focus can be tuned by using different seeding words. The more general seeding words produce a less focused corpus, while specialised seeding words generated a more concentrated corpus with a narrow but deep domain coverage.

In this research, seeding words from the medical domain were more specialised than from the engineering domain, therefore the resulting medical network was more focused. A more focused network is also expressed by a higher network density in

general, tighter network structure in the lower zones, and fewer members in the long tail section. Controlling the ontology corpus structure through seeding terms illustrates the flexibility that SEA can provide to meet specialised requirements.

Flexibility is also enhanced by the time efficiency of SEA approach. The automated process in SEA built medical and engineering test ontology in less than 12 hours each. This accelerated approach in realising of a new ontology significantly reduces the need to consider reusability, if we can quickly create a new customised one.

5.1.3 Reusability

Reusability has been an important concept in the ontology domain, largely because of the level of effort required to generate ontology. Even if the effort is not significantly reduced, there are still benefits in incorporating it.

The ontology built via SEA should accommodate sufficient concepts (known as "brokers") and relationships to connect to wider areas. 'Brokers' play the role of "bridging" members, so the new ontology can be referenced by other ontology. The derived ontology could be centralised through brokers with other ontology, so that they can become a subtree that permits directed fitting to other ontology (as recommended by SENSUS, see section 2.5).

The snowball sampling mechanism devised could be seen as ontology reuse mechanism: with every new round of experiment directly reusing previous concepts and relationships output as the new input. Similarly, multiple versions of ontology

built via this mechanism could reuse the existing concepts and relationships. Importantly semantic relationship may change over time as new knowledge is added to the domain, and it thus is desirable to add concepts and relationships as required to provide coverage of the new domain knowledge.

SEA provides richer internal relationships than existing methods, and more internal relationships may help existing ontology where they require more relationships for enhanced system operation. For example, a faceted classification requires different concepts and relationships from different observing "facets". The derived ontology can provide different ontological output from different "facets".

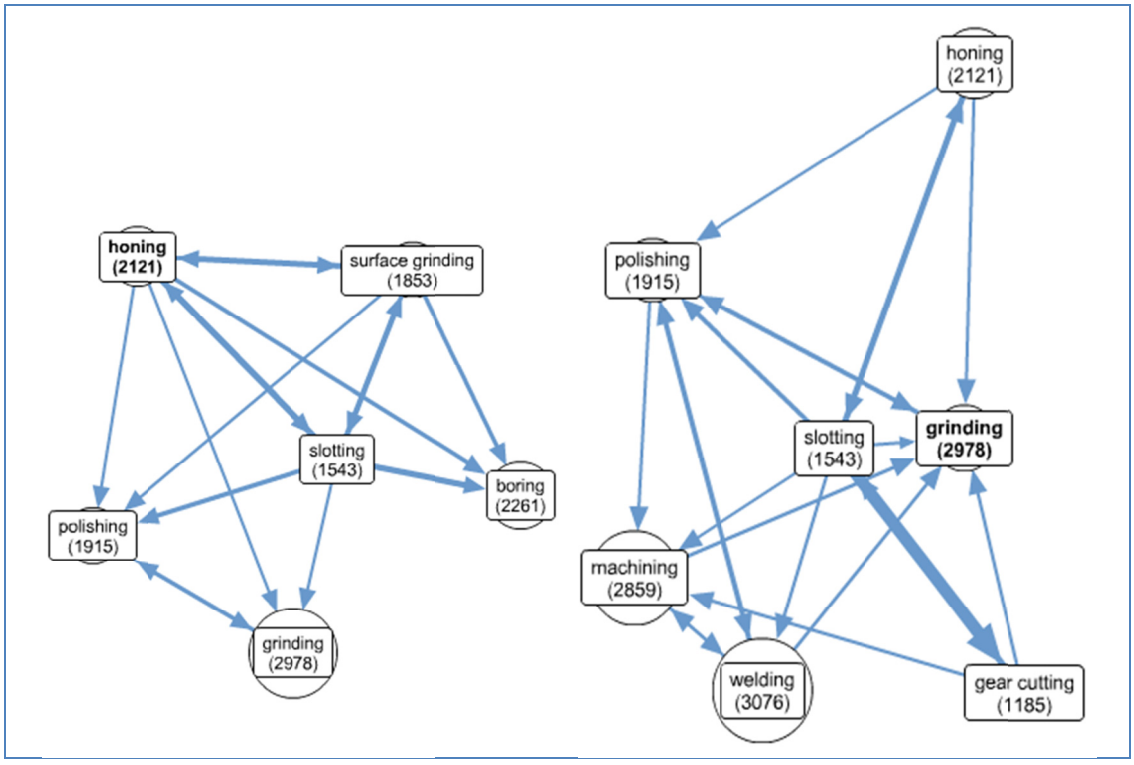


Figure 5.1: Sample facets information from engineering ontology

Figure 5.1 shows that when examining the concept "slotting", observing it from "honing" and "grinding" shows different "views". In this figure, the numbers in the brackets indicate the overall centrality in the whole engineering ontology; the length of the lines indicates the relevant distance from the observing concepts – those in the bold font; the thickness of the lines indicates the weight of the relationship; and the arrow indicates the direction of the relationship.

From the "honing" perspective (network on the left in Figure 5.1), "slotting" and "honing" have more "common" interest in surface treatment such as "surface grinding", "polishing" and "grinding". From the "grinding" perspective (network on the right in Figure 5.1), "slotting" and "grinding" also share some cutting commonalities besides surface grinding. Thus, this network also includes "gear cutting" and "welding" in the network.

In addition to concept changes (and changes around the new concept), relationships between the common concepts in both "facets" also changed: from the "honing" perspective (network on the left in Figure 5.1), the strongest relationship in the network is the honing-slotting relationship, and the relevant distance honing-slotting is slightly further than the distance of slotting-grinding. From the "grinding" perspective, a honing-slotting relationship may still be a strong but may no longer be the strongest in the network, as "gear cutting" clearly shows a much strong connection to "slotting"; at the same time, relevant distance honing-slotting is more than five times further than the distance of slotting-grinding.

The ability to group different concepts and relationships from different facets implies that the derived ontology may be reused by faceted systems. Furthermore, the weight specified relationships among concepts could also assist fuzzy ontology in their fuzzy matching logic.

Fuzzy logic represent relationships effectively when there are uncertainties among them (Lau, 2007). It requires numerical/quantified binary relationships that are normally presented in numbers between 0 to 1, rather than simple logical data that are normally presented as "linked" or "not linked" in a hierarchy ontology structure. Using numerical data enables the probability calculation of "matching" among concepts (Muhammad and Lipika, 2007). The key to enabling such probability calculation is to convert the logical type of data to a numerical type of data. Normally additional work is required to quantify a descriptive relationships based on their linguistic measures in the specified sources, such as a in specific text context from a particular organisation (Zhai et al., 2008), or in existing ontology (Lau, 2007).

SEA can directly produce numerical data for relationships by giving a numerical figure from the overall ontology perspective according to the specified weight between them, and by providing quantified data from every concept's individual perspective based on distance of relevance between them.

Table 5.1 showed that the numeric relationship ($f_d(m,k)$) is provided by derived engineering ontology from "tool grinding" to two other concepts. In this case, the relevant distance has been converted to quantitative data between zero to one

($f_d(\text{tool grinding}, \text{honing}) \approx 0.04$ and $f_d(\text{tool grinding}, \text{folding}) \approx 0.05$), thus these figures can be referenced directly (with no extra work) by fuzzy ontology to assign "tool grinding's" "likelihood" probability to "honing" and "folding".

Table 5.1: Sample fuzzy ontology from engineering ontology

Seeding Words (k)	Predict(m)	$f_c(m,k)$	$f_c(m)$	$f_d(m,k)$
tool grinding	Honing	83	2121	0.039132
tool grinding	Folding	58	1131	0.051282

In summary, the characteristics of hospitality, flexibility and reusability have offered SEA potential generalisability to other domains and compatibility to wider areas of applications.

5.2 GENERALISABILITY FOR ONTOLOGY BUILDERS

Generalisation should also concern whether the research approach can be learnt and used easily and economically by other ontology developer (Ward-Schofield, 1993).

The generalisability of SEA is partially reflected by its economic building processes. As discussed (section 2.2, 2.3), the heavy reliance on domain experts is a barrier to reusing many existing ontology engineering methodologies. SEA has reduced the initial knowledge acquisition to only collecting three pairs of keywords, and has automated the remaining processes via a set of software programmes (from forming the ontology structure via corpus construction to providing ontology output through ontological analysis). By applying this SEA approach, small organisations or even individuals can afford to build large ontology.

Generalisability for other ontology builders is also shown through reduced time consumption on ontology building. The feasibility study is capable of prototyping target ontology within an hour, with the full building processes taking less than 12 hours. This allows ontology builders to spend more time on alternate ontology configuration, ontology structure customisation and output validation according to their application specification.

The generalisability of SEA is additionally strengthened by providing a collaborative and distributed ontology engineering capability. As the source, Google Sets is accessible by various ontology builders from different geographical locations. At a certain period, the same seeding words will form the same ontology structure (as discussed in consistency evaluation section 4.2.3). Such consistency allows multiple ontology building processes to work collaboratively on the same project regardless of their locations.

For example, each ontology builder may be in charge of part of the ontology building, such as building a corpus from one pair of seeding words. The collaborative and distributed ontology building may further lead to greater potential for "cloud" ontology building. As a result, more complex ontology, such as ontology covering multi-facets or multi-domains, could be generated in a fast, economic and consistent manner.

5.3 ONTOLOGY ENGINEERING METHODOLOGY GENERALISATION

The discussion in the previous section has largely focused on the external generalisation of the SEA methodology. Generalisability also requires support from internal validation (Ward-Schofield, 1993) – generalisation and formalisation of the research life cycle. This view is shared by Lee and Baskerville, and is seen as "the other" part of generalisability validation (Lee and Baskerville, 2003).

The discussion of several existing ontology engineering methodologies (in section 2.3) has identified that lack of the detail on the key techniques, inadequate life cycle coverage and poor project management guidelines might undermine the practical application of such methodologies. Fernández-López (2002) and Gómez-Pérez (2003) draw a similar conclusion: even though these ontology engineering methodologies have been "externally validated" (had been applied in multiple domains or in multiple projects), their overall generalisability is limited by internal validation.

Table 5.2 summarises a comparison among ontology engineering methodologies, which are examined from an internal validation perspective. In the table, each row represents a methodological feature. Fields highlighted in blue indicate that the corresponding methodology can provide sufficient detail in the category proposed towards generalisability.

Table 5.2: Comparison of existing ontology methodologies (Davies et al., 2003, Fernández-López and Gómez-Pérez, 2002, Gómez-Pérez et al., 2004)

Methodology / Feature	Cyc	TOVE	KACTUS	METHON-TOLOGY	SENSUS	On-To-Knowledge
Detail of the methodology	Low	Low	Low	Medium	Low	Low
Recommendations for knowledge formalization	NA	Logic	Logic	NA	Semantic networks	NA
Life cycle and project management	Little	Partial	Yes	Yes	Partial	Yes
Strategy for identifying concepts	NA	Yes	Yes	Yes	Yes	Yes

In general, ontology development processes are based on a particular domain or application environments and their conditions. De Hoog summarised "*that it is difficult to value the generalisability of a methodology from its own perspective*" (De Hoog, 1998), so objective validation should be carried out by comparing a methodology to more mature or more generally applicable methodologies. SEA started from a prototype which contained mixture of processes from several ontology engineering methodologies. These processes need to be reassessed for generalisability.

As a sub-division of knowledge engineering, methodologies for ontology engineering should have a direct heritage from the methodologies for knowledge engineering (Waterman, 1986, Wielinga et al., 1992, Gómez-Pérez and

Rojas-Amaya, 1999, Nicola et al., 2009). In turn most knowledge engineering methodologies inherit or adapt processes from the methodologies used in software engineering (Downs et al., 1992, Pressman, 2010), such as IEEE (Institute of Electrical and Electronics Engineers) software engineering model (IEEE, 1996). This is used as a reference example for evaluating the derived methodology for generalisation in this research.

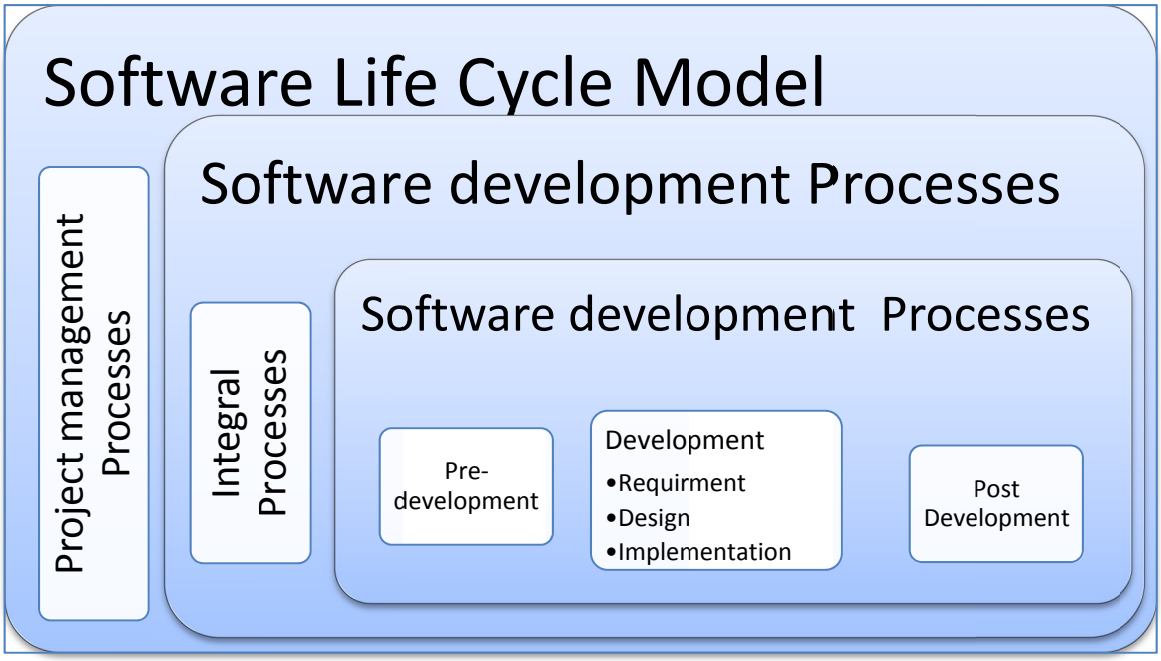


Figure 5.2: IEEE software development methodology

Compared with ontology engineering, Figure 5.2 shows that the IEEE advocates a more mature model which manages and controls processes during the development of software. Fernández-López and colleagues suggested that ontology engineering approaches should inherit these processes in order to form better ontology engineering methodologies (Fernández-López and Gómez-Pérez, 2002). These processes include:

1. Life cycle and project management

   On a higher level, ontology engineering methodology should have a life cycle and project management plan to define how different processes are initiated, scheduled and controlled through the life time of the ontology building (such as in METHONTOLOGY).

2. Pre-development process

   A methodology requires pre-development studies before the actual ontology development, such as studies on the application environment and feasibility studies on the development processes (such as in On-To-Knowledge).

3. Development process

   The actual ontology development processes and the detailed techniques employed should be clearly described. This is important to enable reuse, continuous improvement and collaborative development.

4. Post development process

   After building the ontology, post development processes help to implement the ontology developed. They transform the resulting ontology into a compatible format for the target application, and install the ontology in the application environment.

5. Integral process

   Finally, a completed methodology should be able to verify its result. Integral processes evaluate the ontology developed. Additionally, they could also validate the whole development process. Such an evaluation may uncover issues in the processes, and stimulate maintenance or further development actions

SEA was initially proposed on the basis of SENSUS's processes, since it best met the identified needs. However, SENSUS's processes missed project management, pre-development, post-development processes and integral process (ontology evaluation). Even the key development processes lack details on the techniques, so that they could not be replicated.

In order to meet the IEEE's requirement for a methodology, SEA should detail the techniques utilised, address the missing stages, identify life cycle and project management.

5.3.1 Development Process and Detailed Techniques

As shown in Figure 5.3, the development process starts from a structure similar to SENSUS by identifying seeding words and linking them to a knowledge base.

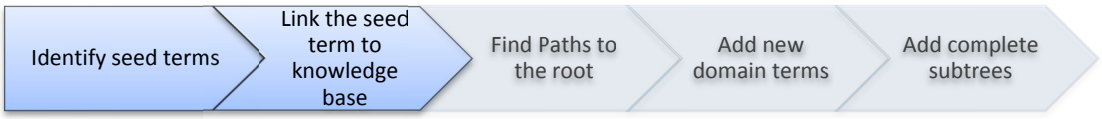


Figure 5.3: Normal SENSUS Approach

5.3.1.1 Seeding words

This research proposed two methods for seeding word(s) generation: extracting seeding words from existing cases, and obtaining seeding words from domain experts (Delphi method)

1. Extracting seeding words by ontology builders from existing cases

Terms and the relations between them were considered from both the focal domain and the application environment, since the definition of terms in natural language may be different due to the meaning they have in the different environment. This research conducted studies in both engineering and medical domains associated with WMCCM and TMG application environment to find semantically related terms.

