


|                    |   |
|--------------------|---|
| University Library |  Loughborough University |
| Class              | T   |
| Acc. No.           | 0402594428  |
| Date               | 11/11   |

0402594428



**Reducing Information Overload by Optimising Information Retrieval Approaches**

by

**Stephen Smith**

**A Doctoral Thesis**

**Submitted in partial fulfilment of the requirements**

**for the award of**

**Doctor of Philosophy of Loughborough University**

**November 2010**

**© by Stephen Smith 2010**

# Abstract

---

The information within an organisation forms a fundamental part of its success. In recent years the volume of information housed and processed by organisations has increased exponentially and grown to such a rate that it can be difficult to harness and make successful use of that information. This growth of information has led to the increasing prevalence of the concept of information overload. Although information overload is not a new concept, it is still considered a large-scale problem, with its effect upon the workplace and employees becoming increasingly detrimental. With the increase in available information comes the potential for increased overload.

This research addresses some of the potential barriers that may exist preventing effective discovery, storage and sharing of information and thus increasing the information overload problem. The research presents a framework to investigate several areas of information retrieval and highlights that by reducing the barriers that may exist at each stage the problem of information overload can be addressed in a systematic way, presenting a number of potential solutions.

The research looks at several key areas, knowledge sharing, searching and tagging and the use of ontology's to provide potential solutions to barriers that prevent the effective communication of relevant information. With a reduction of these barriers comes an increase in relevant information helping to decrease the problem of information overload and the pitfalls associated with it.

# Table of Contents

---

|          |  |            |
|----------|--|------------|
| <b>1</b> | <b>Introduction .....</b>  | <b>1</b>   |
| 1.1      | <i>Preface .....</i>   | <i>1</i>   |
| 1.2      | <i>Background .....</i>  | <i>1</i>   |
| 1.3      | <i>Aim and Objectives .....</i>  | <i>5</i>   |
| 1.4      | <i>Research Environment.....</i>   | <i>6</i>   |
| 1.5      | <i>Thesis Outline.....</i>   | <i>7</i>   |
| 1.6      | <i>Summary.....</i>  | <i>8</i>   |
| <b>2</b> | <b>Literature Review .....</b>   | <b>9</b>   |
| 2.1      | <i>Preface .....</i>   | <i>9</i>   |
| 2.2      | <i>Information Overload.....</i>   | <i>9</i>   |
| 2.3      | <i>Knowledge.....</i>  | <i>21</i>  |
| 2.4      | <i>Knowledge Sharing .....</i>   | <i>32</i>  |
| 2.5      | <i>Retrieving and Navigating Information .....</i>                       | <i>40</i>  |
| 2.6      | <i>Search Engines and User Performance.....</i>                          | <i>41</i>  |
| 2.7      | <i>Tagging.....</i>  | <i>58</i>  |
| 2.8      | <i>Semantic Technologies and Ontology.....</i>                           | <i>65</i>  |
| 2.9      | <i>Conclusion .....</i>  | <i>75</i>  |
| <b>3</b> | <b>Research Methodology.....</b>   | <b>78</b>  |
| 3.1      | <i>Research Philosophies.....</i>  | <i>78</i>  |
| 3.2      | <i>Research Methodologies and Approaches.....</i>                        | <i>80</i>  |
| 3.3      | <i>The Research Framework.....</i>                                       | <i>95</i>  |
| 3.4      | <i>Summary.....</i>  | <i>105</i> |
| <b>4</b> | <b>Assessing the Knowledge Sharing Environment.....</b>                  | <b>107</b> |
| 4.1      | <i>Chapter Preface .....</i>   | <i>107</i> |
| 4.2      | <i>The Need to Assess Knowledge Sharing Barriers.....</i>                | <i>108</i> |
| 4.3      | <i>A Diagnostic Tool – Assessing the Knowledge Sharing Barriers.....</i> | <i>109</i> |
| 4.4      | <i>Capturing the Extent of the Barriers to Knowledge Sharing.....</i>    | <i>110</i> |
| 4.5      | <i>Questionnaire Deployment .....</i>                                    | <i>117</i> |
| 4.6      | <i>Results.....</i>  | <i>122</i> |
| 4.7      | <i>Conclusions .....</i>   | <i>139</i> |