2. Extracting seeding words from domain experts

To maximise the opportunity for picking objective seeding words and minimise bias, the Delphi method was applied to collect seeding words from domain experts. The opinions of domain experts are collected from their particular viewpoints individually, instead of collection on consulting a large group of people at the same time. Therefore the process can be conducted "economically and quickly", and avoided the problem of disagreement among a large group of individuals

A combination of seeding words generation techniques provided techniques (details in section 3.1.2) for the seeding terms extraction process (Figure 5.4).

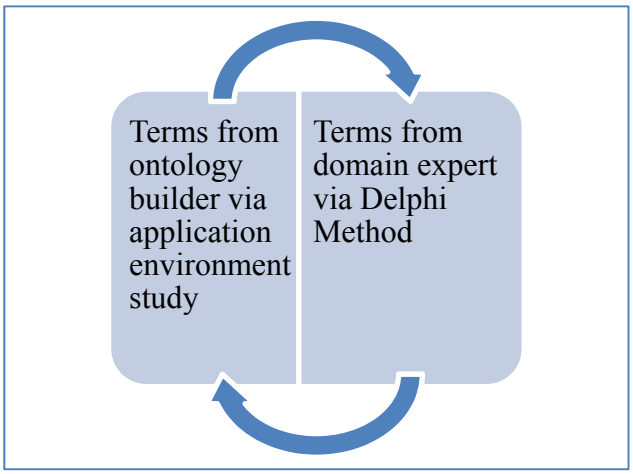


Figure 5.4: Detailed techniques for seeding words extraction

5.3.1.2 Linking to the knowledge base

The process of linking seeding words to the knowledge base depends on a clear understanding of the expected relationship between the seeding word and the corresponding terms in the knowledge base. This normally involves manually linking the appropriate members in the knowledge base to the seeding words. These linkages may be proposed by experts from their understanding of the domain.

For this research, the binary relationships between concepts were proposed specifically as a semantic based word clustering (word co-occurrence in the knowledge base) relationship - semantic relatedness (Section 3.2). The semantic relatedness based web application (Google Sets) is proposed as a key tool (illustrated in Figure 5.5) to enable linking seeding terms to the knowledge base (Google index). This creates a mechanism to convert a human dependant "linking seeding words to the knowledge base" task to a machine executable task.

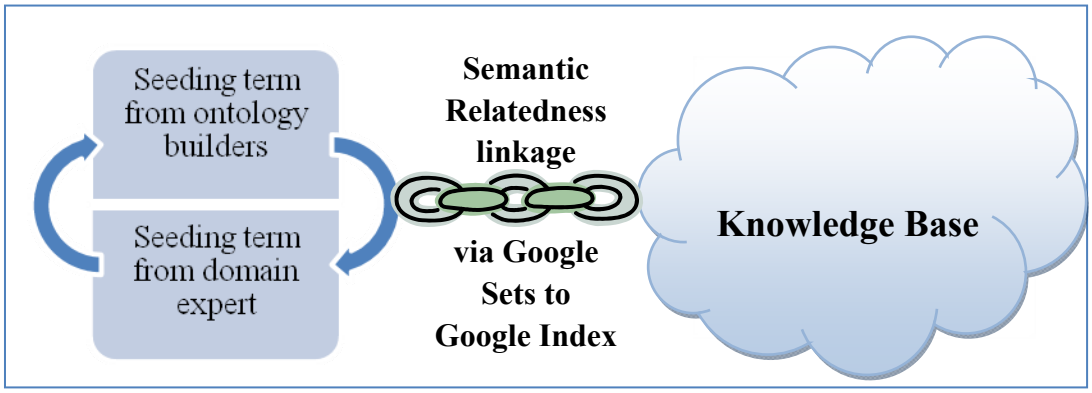


Figure 5.5: Linking seeding words to the knowledge base

Using Google Sets can generate many concepts on the path from the seeding words to the knowledge base, but not necessarily to the "root" of ontology as recommended by SENSUS. Hence, this requires SEA to define root members from a large collection of concepts, and to provide a corpus construction process.

5.3.1.3 Corpus Construction

A corpus construction process collects terms and relationships for the target ontology. In most ontology engineering methods, corpus construction processes are either conducted manually, such as Cyc, METHONTOLOGY, and SNOMED CT; or operated by traversing the knowledge bases structure, such as in SENSUS. Manual coding may possess advantages in knowledge conciseness, but it relies on domain experts. Traversal of an existing knowledge base could reduce the reliance on domain experts, but for any real size problem it is generally computationally inefficient and usually requires the addition of heuristic rules to make it realistic. Such heuristic

rules are often domain dependent and thus again require contributions from experts, who may not all agree.

This research tackled these issues by limiting the number of seeding words and searching only for their directly related terms/nodes in the search space without visiting all the nodes in the source. It then devised a snowball sampling mechanism to expand the scale of corpus construction. As a result, a large number of both generic and domain specific terms with their relationships can be generated to form the corpus of the target ontology (Figure 5.6).

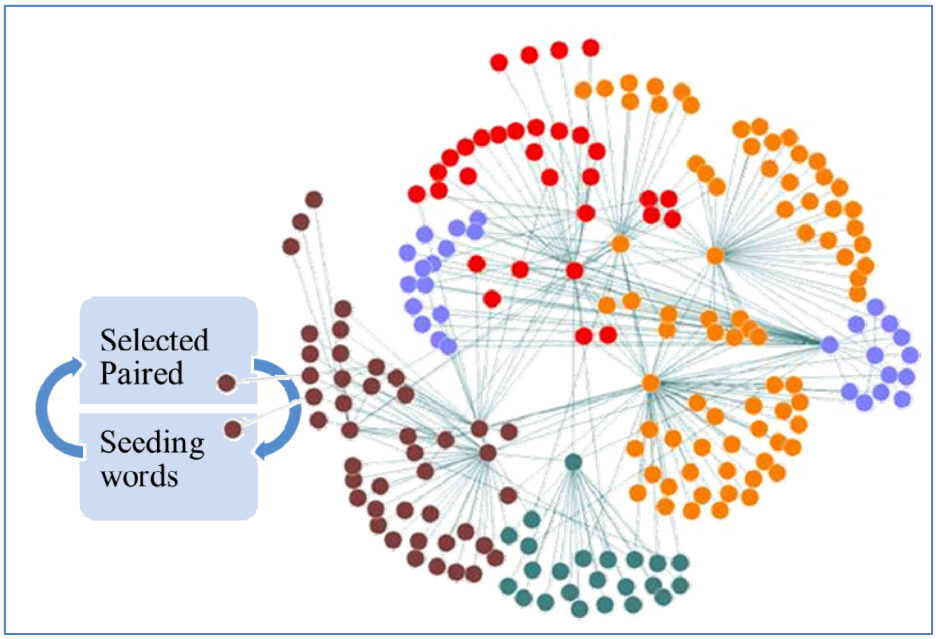


Figure 5.6: Illustration of detailed technique in corpus construction

Having built the main corpus, ontological analysis is used to identify the domain concepts and to clarify the internal structure. Methodologies such as Cyc,

On-To-Knowledge, METHONTOLOGY proposed ontological analysis processes as Figure 5.7 demonstrates:

1. to find the root of the focused domain and extract concepts (Maedche and Staab, 2000),
2. to understand structure (the relationships) of the ontology (Aussenac-Gilles et al., 2000, Faure and Poibeau, 2000), and
3. to prune the ontology output (Kietz et al., 2000).

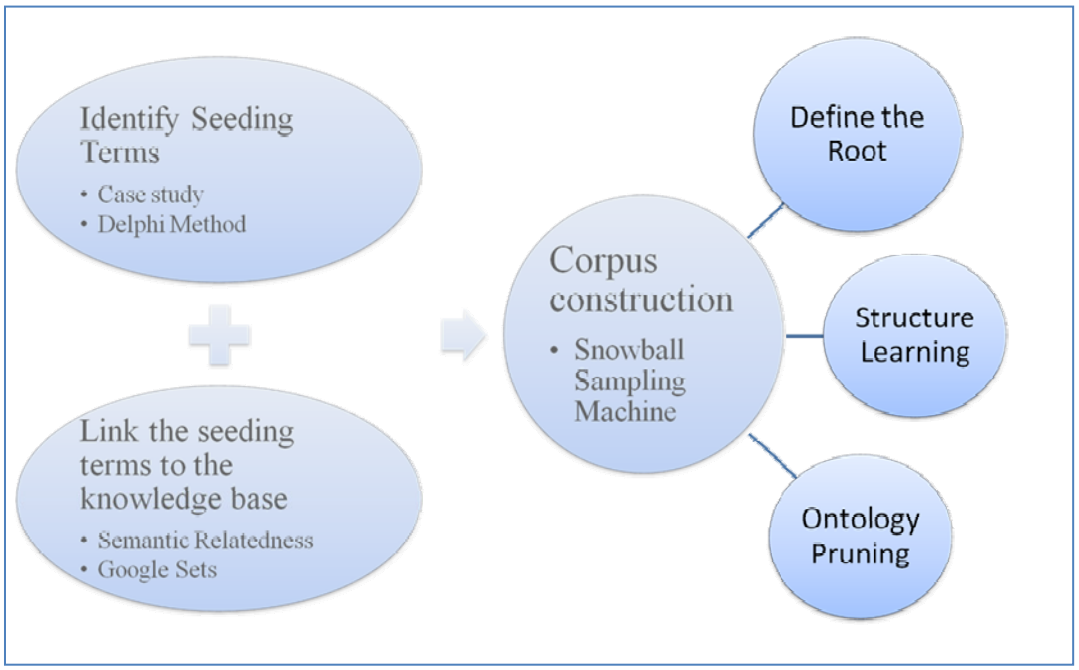


Figure 5.7: Detailed techniques for linking seeding words to the knowledge base

However, these processes are mostly descriptive, i.e.: they tell you what needs to be done rather than how to do it. The author suggests that since the corpus generated could be considered as a "social network" of concepts (built via snowball sampling of related words), social network analysis techniques can be used.

5.3.1.4 Define the root and Conceptual Learning

The root members of the ontology are those that can represent the network, and are more centrally located than the other members. Centrality analysis was used to assess every member's representativeness to the ontology domain by calculating the aggregation of their nominations in the network. It reveals that concept centrality is widely distributed, as members may be nominated by most of the concepts in a network or nominated only by one. The distinguishing difference on the concept centrality may identify highly centralised members in the ontological network, and these members may act as the core concepts in a derived ontology.

In addition to the core members in the definition zone, centrality analysis also identified and separated the less representative concepts (descriptive concepts of the core concepts) from the remotely connected concepts (least representative) in the connection zone.

5.3.1.5 Clarifying Structure

Starting from the definition zone members, binary relationships between concepts may be studied via closeness analysis. Closeness analysis provides insight into how concepts link to and mutually describe each other. It can provide a vector to determine how "far away" other concepts may be.

This analysis may also discover lower centralised conceptual clusters, called "subtrees" by SENSUS. In them, members may be tightly bound with each other to

represent a sub cluster concept. Locating lower centralised conceptual clusters enriches the resulting ontology by better structuring (and hence more efficient search) the problem domain.

However, sometimes conceptual clusters (or concepts) are not directly connected, and gaps can be left between concepts. Under such circumstances, closeness analysis can not fully explore the connectivity of the corpus. To tackle the disconnection, "Betweenness" analysis is introduced. "Betweenness" analysis locates the "brokers" for concept clusters to find alternative connections between unlinked concepts. Brokers can bridge concepts, but their existence also imply boundary between concepts. Drawing boundaries can assist ontology pruning in forming the final ontology.

5.3.1.6 Ontology Pruning

Ontology pruning plays the role of examining the whole ontology to eliminate poor paths and nodes, and finalise the ontology building process (Figure 5.8).

Firstly, peripheral player analysis applies external observation to define boundary players for the whole ontological network. It also conducts an internal observation to delineate concepts and conceptual clusters.

Secondly, network reach analysis helps to limit the corpus size by eliminating members which are "too far away" from core concepts. Together with peripheral player analysis, they shape the final ontology structure.

5.3.1.7 Summary of technique details

The previous discussion has sought to explain the techniques utilised, their place and function in the overall methodology and how they all operate to provide an economic, quick, reliable and repeatable methodology. The full structure of the derived methodology is shown in Figure 5.8.

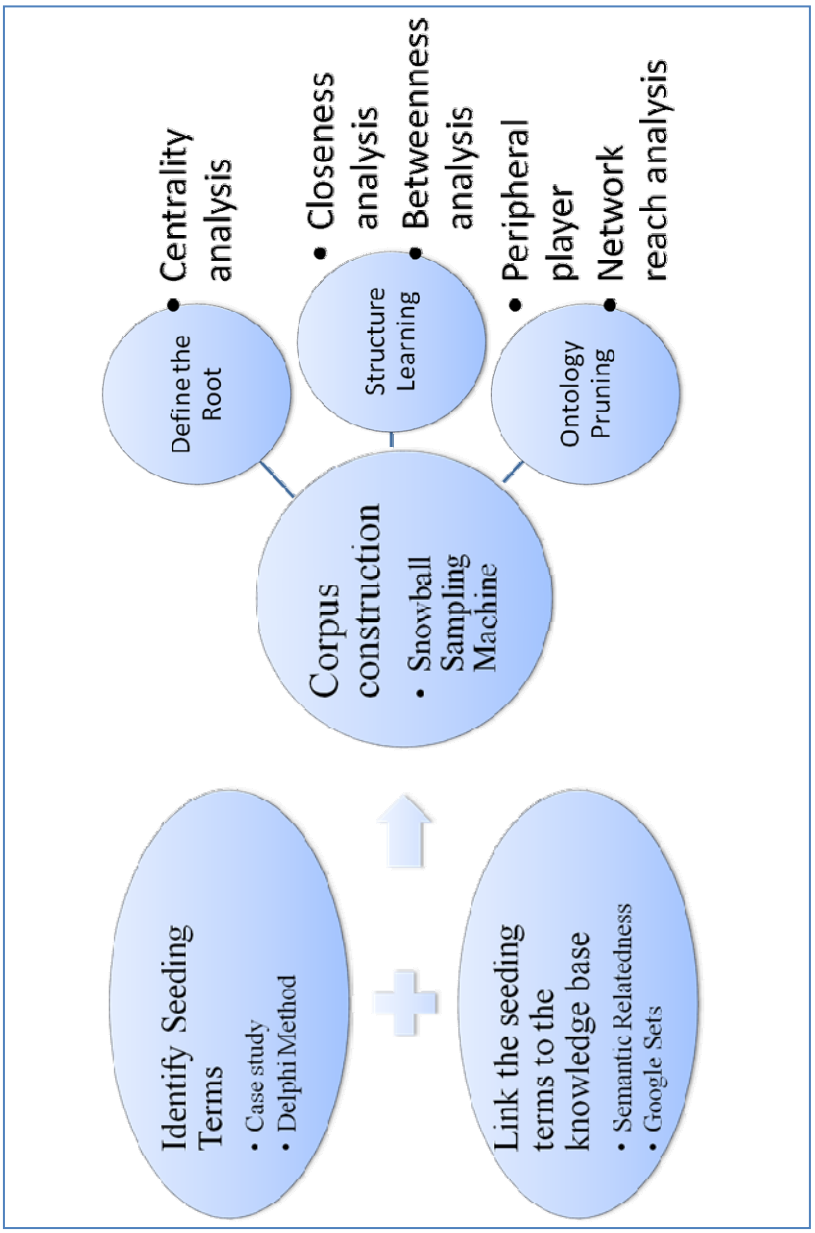


Figure 5.8: Details of the Techniques

5.3.2 Pre-development processes

The IEEE also recommended that a formal methodology should provide pre-development processes, such as application environment and feasibility studies to ensure the correct deployment of development processes. Pre-development processes for SEA have not been discussed as such so far.

5.3.2.1 Environment study

The environment study is initiated by investigating cases in the target domain and/or the seeding words collection and evaluation using Delphi methods. These techniques can produce seeding words for the corpus construction, but also set a vision for the target ontology by addressing the ontology requirement in the given environments (issues with current system and expected characteristics of the target ontology).

With a vision for the target ontology, a pre-development stage test can be conducted to assess the methodology's initial suitability to the ontology engineering requirement.

5.3.2.2 Feasibility study

The purpose of the feasibility study is to help define the settings required when executing the processes for the derived methodology. For example if the application requires a tight domain focus, the number of seeding words in a set and the sets of seeding words used will need to be larger. If one is looking for a system to be able to

better interpret more natural English in a domain, fewer seeding words and a greater number of snowballing runs will be more appropriate.

The feasibility study applies the selected seeding words and executes a shortened corpus building run (this research suggests an initial two rounds of snowball sampling). This quickly shows whether the keywords combination and experiment settings can generate suitable concepts and whether the results can be statistically appropriate for the target domain. The result will help ontology builders to assess if deploying SEA could be feasible, or what further decisions should be made. Such decisions may be to conduct the full scale exercise, modify the experiment settings, pick alternative seeding words, or maybe even abandon this approach as it is simply not producing expected results.

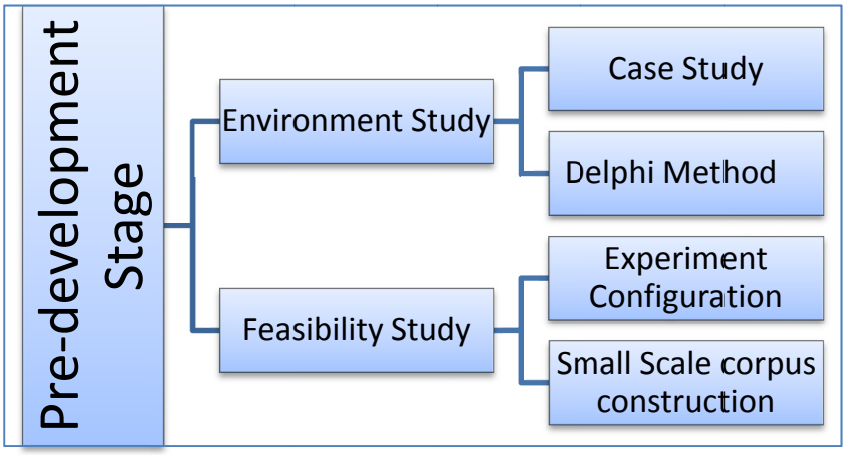


Figure 5.9: Pre-development stage techniques

The environment and feasibility study (Figure 5.9) help the SEA approach to assess the potential for success, and play a role of triggering the ontology development

processes within the whole ontology engineering life cycle. Having identified technique details in pre-development and development stage, SEA should also have a post development stage to support ontology implementation.

5.3.3 Post development process – Ontology Implementation

The ontology developed via SEA can be implemented in multiple ways (Figure 5.10):

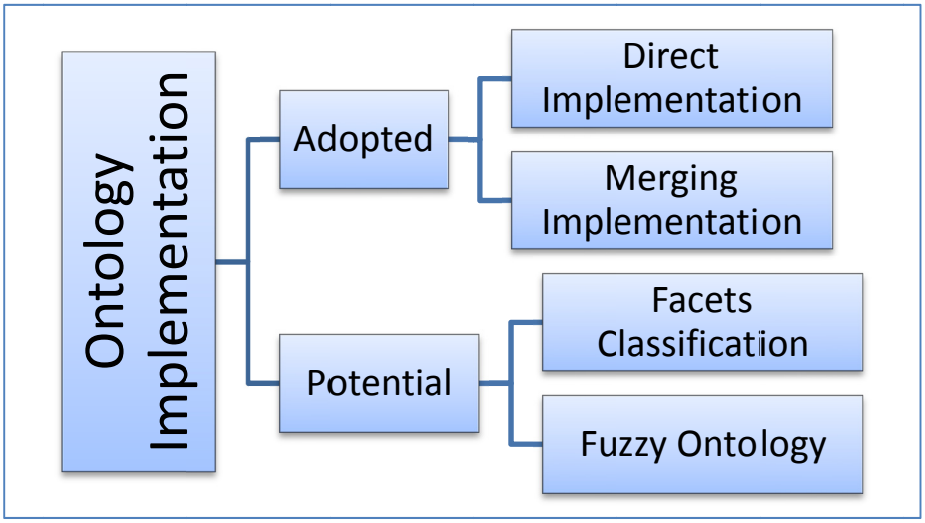


Figure 5.10: ontology implementation methods

1. Direct Implementation: Ontological analysis can form a hierarchical structure output, which can be graphical displayed and presented by OWL (Dean and Schreiber, 2003) for machine processing.
2. Merging: Ontology development can be merged with existing ontology: if the existing ontology requires its basic concepts to remain unchanged, the new ontology can be centralised on those concepts. Alternatively, the derived methodology could enrich the existing ontology structure by adding descriptive

concepts and relationships found by the SEA approach. This merging implementation was tested in the case study domains used in this research.

In addition, SEA can build a network structure and enable directional weight specified binary relationships, these can help expand the implementation methods:

3. A network structure allows ontology observation and output against given concepts from different perspectives. Considering the multiple output perspectives as facets to describe the given concepts, the ontology can be implemented as a facets classification system.
4. Specified weights for the relationships can be calculated and converted to the probability of "fuzzy matching" between concepts. Hence, the derived ontology may be potentially implemented as a fuzzy ontology system.

5.3.4 Ontology Evaluation

This research proposes to integrate several existing ontology evaluation processes (Figure 5.11). The evaluation measures the ontology structure, observes their content and tests their applicability in the application environment.

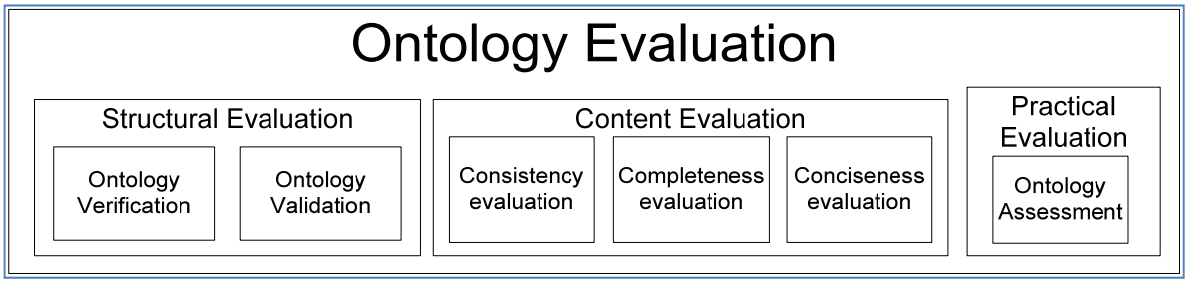


Figure 5.11: Ontology Evaluation during the Observation stage

Ontology evaluation finalises the ontology building processes, but this might not be the end of the whole ontology engineering life cycle. Further actions could be triggered by problems observed at the evaluation stage.

5.3.5 Life cycle and project management

A stage for reflection allows the researchers to reflect critically on the results of the evaluation, keep the whole life cycle sustainable and reveal the reasons behind any problems that may appear in the previous processes in the life cycle.

5.3.5.1 Reflection

Reflection may start from the evaluation results on the generated ontology: the issues that appear in a practical evaluation could result from an incompatible ontology implementation (post development process) for the application environment. Incompatibility may require reconfiguration of ontology output. For example, the initial medical ontology implementation for TMG did not satisfy the later requirement for a collaborative project (based on structure recommendations from Harvard Business School - HBS). Therefore, a new ontology was reconfigured to centralise around the agreed concepts in order to keep these concepts at the top level,

Such reconfiguration may involve altering the zone definitions and ontological analysis. This implies that issues appearing in the practical evaluation can be related to the ontology development stage settings. For example, unsatisfactory ontology structure and content may imply misleading/inappropriate seeding words. Or

sometimes new knowledge may emerge at a later stage, thus the development process may be required to include these changes (as in On-To-Knowledge's maintenance proposal).

It is also possible for the same issues to be traced back to the pre-development stage, for example, the initial feasibility study. The feasibility study may produce a smaller scale ontology corpus, which may demonstrate whether the seeding words combination can bring back the correct concepts. However, they may not be ideal seeding keywords for further expansion to meet the ontology requirement, as the result may not be statistically reliable for the desired domain. This could be an early sign of an unsuitable application environment, or may require a change in the ICT processing undertaken by the application.

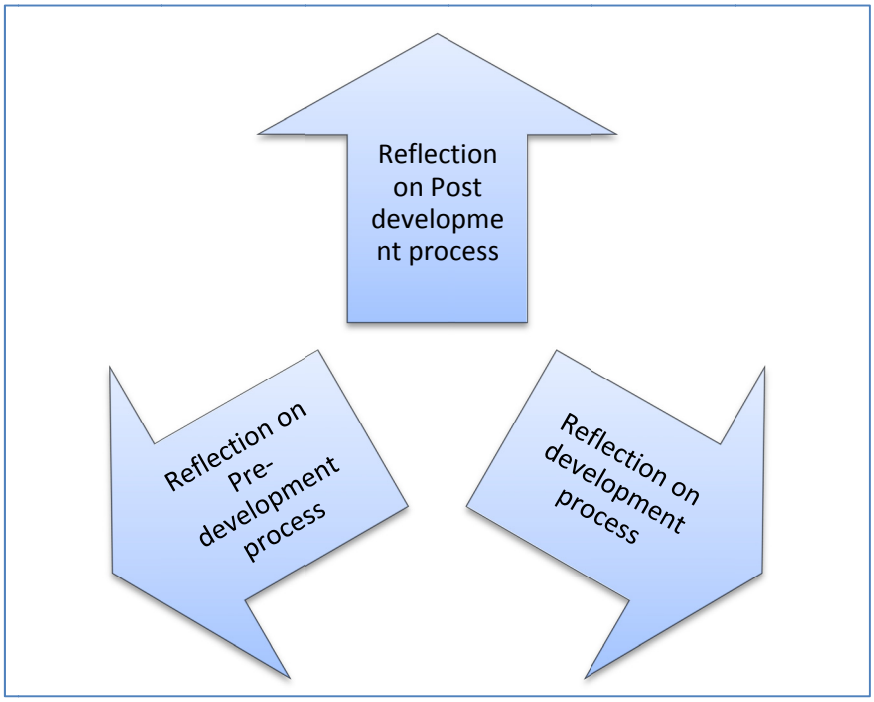


Figure 5.12: The Reflection Stage tasks

Reflection should be applied to all stages from pre-development to post development, (Figure 5.12) and the result of it may trigger further actions from the ontology development life cycle.

5.3.5.2 Life Cycle and Project management

Reflection can actually play the role of initiating a new cycle of ontology engineering (Hughes and William, 2001) and manage the actions on the processes in each development stage. The reflection stage turns a linear methodology process into a development cycle (Figure 5.13), and enables SEA to provide continuous improvement of the resulting ontology in terms of maintenance and information update. This helps meet the IEEE requirement for life cycle and project management in ontology engineering methodology.

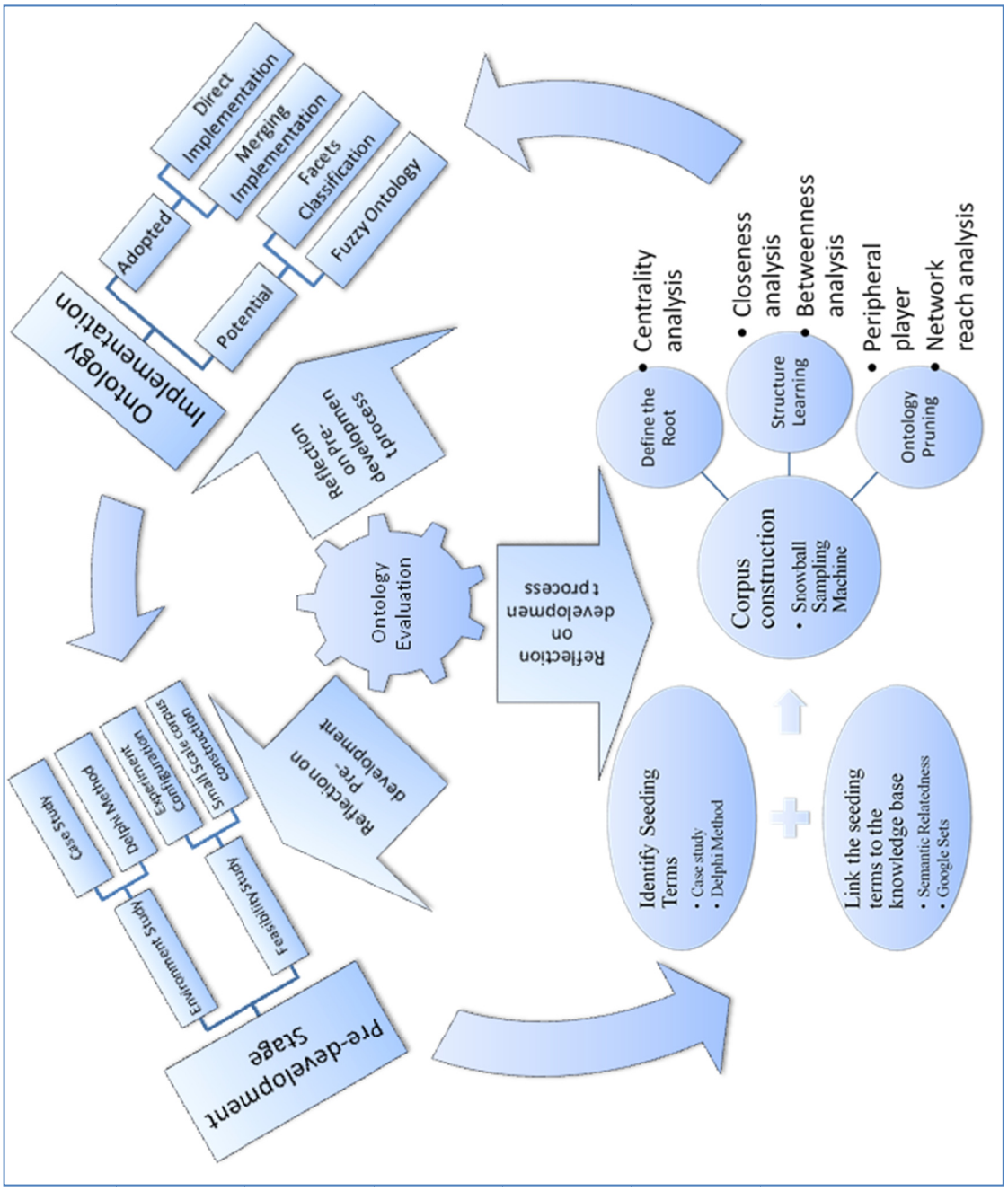


Figure 5.13: Ontology engineering life cycle with process management

5.4 SUMMARY

It appears that SEA could be extended to a wider range of ontology building needs from an external perspective, and validated to match the IEEE methodology requirements internally. In addition, it is economic and easy to use by other ontology builders. This new approach is not only a generalisable new methodology for ontology engineering, but also demonstrates extensive impact on wider subject areas.

CHAPTER 6: FURTHER DISCUSSIONS

The previous chapter has demonstrated that the SEA approach meets the requirement of "generalisability" as an ontology engineering methodology. The author believes that it is a "*methodology to quickly, reliably and economically generate ontology for a specific area or areas that can provide the breadth and depth of coverage required for many automated systems*".

This chapter provides discussion on aspects of SEA that have not been addressed in previous chapters. The author has decided that use of a tool can help encourage neutral thinking about these aspects was a more productive approach. A number of such tools are available with Edward De Bono's PMI (Plus, Minus, Interesting) methodology a prominent one (Bono, 2006), which examines the Plus, Minus and Interesting points regarding an issue or proposal. The tool basically tries to neutralise the rather impulsive decision making that normally occurs in humans. This trait is exemplified by the phenomena discovered in job interviews, namely that the impression to accept or reject a candidate is normally made in the first 30 seconds of the interview, with the rest of the time being spent on trying and justifying the decision (Guerrero et al., 1999).

Figure 6.1: Research findings organisation as PMI

The use of such a tool may improve the judgement and assessment of the research undertaken and the SEA methodology. Thus the various research issues are discussed in terms of "plus", "minus" and "interesting" points (Figure 6.1) regarding their impact on the utilisation of SEA methodology and the ontology derived from it.

6.1 "PLUS" FINDINGS

"Plus" findings show the positive impact that SEA may have on ontology engineering and wider fields. They could also indicate the contribution of this research to the knowledge domain.

6.1.1 Ontology definition

There are a number of definitions of ontology, some of which have been discussed in chapter 2. The derivation of ontology within SEA uses methods from both the ontology engineering and network analysis areas, and the derived ontology possess some different characteristics from other ontology, such as a larger number of concepts, the use of semantic relatedness, and a network type internal structure. These differences raise a concern as to whether SEA generated ontology still qualifies as ontology.

Ontology in knowledge engineering is recognised as an "explicit formal specifications of the terms in a domain and relations among them" (Gruber, 1993). Researchers who developed the Standard Upper Ontology (SUOWP, 2003) proposed a more practical explanation for ontology as "a set of concepts and relations that

describe a domain of interest". Both definitions highlight the fact that ontology should formalise concepts and relationships to represent the domain of interest.

controlled vocabularies	glossaries	thesauri	Informal is-a	Formal is-a	formal instances	frames	value restriction	general logical constraints

Figure 6.2: Ontology formalisation scale

Figure 6.2 shows a scale of possible formalisation (Lassila and McGuinness, 2001), from loosely controlled vocabulary to logically axiomitised taxonomy. Semantic relatedness can be understood as "is related to", which links concepts in a similar manner to a thesaurus (similar words). Meanwhile, "is related to" may not be as restricted as a "formal instances" of "frame" relationship such as in SIC. Thus, SEA's knowledge formalisation leans slightly towards the left half of the formalisation scale. SEA's formalisation lies between "thesauri" and "formal is-a". This suggests that the SEA ontology may be recognised as "lightweight" ontology, SEA has also developed other characteristics beyond the traditional ontology definition, such as a network structure with richer internal structure and configurable customised output.

6.1.1.1 A network structure and richer internal relationships

The ontology studied in this research are normally built upon a typical hierarchy structure in two dimensions, such as the GALEN ontology (Figure 2.3). Other ontology, such as SNOMED CT and UMLS, have increased their internal structure

richness by creating extra connections among concepts. These have had a limited impact on the significant enhancement of relationship richness. The internal structure of SNOMED CT, shown in Figure 6.3.a, shows that many of concepts are only linked to their direct "parent" and "child", as in a hierarchy structure.

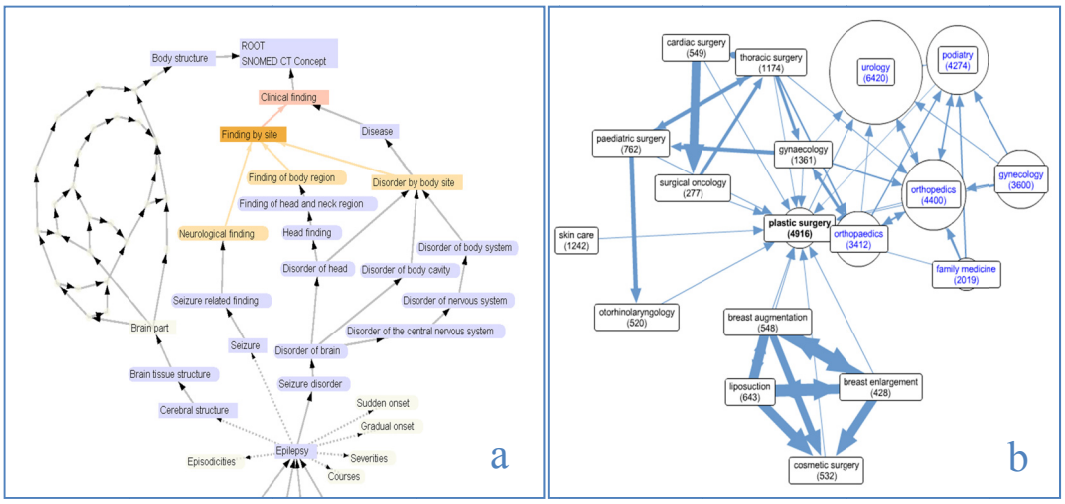


Figure 6.3: Examples of internal structure in SNOMED CT and SEA ontology

In comparison, SEA has exponentially increased the number of internal relationships and quantified them. Thus, the structure has grown into a multi-dimensional network structure (Figure 6.3.b), which substantially improves the correlation between concepts. The ability to generate multiple rich internal relationships is naturally limited in any approach which relies heavily on expert contribution, as a consequence of the "cost" of that expert contribution. Though SEA's internal structure is much richer and more complex than traditional ontology, this richness could be adjusted by setting lower or higher thresholds for quantified relationship (see section 4.2). Therefore, SEA's internal network structure can be simplified to a traditional hierarchy structured output if it is required for an application.

6.1.1.2 Configurable customised output

SNOMED CT provides subsets for different sub domains, as well as providing control for its users to form individual "FavorSet" – a favourite set of concepts and relationships. "FavorSet" allows users form a personal, needs oriented ontology. To achieve this users have to manually pick, organise and manage all their favourite concepts and relationships. SEA also supports the provision of a customised ontology output, and offers configurable ontology output in two further ways:

1. Centralising on required concepts and relationships

SEA ontology provides the advantages of a network structure, which enables multi-angle (multi-viewpoint) observation. This enables SEA to centralise the ontology output on required concepts and relationships (section 3.3).

2. Visualised control of concept representation level

An analysis of the network also offers a visualisation of concept centrality distribution. As Figure 6.4 shows, the distribution may naturally form a tri-sectional understanding, but cut-off points are controllable according to application specifications or users' individual requirement. As a result, more control can be passed on to the ontology builders to alter ontology output as practically required.

Richer internal relationships have also enabled SEA users to apply a series of statistical network analysis methods to help understand the network structure of the derived ontology.

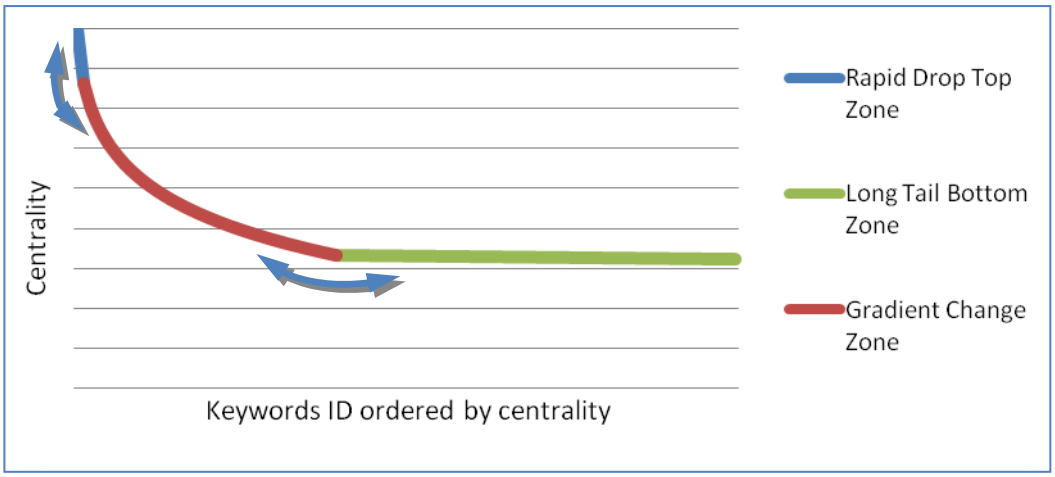


Figure 6.4: Illustration of centrality distribution

6.1.2 A "Statistical Network Analysis" approach

SEA's statistical network analysis monitors how a target concept may be described by a number of relevant concepts, and builds the relevant concepts into a "possible description network". SEA "understands" the context of the target concept in the network by identifying neighbouring phrases, and utilises statistical analysis to propose the most "accepted" description for the target term in the context. This statistical network analysis approach is new compared to the rule based mechanism adopted in traditional ontology.

This approach has been successfully applied in other fields. For instance, Google Translate (Google, 2011) adopts a similar approach to propose the "most accepted" translation by building and analysing a "possible translation network". Google Translate has the best performance, according to a machine translation evaluation

held by the US National Institute of Standards and Technology (NIST, 2005), although it does not apply the classic natural language rule based methods.

The Google Translate example shows that a statistical network analysis approach could provide better results in terms of information correlation identification than a natural language rule based mechanism. The same conclusion has been drawn after implementing the statistical network analysis based SEA generated ontology to the TMG and WMCCM application. The SEA generated ontology have provided more accurate reasoning between general information and professional terminology (section 4.3).

SEA's statistical network analysis mechanism could also benefit areas that require clarifying information correlation, such as natural language processing, machine learning and neural network studies. Identifying the "best" translation also applies to translating professional terminology to non-experts' vocabulary, which has been a main target for this research.

6.1.3 Bridging reusability and usability of knowledge

Professional terminology is often the reusable "representatives" of consensus domain knowledge. For example "*Rhinoplasty" is a professional term* in cosmetic treatment, while non-experts better understand vocabulary such as *"Nose Re-shaping"*. Different preferences for terminology between professionals and non-experts have created a gap between highly reusable knowledge and practically usable knowledge.

Such a gap is also reflected by the relative failure of the domain ontology (reusable knowledge) in classifying the user information for WMCCM and TMG. A reusability-usability matrix of the various ontology (Figure 6.5) demonstrates that usable application ontology (such as WMCCM and TMG) in the top left square do not share "borders" to directly interact with reusable ontology (such as GALEN and SIC) in the bottom right square. Usability and reusability can only be bridged either by expanding domain ontology leftwards to "generic terms" (the "green" route); or by identifying a highly usable section of domain ontology (the "blue" route).

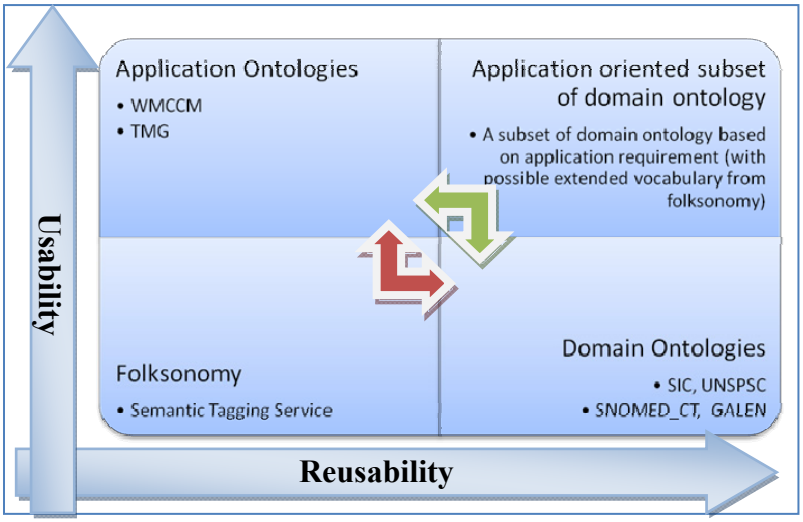


Figure 6.5: Matrix of ontology usability and reusability

The DOGMA approach (Jarrar and Meersman, 2009) bridges usability and reusability through the "green" route by pruning a subset of domain ontology that matches the usability needs of specific concepts and relationship requirements within a given application (Figure 6.6.a). Such an approach is "asking" non-experts (application ontology) to "speak" part of a professional language (domain ontology).

However, non-experts may purposely avoid using professional language (Powell, 1994). Even if they attempt to speak "a part of a professional language", such speech may not naturally reflect their specific intention commonly expressed in non-experts terms, due to the change of vocabulary (using unfamiliar professional terminology).

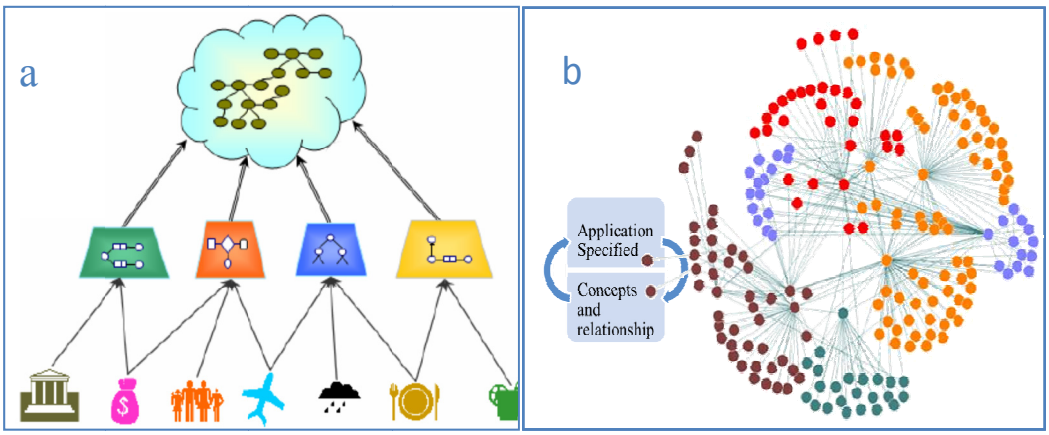


Figure 6.6: DOGMA approach vs. SEA approach

SEA provides an alternative "red" route (Figure 6.5). It identifies the application specified terms and relationships (usable knowledge, illustrated in blue rectangles as shown in Figure 6.6.b) in their natural forms, instead of seeking applicable parts from domain ontology. It then expands the knowledge coverage based on these specified terms by only retrieving relevant knowledge (semantically related terms and relationships). This ensures the usability of the derived expanded knowledge base. The expansion could build a concepts network (right side in Figure 6.6.b), which contains a mixture of concept descriptive terms that cover both the domain level (professional language) and application level (non-expert language). SEA's statistical network analysis could identify professional terminology from the

concepts network (Discussed in section 4.2 and 4.3). This allows SEA to be an alternative method for bridging the gaps between knowledge usability and reusability.

A broad coverage of generic terms and reasoning ability towards domain representatives, establishes a foundation for semantically tagging web content.

6.1.4 Semantic Tagging

Tagging unstructured web content, for example HTML pages, could provide machines with semantic metadata to help with tasks such as identifying relevant content, filtering out irrelevances, and summarising the essence information. In recent years, folksonomies have been utilised to tag web content (Kim et al., 2008), but they can be too ambiguous and lacking domain focus for general content tagging (see section 2.2.2). Moreover, there is still much web content that is not tagged yet.

SEA initially forms a "concept cloud" which is similar to a folksonomy to cover a large number of generic terms, and it also contains domain focused terms. The large number of generic terms increases the probability for SEA ontology to "capture" (through keywords matching) unstructured content. Once SEA locates the matching keywords in its derived ontology, the statistical network analysis helps identify the optional paths from these generic terms to the domain representatives (through configurable zone definitions) or given concepts (through customised outputs). Therefore, the SEA approach could support semantic tagging with a domain focus by

bridging the generic terms in the "concept cloud" to core domain concepts (see Figure 4.13).

SEA not only contributes a method to enable domain focused semantic tagging, but also supports multi-disciplinary tagging (section 4.3) with the ability of fuzzy matching (section 4.1.1.3).

6.1.5 Knowledge discovery

SEA generated ontology can be considered as the "universe", where concepts (stars) may be clustered by their relationships (gravity between stars) into different "galaxies" (Figure 6.7). Studying stars from different galaxies' viewpoints may identify unknown characteristics of galaxies or the stars. In a similar way, studying concepts in an ontological network from different perspectives may discover new patterns that can represent unknown knowledge.

Figure 6.7: An illustration of galaxies in the Universe (Fruchter et al., 2008)

Discovering knowledge through new patterns in existing data has been conceptualised as a framework (Figure 6.8) by Fayyad and colleagues (Fayyad et al., 1996). It highlights three essential abilities:

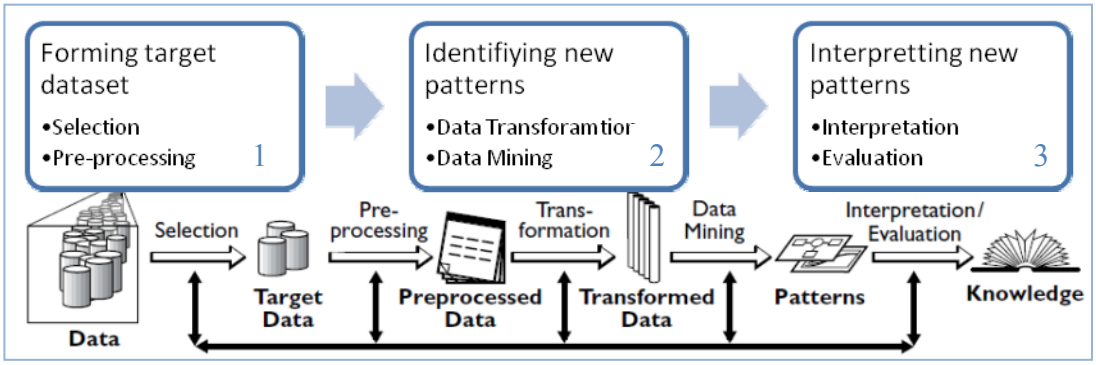


Figure 6.8: Illustration of steps discovering new knowledge (Fayyad et al., 1996)

1. Forming subsets of a large data collection for application domain: "Selection" creates the target dataset and "pre-processing" cleans the datasets such as removing irrelevant data and outliers.
2. Identifying new patterns: "Transformation" addresses different patterns to represent data, and "data mining" searches the dataset for the identified patterns.
3. Interpreting the new patterns: "Interpretation"/"evaluation" ensures the new patterns can be translated to useful knowledge and linked to existing knowledge.

Using semantic relationships to expand existing ontology has been adopted as ontological mechanisms for knowledge discovery. For instance, "co-appearing" unknown terms of known concepts can be retrieved from domain focused documents (Alfonseca and Manandhar, 2002) or a collection of search engine search results (Agirre et al., 2000, Qi et al., 2009). By calculating the appearance of unknown terms, such mechanism measures the "relevancy" (equivalent to SEA's closeness) between unknown terms and known concepts as a "pattern". It then interprets the unknown terms as the new concept description of the most "relevant" concepts.

The SEA approach could enhance the ontological mechanisms of knowledge discovery above in all three characteristics:

1. Forming dataset: Extending existing ontology may result a hierarchy structure of the dataset in reasoning towards a few given concepts. Instead, SEA's network structure can form a conceptual cluster for any given concept (subsets of ontological network) from multiple perspectives.
2. New patterns identification: The use of "relevancy" measurement between unknown concepts and existing concepts can define only one pattern. The new patterns identified could influence relationships between unknown concepts and known concepts, and they also have impacts on the relationships among existing concepts. However, extending existing ontology may recognise the latter impact, since the "relevancy" among existing concepts has been defined by the hierarchy structure (Alfonseca and Manandhar, 2002). In contrast, SEA is capable of identifying a series of new patterns through combining various relationship values such as centrality, closeness, relevant distance, and different observation perspectives of the network structure. These patterns also influence all concepts in SEA's ontological network, regardless their existence in the derived ontology.
3. New knowledge interpretation: Mapping concepts to hierarchy structure (Alfonseca and Manandhar, 2002, Qi et al., 2009) may restrain the interpretation of new concepts only to the given key concepts. In comparison, SEA's faceted output ability offers multiple interpretation directions. In addition, enriched

internal relationship can identify neighbouring concepts for unknown concepts to assist the understanding of the interpretation direction.

Therefore, SEA potentially provides an enhanced ontological approach for knowledge discovery. This has also been observed in practice. For example, SEA identified a pattern that concepts (such as "b2up" – a gum for breast cosmetic) with low centrality values in the whole network may have close relevant distance to a certain core concept (such as "cosmetic treatment"). This new patterns is then interpreted as a potential emerging new product/service for the related core concept. By the time this pattern was identified, TMG had not discovered the strong linkage between the two concepts. The knowledge discovery benefited TMG in preparing for relevant services six months ahead of the peak of "b2up" enquiries.

6.1.6 IT system architecture design

Building ontology for business can assist the componentisation and conceptualisation of business process, since ontology can describe a concept (a business process) by a group of sub-level concepts (relevant business functions). A business IT system may thus be constructed towards componentised processes, and using the ontology and the functionalities of the system can be defined according to the specified components (Ross and Westerman, 2004).

For instance, a part of the WMCCM business process is to match tender information with company profiles. Traditional IT architectures struggle to handle the complexity

of various tenders sources, information structures, business capabilities and company profiles. WMCCM applies ontology to facilitate componentising and conceptualising this business process as shown in Figure 6.9.

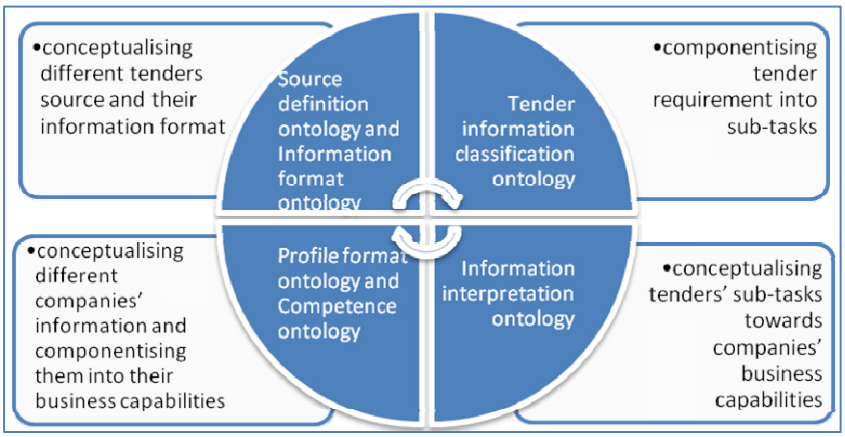


Figure 6.9: Examples of WMCCM ontology

The WMCCM IT architecture is thus componentised as shown in Figure 6.10.a. This architecture has proved successful in matching company capability with business opportunity (semantically tagging tenders).

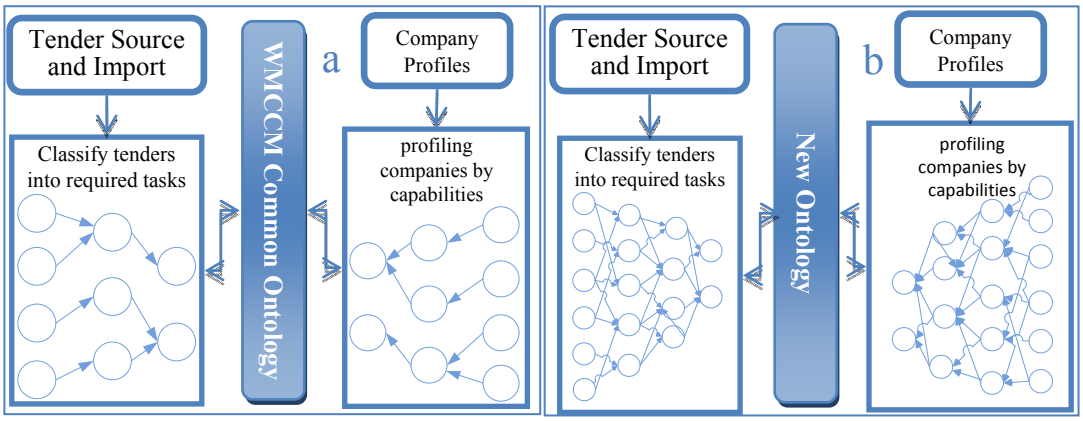


Figure 6.10: Illustration of WMCCM system architecture and improved architecture

WMCCM IT architecture processes information from over 10,000 companies in various industries with more than 50,000 tenders from different sources each year, and has successfully matched over 10 billion pounds worth of tenders for member companies from 2008 to 2010.

SEA has been utilised to re-engineer WMCCM's information interpretation ontology. The new derived ontology further componentises existing tender tasks into more detailed modular expressions, and re-conceptualises such modular expressions towards more business capabilities. Through the new derived ontology, WMCCM IT architecture has become more flexible and modularised, in order to meet wider range of requirements as shown in Figure 6.10.b.

SEA has also been applied to address business requirements from the Science and Technology Facility Council (STFC) in the UK. Ontology customised to STFC capabilities was generated. The SEA ontology has managed to modularise the STFC business processes (Figure 6.10.b), and thus focussed the WMCCM general engineering industry architecture for the specific capability and needs of STFC (STFC, 2010).

6.2 SUMMARY OF CONTRIBUTION TO KNOWLEDGE

This research contributes to ontology engineering and its relevant fields in the following ways:

1. The study provides an insight into ontology engineering methodologies from the perspective of building ontology for information filtering in automated systems. In this research, the desirable ontology characteristics are provided by the studies on existing ontology in the medical and engineering sectors.
2. This research contributes a novel methodology SEA to generate multi-disciplinary ontology that can provide the breadth and depth of coverage required for automated systems. SEA provides a quick and reliable process, and requires much less of a contribution from domain experts than current approaches.
3. The SEA approach constructs unconventional network ontological structures, and utilises statistical network analysis methods to reason the relationships among concepts. This provides SEA with a unique mapping of concepts and relationships, the ability to form visually configurable faceted ontology output, and quantified binary relationships to support fuzzy matching.
4. The SEA approach contributes a route to bridge reusable domain representative knowledge and usable application specified knowledge, since it is capable of retrieving both professional and non-expert terms, and identify their domain representativeness. This also assists semantic tagging with domain focus.
5. SEA's fast, economical and reliable ontology building for various subject areas provides a possible mechanism for knowledge discovery, as well as a mechanism for conceptualising and componentising business process and IT functions in assisting IT system architecture design.

6.3 MINUS FINDINGS

The "minus" findings indicate the limitation of the research.

6.3.1 SEA representativeness

SEA has been deployed to build ontology in the engineering (for WMCCM) and medical (for TMG) sectors. Additional pilot studies have also been conducted in E-Security, ICT and business management areas by the author. However, these examples of SEA deployment may not be sufficiently representative to be applied to other subject areas. Theoretically, SEA appears to be generalisable to other subject domains, because of the likely existence of semantic relatedness relationship in the Google index source. Some judgement may be made about its applicability by considering the amount of information that may be present about the domains of interest on Google.

6.3.2 Reliance on Internet information

The retrieval of semantic related terms from published information on the Internet may bring limitation to SEA approach and its derived ontology.

The knowledge base for this new methodology is the Google index, which archives publicly available information on the Internet. With the web 2.0 trend of user generated content, Google crawls through many types of pages from all sorts of authors, such as a new web page, a post from a blogger, or even a new twitter

message. The reliability of the source information may be questionable, particularly the reliability of web sites providing medical advice. Google itself gives higher ratings to sites that it considers more reliable, such as .gov, or .ac.

The amount of information available online around a certain topic may influence the "snowballing sampling" mechanism in SEA when obtaining semantically related terms. If there is poor information online around selected seeding terms, the selected tool – Google Sets may not be able to produce semantic related terms. This aspect is exacerbated by the trend for certain services to close off their content from Google indexing, Facebook being a prime example.

Moreover, a large quantity of information around seeding terms does not automatically imply a better "quality" of the predictions. Information published in the "wrong" direction may mislead the semantic relatedness. For example, "blackberry" and "apple" are generically recognised as fruits, but they have been heavily discussed in their other meanings – mobile phones. Consequently, Google Sets provides other mobile phone brands instead of other types of fruits when "blackberry" and "apple" are used as seeding words. Such misleading could be identified by the feasibility study, or eliminated by fault tolerance mechanism at an early stage, but it may require further supervision from ontology builders or domain experts.

6.3.3 Complexity of ontology output

The SEA ontology produces a multi-dimensional network structure, but the complexity of the ontological structure will reduce the efficiency of ICT systems for information reasoning.

For example, the original WMCCM ontology (862 concepts with 2,126) could filter a tender within five seconds on average. The customised SEA ontology for WMCCM application takes nearly eight seconds to filter a tender, on average, although it has only a small portion of full SEA ontology (less than 1/20 concepts) and less than 1/50 relationships. It seems that the current WMCCM ICT system may have difficulty in taking full advantage of SEA ontology's enriched structure without a hardware upgrade.

6.4 INTERESTING FINDINGS

These findings demonstrate the potential impact SEA may have on wider fields, which are not specifically addressed due to the research scope limits applied. They highlight possible valuable trails to pursue in future research.

6.4.1 Configuration variability

Within the experiments conducted SEA is configured to produce the most domain-focused terms. A different application environment may require different

settings, and may lead to a different ontology corpus structure and ontological analysis result.

The depth, breadth and quantity of amount of information in a subject area may require a variation in the SEA setting selected. Based on the results from five areas tested when developing SEA, it is difficult to draw a correlation between the information on a subject area and best settings to choose in configuring the ontology generation.

Interestingly SEA could produce concepts in multiple languages if seeding words from different languages are chosen. This aspect was not tested in the research undertaken. SEA could be configured to focus on producing terms in different languages. If the similarity in approach with Google translate holds true, such an approach could bring high performance multi-lingual ontology engineering capably to SEA.

6.4.2 Network structure analysis

The network structure of SEA ontology is analysed through a series of social network analysis methods. There are other analysis techniques that may be able to provide different interpretations of the same ontological structure for various research objectives. For instance, full investigation of the network reach (radiality), prestige of concepts in the directional network, structural coherence and structural hole analysis could draw clearer borders between conceptual clusters and the

reasoning paths between them. A thorough investigation of conceptual clusters could assist in the development of faceted systems capability. However, further investigation of the comparability of SEA ontology with different faceted systems will be required.

6.4.3 Knowledge discovery methods

The SEA approach has shown a level of connectivity and applicability to knowledge discovery mechanisms (discussion in section 6.1.5). Due to the research scope, limited comparison has been drawn between SEA characteristics and knowledge discovery methods. Knowledge discovery has been more thoroughly investigated in Data Mining and Artificial Intelligence. Thus, further investigation of knowledge discovery methods in these fields will be required to address the relationships between SEA's ontological approach and other existing approaches at the detailed process and technique level. This may serve to clarify the advantages of exchangeable processes and techniques, so as to mutually benefit the SEA approach and existing methods (in Data Mining and Artificial Intelligence) operation in knowledge discovery.

6.4.4 Connection zone member re-focusing

SEA currently focuses on highly centralised concepts in the corpus from a domain representativeness perspective. Changing the focus from the top definition zone to the bottom connection zone may provide potentially valuable understandings for

other research. For example, keeping an eye on the most unusual members with low centrality and observing their "movements" in their network over time is a key interest in E-Security related research.

Ontology engineering via SEA is equivalent to a "snapshot" of the knowledge base by retrieving the concepts and their relationships that exist at that particular time. It could trace the footprint of how the domain may develop with time, for example which concept areas are active and growing and which are stagnant. SEA's economical and fast ontology building characteristics allows researchers to take "snapshots" and compare the changes in a domain over a period of time.

The author suggests that connection zone member re-focusing should be prioritised among future research possibilities. Conducting connection zone member re-focusing research will require more storage, capacity due to the archiving of multiple version of the same ontology over time. Connection zone member re-focusing will also require different network structure analysis methods to efficiently analyse relationships from connection zone member perspectives.

CHAPTER 7: CONCLUSIONS

The research presented in this thesis recognises that good ontology can play a key role in ICT systems for "intelligent" processing, knowledge structuring, information interpretation and decision making. Through the investigation of two cases (WMCCM and TMG) and the relevant ontology in their domains, the need is identified to quickly, reliably and economically generate ontology, which can provide the breadth and depth of coverage required for given domain(s).

A new ontology engineering methodology has been proposed to address such needs, and the derived ontology has been implemented to improve the ICT systems' knowledge processing in the two case study applications. In summary, the conclusions drawn from this research are as follows:

1. Ontology driven ICT systems would benefit from broader concepts coverage and richer internal relationships. This would help them better bridge the terminology used by domain professionals and non-expert users, and enable better information matching with appropriate fuzziness and multi-discipline interpretation.
2. Existing ontology engineering methodologies rely heavily on the time of human experts. Since no one is a complete master of any domain, any derived methodology needs to be validated and tested by other domain experts.

3. A new ontology engineering methodology (SEA) is proposed by this research to address issues regarding the cost of generating ontology with sufficient scope and richness. SEA has been demonstrated as building rich multi-disciplinary ontology within 12 hours, with only three pairs of seeding words provided by domain experts. It can produce a high breadth and depth of concept coverage, and derive internal relationships to form a network ontological structure.
4. SEA qualifies as a generalisable ontology engineering methodology.
5. SEA advocates a statistical social network analysis to clarify the derived ontological network structure. A network structure offers rich reasoning paths which aid effective interpretation by ICT systems.
6. SEA enhances traditional logic binary relationships to directional weighted binary relationships, which aid decision making and logical reasoning.
7. The medical and engineering ontology derived by SEA performed better in the automated information filtering applications than the current ontology adopted by TMG and WMCCM.
8. SEA may further benefit problem areas such as machine translation, semantic tagging, knowledge discovery and IT system architecture design.

REFERENCES

Agirre, E., et al. (2000). Enriching very large ontologies using the WWW. *ArXiv Computer Science e-prints* [Online]. Available: http://ol2000.aifb.uni-karlsruhe.de/final/EAgirre_14.pdf [Accessed 22 Jan 2011].

Alfonseca, E. & Manandhar, S. (Year). An unsupervised method for General Named Entity Recognition and Automated Concept Discovery. *In:* Poceedings of the First International Conference on General WordNet, 2002.

Aussenac-Gilles, N., et al. (2000). Corpus analysis for conceptual modelling. *In:* Aussenac-Gilles, N., et al. (eds.) *EKAW-2000 Workshop on Ontologies and Text, Knowledge Engineering and Knowledge Management: Methods, Models and Tools.* Juan-Les-Pins, France.: CEUR Workshop Proceedings.

Benbasat, I., et al. (1987). The Case Research Strategy in Studies of Information Systems. *MIS Quarterly,* 11**,** 369-386.

Bernaras, A., et al. (1996). Building and Reusing Ontologies for Electrical Network Applications. *In:* Wolfgang, W. (ed.) *Proceedings of the 12th European Conference on Artificial Intelligence (ECAI 96): Budapest, Hungary: August 11-16.* Wiley.

Berners-Lee, T. (2000). *Weaving the Web: The Original Design and Ultimate*

Destiny of the World Wide Web, Collins.

Bono, E. D. (2006). *De Bono's thinking course*, BBC Active.

Burt, R. (2004). Structural Holes and Good Ideas. *American Journal of Sociology,* 110**,** 349-399.

Carrington, P., et al. (2005). *Models and methods in social network analysis*, Cambridge University Press.

Collis, J. & Hussey, R. (2009). *Business research: a practical guide for undergraduate & postgraduate students*, Palgrave Macmillan.

Corcho, O. & Gómez-Pérez, A. (2001). Solving Integration Problems of Ecommerce Standards and Initiatives through Ontological Mappings. *In:* Maedche, A., et al. (eds.) *IJCAI'2001 Workshop on Ontology Learning, Proceedings of the Second Workshop on Ontology Learning OL'2001*. Seattle, USA: CEUR-WS.org.

Davies, J., et al. (2003). *Towards the semantic web: ontology-driven knowledge management*, J. Wiley.

Davies, J., et al. (2006). *Semantic Web technologies : trends and research in ontology-based systems,* Chichester, England ; Hoboken, NJ, John Wiley & Sons.

De Hoog, R. (1998). Methodologies for Building Knowledge Based

Systems:Achievements and Prospects. *In:* Liebowitz, J. (ed.) *Handbook of Expert Systems.* Boca Raton, Florida: CRC Press.

Dean, M. & Schreiber, G. (2003). *OWL Web Ontology Language Reference* [Online]. W3C. Available: http://www.w3.org/TR/owl-ref/ [Accessed 20 Sep 2009].

Denzin, N. & Lincoln, Y. (1994). *Handbook of qualitative research*, Sage Publications.

Downs, E., et al. (1992). *Structured systems analysis and design method: application and context*, Prentice Hall.

Easterby-Smith, M., et al. (2002). *Management research: an introduction*, SAGE.

Espinoza, M., et al. (2008). Enriching an Ontology with Multilingual Information. *In:* Bechhofer, S., et al. (eds.) *The Semantic Web: Research and Applications.* Springer Berlin / Heidelberg.

Fairchild, A. & Vuyst, B. d. (2002). Coding Standards Benefiting Product and Service Information in E-commerce. *Proceedings of the 35th Annual Hawaii International Conference on System Sciences (HICSS'02)-Volume 8 - Volume 8.* IEEE Computer Society.

Faure, D. & Poibeau, T. (2000). First experiments of using semantic knowledge learned by ASIUM for information extraction task using INTEX. *In:* Staab,

S., et al. (eds.) *Ontology Learning ECAI-2000 Workshop.*

Fayyad, U., et al. (1996). The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM,* 39**,** 27-34.

Fensel, D., et al. (2000). OIL in a Nutshell. *In:* Dieng, R. & Corby, O. (eds.) *Proceedings of the 12th European Workshop on Knowledge Acquisition, Modeling and Management.* Juan-Les-Pins, France: Springer-Verlag.

Fernández-López, M. & Gómez-Pérez, A. (2002). Overview and analysis of methodologies for building ontologies. *The Knowledge Engineering Review,* 17**,** 129-156.

Fernández-López, M., et al. (1997). METHONTOLOGY: from Ontological Art towards Ontological Engineering. *In:* Farquhar, A., et al. (eds.) *Proceedings of the AAAI'97 Spring Symposium.* Stanford University, California.

Fernández-López, M., et al. (1999). Building a Chemical Ontology Using Methontology and the Ontology Design Environment. *IEEE Intelligent Systems,* 14**,** 37-46.

Friedkin, N. E. & Johnsen, E. C. (1997). Social positions in influence networks. *Social Networks,* 19**,** 209-222.

Fruchter, A., et al. (2008). *A Galaxy Cluster Lens* [Online]. NASA. Available: http://apod.nasa.gov/apod/ap080210.html [Accessed].

Giunchiglia, F., et al. (2009). Faceted Lightweight Ontologies. *In:* Borgida, A., et al. (eds.) *Conceptual Modeling: Foundations and Applications.* Springer Berlin / Heidelberg.

Gjoka, M., et al. (2010). Walking in Facebook: A Case Study of Unbiased Sampling of OSNs. *Proceedings of IEEE INFOCOM '10.*

Goetz, J. & Lecompte, M. (1984). *Ethnography and qualitative design in educational research*, Academic Press.

Gómez-Pérez, A. (1998). Knowledge Sharing and Reuse. *In:* Liebowitz, J. (ed.) *Handbook of Expert Systems.* Boca Raton, Florida: CRC Press.

Gómez-Pérez, A. (2001). Evaluation of ontologies. *International Journal of Intelligent Systems,* 16**,** 391-409.

Gómez-Pérez, A., et al. (2004). *Ontological Engineering : with examples from the areas of Knowledge Management, e-Commerce and the Semantic Web. First Edition (Advanced Information and Knowledge Processing)*, Springer.

Gómez-Pérez, A. & Rojas-Amaya, D. (1999). Ontological Reengineering for Reuse. *In:* Fensel, D. & Studer, R. (eds.) *Proceedings of the 11th European Workshop on Knowledge Acquisition, Modeling and Management (EKAW'99).* Dagstuhl Castle, Germany: Springer-Verlag.

Google. (2011). *Inside Google Translate* [Online]. Google. Available:

http://translate.google.com/about/intl/en_ALL/ [Accessed 10 Mar 2011].

Granada_Research. (2001). *Using the UNSPSC - United Nations Standard Products and Services Code* [Online]. UNSPSC. Available: http://www.unspsc.org/AdminFolder/Documents/UNSPSC_White_Paper.doc [Accessed 17 Sep 2009].

Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition - Special issue: Current issues in knowledge modeling,* 5**,** 199-220.

Grüninger, M. & Fox, M. (1995). Methodology for the Design and Evaluation of Ontologies. *IJCAI'95, Workshop on Basic Ontological Issues in Knowledge Sharing.*

Guarino, N. (1997). Semantic Matching: Formal Ontological Distinctions for Information Organization, Extraction, and Integration. *International Summer School on Information Extraction: A Multidisciplinary Approach to an Emerging Information Technology.* Springer-Verlag.

Guerrero, L. K., et al. (1999). *The nonverbal communication reader: classic and contemporary readings*, Waveland Press.

Gulli, A. & Signorini, A. (2005). The indexable web is more than 11.5 billion pages. *In:* Ellis, A. & Hagino, T. (eds.) *Special interest tracks and posters of the*

14th international conference on World Wide Web. Chiba, Japan: ACM.

Heiner, S. & Michel, K. (2004). Structure-Based Partitioning of Large Concept Hierarchies. *In:* McIlraith, S. A., et al. (eds.) *The Semantic Web – ISWC 2004.* Springer Berlin / Heidelberg.

Hughes, I. & William, R. (2001). *Planning your Action Research Project* [Online]. Available: http://www2.fhs.usyd.edu.au/arow/o/m01/hughes3.htm [Accessed 20 Nov 2009].

IEEE (1996). IEEE Standard for Developing Software Life Cycle Processes. *IEEE Std 1074-1995.* IEEE Computer Society.

IHTSDO. (2009). *SNOMED CT* [Online]. International Health Terminology Standeards development organisation. Available: http://www.ihtsdo.org/snomed-ct/ [Accessed 23 Sep 2009].

ILC-CNR. (2007). *Word Clustering* [Online]. Istituto Di Linguistica Computazionale. Available: http://www.ilc.cnr.it/EAGLES96/rep2/node37.html [Accessed 12 Apr 2009].

Jacob, E. (2004). Classification and categorization : a difference that makes a difference. *Library Trends,* 52**,** 515-540.

Jarrar, M. & Meersman, R. (2002). Formal Ontology Engineering in the DOGMA Approach. *In:* Meersman, R. & Tari, Z. (eds.) *On the Move to Meaningful*

Internet Systems 2002: CoopIS, DOA, and ODBASE. Springer Berlin / Heidelberg.

Jarrar, M. & Meersman, R. (2009). Ontology Engineering – The DOGMA Approach. *In:* Dillon, T., et al. (eds.) *Advances in Web Semantics I.* Springer Berlin / Heidelberg.

Katz, L. (1953). A new status index derived from sociometric analysis. *Psychometrika,* 18**,** 39-43.

Kietz, J., et al. (2000). A method for semi-automatic ontology acquisition from a corporate intranet. *In:* Aussenac-Gilles, N., et al. (eds.) *EKAW'00 Workshop on Ontologies and Texts. CEUR Workshop Proceedings* Juan-Les-Pins, France.

Kim, H. L., et al. (2008). The state of the art in tag ontologies: a semantic model for tagging and folksonomies. *Proceedings of the 2008 International Conference on Dublin Core and Metadata Applications.* Berlin, Germany: Dublin Core Metadata Initiative.

Kunder, M. (2010). *Size of the World Wide Web* [Online]. Available: http://www.worldwidewebsize.com [Accessed 15 Dec 2010].

Lacy, S. (2010). *Coincidence? Or Has this Tiny Indonesian Company Pushed Google into the Market Faster?* [Online]. TechCrunch. Available:

http://techcrunch.com/2010/11/12/coincidence-or-has-this-tiny-indonesian-company-pushed-google-into-the-market-faster/ [Accessed 10 Dec 2010].

Lau, R. (2007). Fuzzy domain ontology discovery for business knowledge management. *IEEE Intelligent Informatics Bulletin,* 8(1)**,** 29-41.

Lee, A. S. & Baskerville, R. L. (2003). Generalizing Generalizability in Information Systems Research. *INFORMATION SYSTEMS RESEARCH,* 14**,** 221-243.

Lenat, D. & Guha, R. V. (1989). *Building Large Knowledge-Based Systems; Representation and Inference in the Cyc Project*, Addison-Wesley Longman Publishing Co., Inc.

Maedche, A. & Staab, S. (2000). Mining Ontologies from Text. *In:* Dieng, R. & Corby, O. (eds.) *12th International Conference in Knowledge Engineering and Knowledge Management (EKAW'00)*. Juan-Les-Pins, France: Springer Berlin / Heidelberg.

Mahesh, K. (1996). Ontology Development for Machine Translation: Ideology and Methodology. *Technical Report.* Computing Research Laboratory, New Mexico State University.

Mayr, E. (1982). *The growth of biological thought: diversity, evolution, and inheritance*, Belknap Press.

Miller, G. (1995). WordNet: a lexical database for English. *Commun. ACM,* 38**,** 39-41.

Miller, G. A., et al. (1990). Introduction to WordNet: An On-line Lexical Database*. *International Journal of Lexicography,* 3**,** 235-244.

Mizoguchi, R., et al. (1995). Task ontology for reuse of problem solving knowledge. *In:* Mars, N. J. I. (ed.) *Towards Very Large Knowledge Bases: Knowledge Building & Knowledge Sharing.* Amsterdam: IOS Press.

Muhammad, A. & Lipika, D. (2007). A Fuzzy Ontology Generation Framework for Handling Uncertainties and Nonuniformity in Domain Knowledge Description. *Computing: Theory and Applications, 2007. ICCTA '07. International Conference on.*

National_Statistics. (2008). *INTRODUCTION TO UK STANDARD INDUSTRIAL CLASSIFICATION OF ECONOMIC ACTIVITIES UK SIC(92)* [Online]. Office for National Statistics. Available: http://www.statistics.gov.uk/methods_quality/sic/default.asp [Accessed 17 Sep 2009].

National_Statistics. (2010). *Standard Industrial Classification (SIC)* [Online]. Office for National Statistics. Available: http://www.statistics.gov.uk/statbase/product.asp?vlnk=14012 [Accessed 01 Aug 2010].

Nicola, A. D., et al. (2009). A software engineering approach to ontology building. *Inf. Syst.,* 34**,** 258-275.

NIST. (2005). *NIST 2005 Machine Translation Evaluation Official Results* [Online]. Available: http://www.itl.nist.gov/iad/mig//tests/mt/2005/doc/mt05eval_official_results_release_20050801_v3.html [Accessed 5 Mar 2011].

Noy, N. & Musen, M. (Year). Anchor-PROMPT: Using non-local context for semantic matching. *In:* Gómez-Pérez, A., et al., eds. Proceeding of IJCAI 2001 workshop on ontology and information sharing, 2001 Seattle, Washington. 63-70.

Noy, N. F. & Musen, M. A. (1999). SMART: Automated Support for Ontology Merging and Alignment. *In:* Gaines, B. (ed.) *Proceedings of the Twelfth Banff Workshop on Knowledge Acquisition, Modeling, and Management.* Banff, Alberta.

Passant, A. & Laublet, P. (2008). Meaning Of A Tag: A collaborative approach to bridge the gap between tagging and Linked Data. *Proceedings of the WWW 2008 Workshop Linked Data on the Web (LDOW2008), Beijing, China, Apr*.

Patrick, J., et al. (2008). Developing SNOMED CT Subsets from Clinical Notes for Intensive Care Service. *Health Care and Informatics Review Online,* 12(3)**,** 25-30.

Pressman, R. (2010). *Software engineering: a practitioner's approach*, McGraw-Hill Higher Education.

Qi, L., et al. (2009). A Study of Relation Annotation in Business Environments Using Web Mining. *In:* IEEE (ed.) *ICSC '09. IEEE International Conference on Semantic Computing.* IEEE.

Ranganathan, S. & Gopinath, M. (1967). *Prolegomena to library classification*, Asia Publishing House.

Rector, A. L., et al. (1995). A Terminology Server for medical language and medical information systems. *Methods of information in medicine,* 34**,** 147-157.

REDMOND, W. O. A. a. A. o. N. (2008). *Statement from Brad Smith, General Counsel, Microsoft* [Online]. Microsoft. Available: http://www.microsoft.com/presspass/press/2008/feb08/02-03statement.mspx [Accessed 16 Nov 2009].

Ross, J. W. & Westerman, G. (2004). Preparing for utility computing: The role of IT architecture and relationship management. *IBM Systems Journal,* 43**,** 5-19.

Salganik, M. (2003). *Sampling, estimation, and variance estimation in hidden populations using respondent-driven sampling*, Cornell University.

Scholten, H., et al. (2007). A methodology to support multidisciplinary model-based water management. *Environmental Modelling & Software,* 22**,** 743-759.

Schreiber, A. T., et al. (1995). The KACTUS View on the 'O' word. *In:* Skuce, D. (ed.) *IJCAI-95 Workshop on Basic Ontological Issues in Knowledge Sharing.* Montreal, Canada: Erasmus University Rotterdam.

Schulz, S., et al. (2009). SNOMED reaching its adolescence: Ontologists' and logicians' health check. *International Journal of Medical Informatics,* 78**,** S86-S94.

Seidenberg, J. & Rector, A. (2006). Web ontology segmentation: analysis, classification and use. *In:* Carr, L., et al. (eds.) *Proceedings of the 15th international conference on World Wide Web.* Edinburgh, Scotland: ACM.

Sommer, A. (2003). Small and medium sized enterprises: no longer just coping with the supply chain. *International Journal of Management and Enterprise Development,* 1(1)**,** 4-10.

Specia, L. & Motta, E. (2007). Integrating Folksonomies with the Semantic Web. *In:* Franconi, E., et al. (eds.) *The Semantic Web: Research and Applications.* Springer Berlin / Heidelberg.

Staab, S. (2001). Knowledge Processes and Ontologies. *IEEE Intelligent Systems,* 16(1)**,** 26-34.

STFC. (2010). *Welcome to SciTech Connections* [Online]. Available: http://www.scitechconnections.com/STFC/DesktopDefault.aspx?tabindex=1

&tabid=1 [Accessed].

Sure, Y., et al. (2002). OntoEdit: Guiding Ontology Development by Methodology and Inferencing. *In:* Meersman, R. & Tari, Z. (eds.) *On the Move to Meaningful Internet Systems 2002: CoopIS, DOA, and ODBASE*. Springer Berlin / Heidelberg.

Swartout, B., et al. (Year). Toward distributed use of large-scale ontologies. *In:* Farquhar, A., et al., eds. AAAI'97 Spring Symposium on Ontological Engineering, 1997 Stanford University, California. 138-148.

Tho, Q. T., et al. (2006). Automatic fuzzy ontology generation for semantic Web. *Knowledge and Data Engineering, IEEE Transactions on,* 18**,** 842-856.

TMG. (2009). *Our Service and Price* [Online]. The Taj Medical Group. Available: http://www.tajmedicalgroup.co.uk/cosmetic-surgery.html [Accessed 10 Oct 2009].

Tong, S. & Dean, J. (2008). *System and methods for automatically creating lists* [Online]. United States Patent. Available: http://patft.uspto.gov/netacgi/nph-Parser?Sect1=PTO2&Sect2=HITOFF&u=%2Fnetahtml%2FPTO%2Fsearch-adv.htm&r=1&p=1&f=G&l=50&d=PTXT&S1=7,350,187.PN.&OS=pn/7,350,187&RS=PN/7,350,187 [Accessed 12 Dec 2010].

UMLS. (2010). *About the UMLS* [Online]. U.S. National Library of Medicine. Available: http://www.nlm.nih.gov/research/umls/about_umls.html [Accessed 22 Sep 2009].

Uschold, M. & Gruninger, M. (1996). Ontologies: principles, methods and applications. *The Knowledge Engineering Review,* 11**,** 93-136.

van Heijst, G., et al. (1997). Using explicit ontologies in KBS development. *International Journal of Human-Computer Studies,* 46**,** 183-292.

Vander Wal, T. (2007). *Folksonomy* [Online]. Available: http://vanderwal.net/folksonomy.htm [Accessed 10 Nov 2010].

Ward-Schofield, J. (1993). Increasing the Generalizability of Qualitative Research. *In:* Huberman, A. & Miles, M. (eds.) *The qualitative researcher's companion.* Sage Publications.

Waterman, D. (1986). *A guide to expert systems,* Boston, Massachusetts, Addison-Wesley.

Watts, D. (2004). *Six degrees: the science of a connected age*, W. W. Norton & Company.

Wellington, J. & Szczerbiński, M. (2007). *Research methods for the social sciences*, Continuum International Pub. Group.

Welty, C. & Guarino, N. (2001). Supporting ontological analysis of taxonomic

relationships. *Data and Knowledge Engineering,* 39(1)**,** 51-74.

Wielinga, B. J., et al. (1992). KADS: a modelling approach to knowledge engineering. *Knowledge Acquisition - Special issue on the KADS approach to knowledge engineering,* 4(1)**,** 5-53.

Wikipedia. (2011). *Ontology (information science)* [Online]. Wikimedia Foundation, Inc. Available: http://en.wikipedia.org/wiki/Ontology_%28information_science%29 [Accessed].

Xu, Z., et al. (Year). Towards the semantic web: Collaborative tag suggestions. *In:* Proceedings of Collaborative Web Tagging Workshop at 15th International World Wide Web Conference, 2006.

Zhai, J., et al. (2008). Fuzzy Ontology Models Based on Fuzzy Linguistic Variable for Knowledge Management and Information Retrieval. *In:* Shi, Z., et al. (eds.) *Intelligent Information Processing IV*. Springer Boston.

Zhang, W. & Yin, J. (2008). Exploring Semantic Web technologies for ontology-based modeling in collaborative engineering design. *The International Journal of Advanced Manufacturing Technology,* 36**,** 833-843.

BIBLIOGRAPHY

Adriaans, P. & Zantinge, D. (1996). *Data Mining,* Harlow, United Kingdom, Addison-Wesley Professional.

Agrawal, R., et al. (1993). Mining association rules between sets of items in large databases. *ACM SIGMOD Record,* 22**,** 207-216.

Allemang, D. & Hendler, J. (2008). *Semantic web for the working ontologist: modeling in RDF, RDFS and OWL*, Morgan Kaufmann Publishers/Elsevier.

ARVIDSSON, F. & Annika, F. (2008). *Ontology I.* [Online]. Available: http://www.ida.liu.se/~janma/SemWeb/Slides/ontology1.pdf [Accessed 05 March 2010].

Ashworth, C. M. (1988). Structured systems analysis and design method (SSADM). *Information and Software Technology,* 30**,** 153-163.

Baker, L. D. & McCallum, A. K. (1998). Distributional clustering of words for text classification. *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval.* Melbourne, Australia: ACM.

Bernaras, A., et al. (1998). Building and using an electrical network ontology for fault diagnosis. *Engineering Intelligent Systems for Electrical Engineering and Communications,* 6**,** 3-11.

Bharat, K. & Broder, A. (1998). A technique for measuring the relative size and overlap of public Web search engines. *Computer Networks and ISDN Systems,* 30**,** 379-388.

Blaikie, N. (2007). *Approaches to Social Enquiry: Advancing Knowledge,* Cambridge, UK, Polity Press.

Blichfeldt, B. S. & Andersen, J. R. (2006). Creating a Wider Audience for Action Research: Learning from Case-Study Research. *Journal of Research Practice,* 2(1)**,** 1-15.

Bollegala, D., et al. (2007). Measuring semantic similarity between words using web search engines. *In:* Williamson, C. L., et al. (eds.) *Proceedings of the 16th international conference on World Wide Web (WWW'07).* Banff, Alberta, Canada: ACM.

Bollegala, D. T., et al. (2009). Measuring the similarity between implicit semantic relations from the web. *In:* Quemada, J., et al. (eds.) *Proceedings of the 18th international conference on World wide web.* Madrid, Spain: ACM.

Bonoma, T. V. (1985). Case research in marketing: Opportunities, problems, and a process. *Journal of Marketing Research,* 22**,** 199-208.

Borst, W. N. (1997). *Construction of Engineering Ontology for Knowledge Sharing and Reuse*, Enschede, The Netherlands.

Brachman, R. (1979). On the Epistemological Status of Semantic Networks. *In:* Findler, N. (ed.) *Associative Networks: Representation and Use of Knowledge by Computers.* London, United Kingdom: Academic Press.

Brickley, D. & Guha, R. (2003). *RDF Vocabulary Description Language 1.0: RDF Schema. W3C Working Draft.* [Online]. W3C. Available: http://www.w3.org/TR/PR-rdf-schema [Accessed 25 Sep 2009].

Brin, S. & Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems,* 30**,** 107-117.

Brydon-Miller, M., et al. (2003). Why Action Research? *Action Research,* 1**,** 9-28.

Calishain, T. (2002). New Google Answer Service Raises a Few Questions of its Own. *Information Today,* 19(6)**,** 50-51.

Calishain, T. (2002). Yahoo! Service Offers More Advice Than Exertise. *Information Today,* 19 (6)**,** 51.

Cheung, Y. P., et al. (Year). A case of sharing competences via a Collaborative Digital Network. *In:* Management of Innovation and Technology (ICMIT), 2010 IEEE International Conference on, 2-5 June 2010 2010. 1067-1072.

Cirasella, J. (2007). You and Me and Google Makes Three: Welcoming Google into the Reference Interview. *Library Philosophy and Practice*.

Cohen, L., et al. (2007). *Research methods in education*, Routledge.

Cornford, T. & Smithson, S. (2006). *Project research in information systems: a student's guide,* London, UK, Palgrave.

Cycorp. (2010). *the Cyc Knowledge Base* [Online]. Cycorp. Available: http://www.cyc.com/cyc/technology/whatiscyc_dir/whatsincyc [Accessed 22 Sep 2009].

Davies, J., et al. (2009). *Semantic Knowledge Management: Integrating Ontology Management, Knowledge Discovery, and Human Language Technologies*, Springer.

Davies, J., et al. (2005). Next generation knowledge management. *Bt Technology Journal,* 23**,** 175-190.

de Boer, V., et al. (2006). Extracting Instances of Relations from Web Documents Using Redundancy. *In:* Sure, Y. & Domingue, J. (eds.) *The Semantic Web: Research and Applications.* Springer Berlin / Heidelberg.

Doll, W. & Deng, X. (2010). A technology empowerment model for engineering work. *SIGMIS Database,* 41**,** 52-74.

Drucker, P. F. (2006). Knowledge-worker productivity: the biggest challenge. *Engineering Management Review, IEEE,* 34**,** 29-29.

Eisenhardt, K. (1989). Building Theories from Case Study Research. *The Academy of Management Review,* 14**,** 532-550.

Estes, W. (1996). *Classification and Cognition*, Oxford University Press.

Falagas, M. E., et al. (2008). Comparison of PubMed, Scopus, Web of Science, and Google Scholar: strengths and weaknesses. *The FASEB Journal,* 22**,** 338-342.

Fensel, D. (2003). *Ontology: A Silver Bullet for Knowledge Management and Electronic Commerce*, Springer-Verlag New York, Inc.

Fensel, D., et al. (2001). Product Data Integration in B2B E-Commerce. *IEEE Intelligent Systems,* 16**,** 54-59.

Friedkin, N. E. & Johnsen, E. C. (1997). Social positions in influence networks. *Social Networks,* 19**,** 209-222.

Gable, G. (1994). Integrating case study and survey research methods: an example in information systems. *European Journal of Information Systems,* 3:2**,** 112-126.

Galliers, R. D., et al. (1991). Choosing appropriate information systems research approaches: a revised taxonomy. *In:* Nissen, H.-E., et al. (eds.) *Information Systems Research: Contemporary Approaches and Emergent Traditions.* North Holland: Elsevier Science Publishers.

Gangemi, A., et al. (2006). Modelling Ontology Evaluation and Validation. *In:* Sure, Y. & Domingue, J. (eds.) *The Semantic Web: Research and Applications.* Springer Berlin / Heidelberg.

Gašević, D., et al. (2006). *Model driven architecture and ontology development*, Springer-Verlag.

Gilchrist, A. (2003). Thesauri, taxonomies and ontology -– an etymological note. *Journal of Documentation,* 59**,** 7-18.

Gjoka, M., et al. (2010). *Multigraph Sampling of Online Social Networks* [Online]. Cornell University. Available: http://arxiv.org/abs/1008.2565 [Accessed 12 Dec 2010].

Greenwood, D. & Levin, M. (2007). *Introduction to action research: social research for social change*, Sage Publications.

Gruninger, M., et al. (2008). Ontology Summit 2007 – Ontology, taxonomy, folksonomy: Understanding the distinctions. *Applied Ontology,* 3**,** 191-200.

Guarino, N. (1994). The Ontological Level. *In:* Casati, R., et al. (eds.) *Philosophy and the Cognitive Science.* Vienna, Austria: Hölder-Pichler-Tempsky.

Guarino, N. (1998). *Formal ontology in information systems: proceedings of the first international conference (FOIS'98), June 6-8, Trento, Italy*, IOS Press.

Guarino, N. & Giaretta, P. (1995). Ontology and Knowledge Bases - Towards a Terminological Clarification. *In:* Mars, N. (ed.) *Towards Very Large Knowledge Bases.* IOS Press, Amsterdam, The Netherlands.

Guizzardi, G. & Halpin, T. (2008). Ontological foundations for conceptual modelling. *Applied Ontology,* 3**,** 1-12.

Gummesson, E. (2000). *Qualitative Methods in Management Research,* Thousand Oaks, CA, Sage.

Hahn, U. & Schulz, S. (2000). Towards Very Large Terminological Knowledge Bases: A Case Study from Medicine. *In:* Hamilton, H. (ed.) *Proceedings of the 13th Biennial Conference of the Canadian Society on Computational Studies of Intelligence: Advances in Artificial Intelligence.* Montréal, Quebec, Canada: Springer-Verlag.

Han, J., et al. (2006). *Data Mining: Concepts and Techniques, Second Edition*, Morgan Kaufmann.

Hatten, R., et al. (1997). *Action Research: Comparison with the concepts of 'The Reflective Practitioner' and 'Quality Assurance'* [Online]. The University of Sydney. Available: http://www.scu.edu.au/schools/gcm/ar/arr/arow/rdr.html [Accessed 25 Nov 2009].

Hearn, G. N. & Foth, M. (2005). Action Research in the Design of New Media and

ICT Systems. *In:* Kwansah-Aidoo, K. (ed.) *Topical Issues in Communications and Media Research.* New York, NY: Nova Science.

Hearst, M. A. (1992). Automatic acquisition of hyponyms from large text corpora. *In:* Zampolli, A. (ed.) *Proceedings of the 14th conference on Computational linguistics - Volume 2.* Nantes, France: Association for Computational Linguistics.

Hepp, M. (2006). Products and Services Ontology: A Methodology for Deriving OWL Ontology from Industrial Categorization Standards. *International Journal On Semantic Web and Information Systems,* Volume 2 - 2006**,** 72 - 99.

Hepp, M., et al. (2007). A quantitative analysis of product categorization standards: content, coverage, and maintenance of eCl@ss, UNSPSC, eOTD, and the RosettaNet Technical Dictionary. *Knowledge and Information Systems,* 13(1)**,** 77-114.

HOCKEY, S. & WALKER, D. (1993). Developing Effetive Resources for Research on Texts: Collecting Texts, Tagging Texts, Cataloguing Texts, Using Texts, and Putting Texts in Context. *Literary and Linguistic Computing,* 8**,** 235-242.

HOCKEY, S. & WALKER†, D. (1993). Developing Effetive Resources for Research on Texts: Collecting Texts, Tagging Texts, Cataloguing Texts, Using Texts, and Putting Texts in Context. *Literary and Linguistic*

Computing, 8**,** 235-242.

Holsapple, C. & Joshi, K. D. (2002). A collaborative approach to ontology design. *Commun. ACM,* 45**,** 42-47.

Holsapple, C. W. (2004). *Handbook of Knowledge Management: Knowledge Matters.*, Springer.

IEEE (1990). IEEE Standard Glossary of Software Engineering Terminology. *IEEE Std 610.121990.* IEEE Computer Society.

Iphofen, R. & Poland, F. (2002). *Sociology in practice for health care professionals*, Palgrave.

Jasper, R. & Uschold, M. (1999). A Framework for Understanding and Classifying Ontology Applications. *In:* Benjamins, V. R. (ed.) *Proceedings of the IJCAI-99 Workshop on Ontology and Problem-Solving Methods: Lessons Learned and Future Trends.* Stockholm, Sweden: CEUR Publications.

Johnson, S. (1967). Hierarchical clustering schemes. *Psychometrika,* 32**,** 241-254.

Kavi, M. & Sergei, N. (1998). Semantic Classification for Practical Natural Language Processing. *In:* Schwartz, R. & Beghtol, C. (eds.) *Advances in classification research: proceedings of the 6th ASIS SIG/CR Classification Research Workshop, held at the 58th ASIS Annual Meeting, Chicago, Illinois, October 8, 1995.* Information Today, Inc.

Klyne, G. & Carroll, J. (2004). *Resource Description Framework (RDF): Concepts and Abstract Syntax.* [Online]. W3C. Available: http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/ [Accessed 17 Sep 2009].

Knight, K. & Luk, S. K. (1994). Building a large-scale knowledge base for machine translation. *In:* AAAI (ed.) *Proceedings of the Twelfth National Conference on Artificial Intelligence, Volume 1*. Seattle, United States: American Association for Artificial Intelligence Press.

Kok, S. & Domingos, P. (2008). Extracting Semantic Networks from Text Via Relational Clustering. *In:* Daelemans, W., et al. (eds.) *Machine Learning and Knowledge Discovery in Databases.* Springer Berlin / Heidelberg.

Kumar, R. (2005). *Research methodology: a step-by-step guide for beginners*, Pearson Education Australia.

Lanzenberger, M. & Sampson, J. (2008). Making Ontology Talk: Knowledge Interoperability in the Semantic Web. *Intelligent Systems, IEEE,* 23**,** 72-85.

Lassila, O. & McGuinness, D. (2001). *The role of frame-based representation on the semantic web* [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.104.3946 [Accessed 03 Oct 2010].

Lau, F. (1997). A review on the use of action research in information systems studies. *Proceedings of the IFIP TC8 WG 8.2 international conference on Information systems and qualitative research.* Philadelphia, Pennsylvania, United States: Chapman \& Hall, Ltd.

Lei, Y., et al. (2006). An Infrastructure for Acquiring High Quality Semantic Metadata. *In:* Sure, Y. & Domingue, J. (eds.) *The Semantic Web: Research and Applications.* Springer Berlin / Heidelberg.

Lewin, K. (1947). Frontiers in Group Dynamics: II. Channels of Group Life; Social Planning and Action Research. *Human Relations,* 1**,** 143-153.

Li, Z., et al. (2009). A methodology for engineering ontology acquisition and validation. *Artificial Intelligence for Engineering Design,* 23**,** 37-51.

Linstone, H. A. & Turoff, M. (2002). *The Delphi Method: Techniques and Applications*, Addison-Wesley.

Lyman, P., et al. (2003). *How much information? 2003* [Online]. Available: http://www2.sims.berkeley.edu/research/projects/how-much-info-2003/ [Accessed 05 Nov 2010].

Ma, X. & Bal, J. (2009). Semantic industrial categorisation based on Search Engine Index. *In:* Arabnia, H. R., et al. (eds.) *Proceedings of the 2009 International Conference on Information & Knowledge Engineering, IKE 2009.* Las Vegas

Nevada, USA: CSREA Press.

Ma, X. & Bal, J. (2009). A Sensemaking-based Information Foraging and Summarization System in Business Environments. *In:* Cunningham, P. & Cunningham, M. (eds.) *eChallenges e-2009 Conference Proceedings.* Istanbul, Turkey: IIMC International Information Management Corporation Ltd.

Maedche, A. & Staab, S. (Year). Semi-automatic engineering of ontology from text. *In:* Chang, S. & Obozinski, W., eds. 12th International Conference on Software Engineering and Knowledge Engineering (SEKE'2000), 2000 Chicago, Illinois.

Marczyk, G. R., et al. (2010). *Essentials of Research Design and Methodology*, John Wiley & Sons.

McEnery, A. M. & Wilson, A. (2001). Corpus linguistics : an introduction. Edinburgh University Press.

Mendonca, D. & Bouwman, H. (2008). Information and communications technology for crisis management: Defining an agenda for scientific research. *In:* Thissen, W. & Weijnen, M. (eds.) *Proceedings of the International Conference on Infrastructure Systems, Building Networks for a Brighter Future (INFRA).*

Moon, M. (2009). Knowledge worker productivity. *Journal of Digital Asset Management,* 5**,** 178-180.

Morik, K. & Kietz, J.-U. (1989). A bootstrapping approach to conceptual clustering. *In:* Segre, A. (ed.) *Proceedings of the sixth international workshop on Machine learning.* Ithaca, New York, United States: Morgan Kaufmann Publishers Inc.

NAICS. (2010). *FAQs (Ask Dr. NAICS)* [Online]. NAICS. Available: ttp://www.census.gov/eos/www/naics/faqs/faqs.html#q5 [Accessed 17 Sep 2009].

Navigli, R. & Velardi, P. (2004). Learning Domain Ontology from Document Warehouses and Dedicated Web Sites. *Computational Linguistics,* 30**,** 151-179.

O'Brien, R. (2001). *Um exame da abordagem metodológica da pesquisa ação [An Overview of the Methodological Approach of Action Research]* [Online]. João Pessoa, Brazil: Universidade Federal da Paraíba. Available: http://www.web.ca/~robrien/papers/arfinal.html [Accessed 20 nov 2009].

Orlikowski, W. J. & Baroudi, J. J. (1991). Studying information technology in organizations: Research approaches and assumptions. *INFORMATION SYSTEMS RESEARCH,* Vol. 2**,** 1-28

Patton, M. Q. (2005). Qualitative Research. *In:* Everitt, B. & Howell, D. (eds.) *Encyclopedia of statistics in behavioral science.* John Wiley & Sons.

PEASE, A. & NILES, I. (2002). IEEE standard upper ontology: a progress report. *The Knowledge Engineering Review,* 17**,** 65-70.

Powell, T. (1994). *Understanding the self-help organization: frameworks and findings*, Sage Publications.

Preece, R. A. (2000). *Starting research : an introduction to academic research and dissertation writing,* London; New York, Continuum.

Pretorius, A. J. (2005). Visual analysis for ontology engineering. *Journal of Visual Languages & Computing,* 16**,** 359-381.

Rabinovich, A., et al. (2007). Objects in Context. *IEEE 11th International Conference on Computer Vision, 2007. ICCV 2007.* .

Raveendranathan, P. (2005). *Identifying Sets of Related Words from the World Wide Web* [Online]. University of Minnesota. Available: http://www.d.umn.edu/~tpederse/Pubs/prath-thesis.pdf [Accessed 07 Jan 2010].

Remenyi, D. & Williams, B. (1996). The nature of research: qualitative or quantitative, narrative or paradigmatic? *Information Systems Journal,* 6**,** 131-146.

Remenyi, D. & Williams, B. (1998). *Doing research in business and management: an introduction to process and method*, SAGE.

RNCOS. (2007). *Opportunities in Asian Medical Tourism (2007-2010)* [Online]. RNCOS. Available: http://www.rncos.com/Report/IM0102.htm [Accessed 11 Nov 2009].

Rowe, G. & Wright, G. (2001). Expert Opinions in Forecasting. Role of the Delphi Technique. *In:* Armstrong, J. (ed.) *Principles of forecasting: a handbook for researchers and practitioners.* Boston, USA: Kluwer Academic.

Roy, W. (1995). Action research: a new paradigm for research in production andoperations management. *International Journal of Operations & Production Management,* 15**,** 6-20.

Schorlemmer, M. & Kalfoglou, Y. (2008). Institutionalising ontology-based semantic integration. *Applied Ontology,* 3**,** 131-150.

Sharp, J., et al. (2002). *The management of a student research project*, Gower.

Sleeman, D., et al. (2008). The role of ontology in creating and maintaining corporate knowledge: A case study from the aero industry. *Applied Ontology,* 3**,** 151-172.

Slonim, N. & Tishby, N. (2000). Document clustering using word clusters via the information bottleneck method. *Proceedings of the 23rd annual international*

ACM SIGIR conference on Research and development in information retrieval. Athens, Greece: ACM.

Sowa, J. (2000). *Knowledge Representation: Logical, Philosophical, and Computational Foundations*, Brooks/Cole.

Sparckjones, K. (2007). Automatic summarising: The state of the art. *Information Processing & Management,* 43**,** 1449-1481.

Spyns, P., et al. (2008). An ontology engineering methodology for DOGMA. *Applied Ontology,* 3**,** 13-39.

Staab, S., et al. (2000). An Extensible Approach for Modeling Ontology in RDF(S). *In:* Constantopoulos, P. (ed.) *ECDL 2000 Workshop on the Semantic Web.* Lisbon, Portugal.

Studer, R., et al. (1998). Knowledge engineering: Principles and methods. *Data & Knowledge Engineering,* 25**,** 161-197.

Stumme, G. & Maedche, A. (2001). FCA-MERGE: bottom-up merging of ontology. *In:* Nebel, B. (ed.) *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence (IJCAI 2001) - Volume 1.* Seattle.: Morgan Kaufmann Publishers Inc.

SUOWP. (2003). *Standard Upper Ontology Working Group Home Page* [Online]. IEEE. Available: http://suo.ieee.org/ [Accessed 22 Sep 2009].

Tague-Sutcliffe, J. M. (1996). Some perspectives on the evaluation of information retrieval systems. *Journal of the American Society for Information Science,* 47**,** 1-3.

Taniar, D. & Rahayu, J. (2006). *Web semantics and ontology*, Idea Group Pub.

Tempich, C., et al. (2006). Ontology Engineering Revisited: An Iterative Case Study. *In:* Sure, Y. & Domingue, J. (eds.) *The Semantic Web: Research and Applications.* Springer Berlin / Heidelberg.

Tempich, C., et al. (2007). Argumentation-Based Ontology Engineering. *Intelligent Systems, IEEE,* 22**,** 52-59.

Trochim, W. & Donnelly, J. (2006). *Research methods knowledge base*, Cengage Learning.

Uschold, M. (1995). Towards a Methodology for Building Ontology. *In:* Skuce, D. (ed.) *IJCAI'95 Workshop on Basic Ontological Issues in Knowledge Sharing.* Montreal, Canada.

van Harmelen, F. & Horrocks, I. (2001). *Reference Description of the DAML+OIL(March 2001) Ontology Markup Language* [Online]. daml. Available: http://www.daml.org/2001/03/reference.html [Accessed 18 Sep 2009].

Völker, J., et al. (2008). AEON – An approach to the automatic evaluation of

ontology. *Applied Ontology,* 3**,** 41-62.

Vossen, P. (2002). *EuroWordNet: general document* [Online]. Amsterdam: Vrije Universiteit. Available: http://dare.ubvu.vu.nl/handle/1871/11116 [Accessed 25 Sep 2009].

Vossen, P. T. J. M. (1998). *EuroWordNet: a multilingual database with lexical semantic networks for European Languages,* Dordrecht, The Netherlands, Kluwer Academic Publishers.

VTSL. (2010). *SNOMED CT Core Browser* [Online]. The Veterinary Medical Informatics. Available: http://terminology.vetmed.vt.edu/SCT/menu.cfm [Accessed 05 Jun 2010].

W3C. (2010). *Resource Description Framework* [Online]. W3C. Available: http://www.w3.org/RDF/ [Accessed 11 May 2009].

Walsham, G. (1995). The Emergence of Interpretivism in IS Research. *INFORMATION SYSTEMS RESEARCH,* 6**,** 376-394.

Wang, T., et al. (2006). Automatic Extraction of Hierarchical Relations from Text. *In:* Sure, Y. & Domingue, J. (eds.) *The Semantic Web: Research and Applications.* Springer Berlin / Heidelberg.

Whitehead, D. (2005). Project management and action research: two sides of the same coin? *Journal of Health Organization and Management,* 19**,** 519 - 531.

WMCCM. (2010). *Partner Search* [Online]. WMCCM. Available: http://www.wmccm.co.uk/WMCCM/DesktopDefault.aspx?tabindex=4&tabid=17 [Accessed 04 Apr 2010].

Yanna, W. & Zili, Z. (2009). Domain Ontology Generation Based on WordNet and Internet. *Management and Service Science, 2009. MASS '09. International Conference on.*

Yin, R. (2009). *Case study research: design and methods*, Sage Publications.

APPENDICES

*All appendixes for this research is stored in the CD provided.

Appendix 1 – Raw data collected for this research and viewer friendly presentation of primary data.

File location: Appendix\Appendix.1

This appendix is constituted of 11 sub collections from appendix 1.1 – appendix 1.11

Appendix 1.1 – Raw data collection.

File location: Appendix\Appendix.1

This appendix is constituted of two parts:

> ***Appendix 1.1.1 – Raw data stored in SQL database.***
>
> *File location: Appendix\Appendix.1\Appendix 1.1\Appendix 1.1.1 - DBbackup20110124*
>
> ***Appendix 1.1.2 - SQL data mapping to view friendly presentations***
>
> *File location: Appendix\Appendix.1\Appendix 1.1\Appendix 1.1.2 - SQL data mapping.pdf*

Appendix 1.2 – Data collection for engineering ontology.

File location: Appendix\Appendix.1\Appendix 1.2

This appendix is constituted of six parts:

Appendix 1.2.1 - Data collected for engineering ontology feasibility study

File location: Appendix\Appendix.1\Appendix 1.2\Appendix 1.2.1 - Data collected for engineering ontology (feasibility study).pdf

Appendix 1.2.2 - Centrality calculation for engineering ontology after feasibility study

File location: Appendix\Appendix.1\Appendix 1.2\Appendix 1.2.2 - Centrality calculation for engineering ontology (feasibility study).pdf

Appendix 1.2.3 - Data collected for engineering ontology

File location: Appendix\Appendix.1\Appendix 1.2\Appendix 1.2.3 - Data collected for engineering ontology (full).pdf

Appendix 1.2.4 - Centrality calculation for engineering ontology

File location: Appendix\Appendix.1\Appendix 1.2\Appendix 1.2.4 - Centrality calculation for engineering ontology (full).pdf

Appendix 1.2.5 - Relationship Calculation for engineering ontology top zone members

File location: Appendix\Appendix.1\Appendix 1.2\Appendix 1.2.5 - Relationship Calculation for engineering ontology (top zone members).pdf

Appendix 1.2.6 - Relationship Calculation for engineering ontology

File location: Appendix\Appendix.1\Appendix 1.2\Appendix 1.2.6 - Relationship Calculation for engineering ontology (full).pdf

Appendix 1.3 – Data collection for medical ontology.

File location: Appendix\Appendix.1\Appendix 1.3

This appendix is constituted of five parts:

Appendix 1.3.1 - Data collected for medical ontology feasibility study

File location: Appendix\Appendix.1\Appendix 1.3\Appendix 1.3.1 - Data collected for medical ontology (feasibility study).pdf

Appendix 1.3.2 - Centrality calculation for medical ontology after feasibility study

File location: Appendix\Appendix.1\Appendix 1.3\Appendix 1.3.2 - Centrality calculation for medical ontology (feasibility study).pdf

Appendix 1.3.3 - Data collected for medical ontology

File location: Appendix\Appendix.1\Appendix 1.3\Appendix 1.3.3 - Data collected for medical ontology (full).pdf

Appendix 1.3.4 - Centrality calculation for medical ontology

File location: Appendix\Appendix.1\Appendix 1.3\Appendix 1.3.4 - Centrality calculation for medical ontology (full).pdf

Appendix 1.3.5 - Relationship calculation for medical ontology

File location: Appendix\Appendix.1\Appendix 1.3\Appendix 1.3.5 - Relationship calculation for medical ontology.pdf

Appendix 1.4 – Data collection for alternative engineering ontology study 1.

File location: Appendix\Appendix.1\Appendix 1.4

This appendix is constituted of four parts:

Appendix 1.4.1 - Data collected for alternative engineering ontology 1feasibility study

File location: Appendix\Appendix.1\Appendix 1.4\Appendix 1.4.1 - Data collected for alternative engineering ontology 1 (feasibility study).pdf

Appendix 1.4.2 - Centrality calculation for alternative engineering ontology 1 after feasibility study

File location: Appendix\Appendix.1\Appendix 1.4\Appendix 1.4.2 - Centrality calculation for alternative engineering ontology 1 (feasibility study).pdf

Appendix 1.4.3 - Data collected for alternative engineering ontology 1

File location: Appendix\Appendix.1\Appendix 1.4\Appendix 1.4.3 - Data collected for alternative engineering ontology 1 (full).pdf

Appendix 1.4.4 - Centrality calculation for alternative engineering ontology 1

File location: Appendix\Appendix.1\Appendix 1.4\Appendix 1.4.4 - Centrality calculation for alternative engineering ontology 1 (full).pdf

Appendix 1.5 – Data collection for alternative engineering ontology study 2.

File location: Appendix\Appendix.1\Appendix 1.5

This appendix is constituted of four parts:

Appendix 1.5.1 - Data collected for alternative engineering ontology 2 feasibility study

File location: Appendix\Appendix.1\Appendix 1.5\Appendix 1.5.1 - Data collected for alternative engineering ontology 2 (feasibility study).pdf

Appendix 1.5.2 - Centrality calculation for alternative engineering ontology 2 after feasibility study

File location: Appendix\Appendix.1\Appendix 1.5\Appendix 1.5.2 - Centrality calculation for alternative engineering ontology 2 (feasibility study).pdf

Appendix 1.5.3 - Data collected for alternative engineering ontology 2

File location: Appendix\Appendix.1\Appendix 1.5\Appendix 1.5.3 - Data collected for alternative engineering ontology 2 (full).pdf

Appendix 1.5.4 - Centrality calculation for alternative engineering ontology 2

File location: Appendix\Appendix.1\Appendix 1.5\Appendix 1.5.4 - Centrality calculation for alternative engineering ontology 2 (full).pdf

Appendix 1.6 – Data collection for alternative engineering ontology study 3.

File location: Appendix\Appendix.1\Appendix 1.6

This appendix is constituted of four parts:

Appendix 1.6.1 - Data collected for alternative engineering ontology 3 feasibility study

File location: Appendix\Appendix.1\Appendix 1.6\Appendix 1.6.1 - Data collected for alternative engineering ontology 3 (feasibility study).pdf

Appendix 1.6.2 - Centrality calculation for alternative engineering ontology 3 after feasibility study

File location: Appendix\Appendix.1\Appendix 1.6\Appendix 1.6.2 - Centrality calculation for alternative engineering ontology 3 (feasibility study).pdf

Appendix 1.6.3 - Data collected for alternative engineering ontology 3

File location: Appendix\Appendix.1\Appendix 1.6\Appendix 1.6.3 - Data collected for alternative engineering ontology 3 (full).pdf

Appendix 1.6.4 - Centrality calculation for alternative engineering ontology 3

File location: Appendix\Appendix.1\Appendix 1.6\Appendix 1.6.4 - Centrality calculation for alternative engineering ontology 3 (full).pdf

Appendix 1.7 – Data collection for alternative medical ontology study 1.

File location: Appendix\Appendix.1\Appendix 1.7

This appendix is constituted of five parts:

Appendix 1.7.1 - Data collected for alternative medical ontology 1 feasibility study

File location: Appendix\Appendix.1\Appendix 1.7\Appendix 1.7.1 - Data collected for alternative medical ontology 1 (feasibility study).pdf

Appendix 1.7.2 - Centrality calculation for alternative medical ontology 1 after feasibility study

File location: Appendix\Appendix.1\Appendix 1.7\Appendix 1.7.2 - Centrality calculation for alternative medical ontology 1 (feasibility study).pdf

Appendix 1.7.3 - Data collected for alternative medical ontology 1

File location: Appendix\Appendix.1\Appendix 1.7\Appendix 1.7.3 - Data collected for alternative medical ontology 1 (full).pdf

Appendix 1.7.4 - Centrality calculation for alternative medical ontology 1

File location: Appendix\Appendix.1\Appendix 1.7\Appendix 1.7.4 - Centrality calculation for alternative medical ontology 1 (full).pdf

Appendix 1.7.5 - Relationship calculation for alternative medical ontology 1

File location: Appendix\Appendix.1\Appendix 1.7\Appendix 1.7.5 - Relationship calculation for alternative medical ontology 1.pdf

Appendix 1.8 – Data collection for e-security ontology study 1.

File location: Appendix\Appendix.1\Appendix 1.8

This appendix is constituted of five parts:

Appendix 1.8.1 - Data collected for e-security ontology 1 feasibility study

File location: Appendix\Appendix.1\Appendix 1.8\Appendix 1.8.1 - Data collected for e-security ontology 1 (feasibility study).pdf

Appendix 1.8.2 - Centrality calculation for e-security ontology 1 after feasibility study

File location: Appendix\Appendix.1\Appendix 1.8\Appendix 1.8.2 - Centrality calculation for e-security ontology 1 (feasibility study).pdf

Appendix 1.8.3 - Data collected for e-security ontology 1

File location: Appendix\Appendix.1\Appendix 1.8\Appendix 1.8.3 - Data collected for e-security ontology 1 (full).pdf

Appendix 1.8.4 - Centrality calculation for e-security ontology 1

File location: Appendix\Appendix.1\Appendix 1.8\Appendix 1.8.4 - Centrality calculation for e-security ontology 1 (full).pdf

Appendix 1.8.5 - Relationship calculation for e-security ontology 1

File location: Appendix\Appendix.1\Appendix 1.8\Appendix 1.8.5 - Relationship calculation for e-security ontology 1.pdf

Appendix 1.9 – Data collection for e-security ontology study 2.

File location: Appendix\Appendix.1\Appendix 1.9

This appendix is constituted of five parts:

Appendix 1.9.1 - Data collected for e-security ontology 2 feasibility study

File location: Appendix\Appendix.1\Appendix 1.9\Appendix 1.9.1 - Data collected for e-security ontology 2 (feasibility study).pdf

Appendix 1.9.2 - Centrality calculation for e-security ontology 2 after feasibility study

File location: Appendix\Appendix.1\Appendix 1.9\Appendix 1.9.2 - Centrality calculation for e-security ontology 2 (feasibility study).pdf

Appendix 1.9.3 - Data collected for e-security ontology 2

File location: Appendix\Appendix.1\Appendix 1.9\Appendix 1.9.3 - Data collected for e-security ontology 2 (full).pdf

Appendix 1.9.4 - Centrality calculation for e-security ontology 2

File location: Appendix\Appendix.1\Appendix 1.9\Appendix 1.9.4 - Centrality calculation for e-security ontology 2 (full).pdf

Appendix 1.9.5 - Relationship calculation for e-security ontology 2

File location: Appendix\Appendix.1\Appendix 1.9\Appendix 1.9.5 - Relationship calculation for e-security ontology 2.pdf

Appendix 1.10 – Data collection for ICT ontology building.

File location: Appendix\Appendix.1\Appendix 1.10

This appendix is constituted of four parts:

Appendix 1.10.1 - Data collected for ICT ontology feasibility study

File location: Appendix\Appendix.1\Appendix 1.10\Appendix 1.10.1 - Data collected for ICT ontology (feasibility study).pdf

Appendix 1.10.2 - Centrality calculation for ICT ontology after feasibility study

File location: Appendix\Appendix.1\Appendix 1.10\Appendix 1.10.2 - Centrality calculation for ICT ontology (feasibility study).pdf

Appendix 1.10.3 - Data collected for ICT ontology

File location: Appendix\Appendix.1\Appendix 1.10\Appendix 1.10.3 - Data collected for ICT ontology (full).pdf

Appendix 1.10.4 - Centrality calculation for ICT ontology

File location: Appendix\Appendix.1\Appendix 1.10\Appendix 1.10.4 - Centrality calculation for ICT ontology (full).pdf

Appendix 1.11 – Data collection for MAS ontology building.

File location: Appendix\Appendix.1\Appendix 1.11

This appendix is constituted of four parts:

Appendix 1.11.1 - Data collected for MAS ontology

File location: Appendix\Appendix.1\Appendix 1.11\Appendix 1.11.1 - Data collected for MAS ontology (full).pdf

Appendix 1.11.2 - Centrality calculation for MAS ontology

File location: Appendix\Appendix.1\Appendix 1.11\Appendix 1.11.2 - Centrality calculation for MAS ontology (full).pdf

Appendix 2 – Supportive information for chapter 2

File location: Appendix\Appendix.2

This appendix is constituted of:

Appendix 2.1 - Medical Tourism Research

File location: Appendix\Appendix.2\Appendix 2.1 - Medical Tourism Research.pdf

Appendix 2.2 - Type of Ontology

File location: Appendix\Appendix.2\Appendix 2.2 - Type of Ontology.pdf

Appendix 3 – Supportive information for chapter 3

File location: Appendix\Appendix.2

This appendix is constituted of:

Appendix 3.1 - Source Selection

File location: Appendix\Appendix.3\Appendix 3.1 - Source Selection.pdf

Appendix 3.2 - Word Clustering

File location: Appendix\Appendix.3\Appendix 3.2 - Word Clustering.pdf

Appendix 3.3 - Google Sets Configuration

File location: Appendix\Appendix.3\Appendix 3.3 - Google Sets Configuration.pdf

Appendix 3.4 - Detailed Coding for this research's experiment

File location: Appendix\Appendix.3\Appendix 3.4 - Detailed Coding.pdf

Appendix 3.5 - Experiments Results

File location: Appendix\Appendix.3\Appendix 3.5 - Experiments Results.pdf