|          |  |            |
|----------|--|------------|
| <b>5</b> | <b>Alternative Search Visualisation – Concept Clouds.....</b>  | <b>142</b> |
| 5.1      | <i>Chapter Preface .....</i>   | 142        |
| 5.2      | <i>The Need to Reduce Information Overload .....</i>   | 143        |
| 5.3      | <i>Conceptualisation of Visualisation Techniques for Improving the Presentation of Search Results.....</i> | 145        |
| 5.4      | <i>Concept Cloud Development.....</i>  | 148        |
| 5.5      | <i>Assessing the Potential and Performance of Concept Clouds .....</i>                                     | 153        |
| 5.6      | <i>Reviewing the Recommendations for Building and Implementing Visualisation Systems.....</i>              | 167        |
| 5.7      | <i>Conclusions .....</i>   | 168        |
| <b>6</b> | <b>Using Tagging to Discover Networked and Local Information.....</b>                                      | <b>171</b> |
| 6.1      | <i>Chapter Preface .....</i>   | 171        |
| 6.2      | <i>Searching for Local and Networked Information.....</i>  | 172        |
| 6.3      | <i>Could Tagging be applied to a Traditional File System?.....</i>   | 175        |
| 6.4      | <i>Proof of Concept: Building a Tag Based File System .....</i>  | 176        |
| 6.5      | <i>The Application of Tagging in a Business Environment.....</i>   | 183        |
| 6.6      | <i>Conclusions .....</i>   | 193        |
| <b>7</b> | <b>Ontology Development .....</b>  | <b>195</b> |
| 7.1      | <i>Preface.....</i>  | 195        |
| 7.2      | <i>Introduction.....</i>   | 195        |
| 7.3      | <i>Understanding the Requirements.....</i>   | 198        |
| 7.4      | <i>Development of the OntoFarm System .....</i>  | 200        |
| 7.5      | <i>Assessment .....</i>  | 219        |
| 7.6      | <i>Conclusions .....</i>   | 231        |
| <b>8</b> | <b>Conclusions and Recommendations Framework .....</b>   | <b>235</b> |
| 8.1      | <i>Preface.....</i>  | 235        |
| 8.2      | <i>Introduction.....</i>   | 235        |
| 8.3      | <i>Introduction to the Recommendations Framework.....</i>  | 237        |
| 8.4      | <i>Breakdown of the Recommendations Framework.....</i>   | 239        |
| 8.5      | <i>The Recommendations Framework Summary.....</i>  | 243        |
| 8.6      | <i>Meeting the Aims and Objectives.....</i>  | 243        |
| 8.7      | <i>Limitations of research and potential future work.....</i>  | 244        |
| 8.8      | <i>Final Summary.....</i>  | 250        |
| <b>9</b> | <b>References.....</b>   | <b>254</b> |

# 1 Introduction

## 1.1 Preface

This chapter explores the growth of information and information sources within the workplace and introduces the issues associated with information retrieval. The aims and objectives of the proposed research are detailed together with an outline of the thesis.

## 1.2 Background

Information is a fundamental part of an organisation's success. It forms the basis of an employee's knowledge and is vital to the success of the organisation. To illustrate the importance of information within an organisation, Nelson (Nelson 1994) highlights "In today's society, the success and survival of many companies and individuals hinges upon their ability to 'locate, analyze, and use information skilfully and appropriately'."

The amount of information available to organisations is growing rapidly (Nelson 1994). Huge volumes of information are generated every day in even the smallest of organisations. Murray (Nelson 1994) estimated that "In every 24-hour period approximately 20,000,000 words of technical information are being recorded. A reader capable of reading 1,000 words per minute would require 1.5 months, reading 8 hours every day, to get through 1 day's technical output, and at the end of that period, he would have fallen 5.5 years behind in his reading!"

Harvesting and using information is not straightforward process. Kirsh (Kirsh 2000) states "Information is mediated by an ill understood array of technologies, at hand resources and shifting teams of people". Information comes from a wide variety of sources and with this the office is no longer a straightforward and procedural place (Kirsh 2000). People are constantly interrupted, partake in a number of tasks at once and are constantly creating and consuming information in one form of another. As Savolainen (Savolainen 2007) states this problem has become more topical and intensified in recent years, especially due to the Internet.

With such an abundance of information in the workplace it is little wonder that people may become overloaded. The information that is encountered by the average employee can severely outweigh their ability to process that information. Nelson states "Our proficiency at generating information has exceeded our abilities to find, review and understand it." (Nelson 1994).

With so much data available, the modern workplace is deemed overloaded with information and as such the idea of information overload has evolved over the years and is becoming more relevant today. In 1967 Ackoff (Ackoff 1967) described one of the deficiencies with Management Information Systems. "Most MIS [Management Information Systems] are designed on the assumption that the critical deficiency under which most managers operate is the lack of relevant information" (Ackoff 1967). The problem it seems is not a lack of information that causes a dilemma, but too much. The problem that Ackoff described was "an over abundance of irrelevant information" (Ackoff 1967).

Ackoff described the problem of information overload and although this work dates back to 1967, the problem is just as relevant today. Information overload has become a common occurrence in the modern workplace. Academic studies have shown that information overload affects managers in organisations on a daily basis (Farhoomand, Drury 2002)(Kirsh 2000) and can have dramatic effects within an organisation. With the growth of intranets and the Internet, the information overload problem continues to grow. It is becoming more and more challenging for employees to find the information that they need in order to perform their daily activities, especially in knowledge intensive fields (Chen, Dumais 2000)(Kobayashi et al. 2006). This growth of information accentuates the need for employees to be able to filter and find the relevant information amongst all of the information available to them.

In our modern information rich era, search engines are heavily relied upon to help employees find the content that is relevant to them and remove the irrelevant information that leads to overload. Whether searching the entire Internet or a small document store, the process of searching is often extremely similar, especially with regards to the presentation of results. In 2005 Gulli and Signorini

estimated that there were over 11.5 billion indexable pages on the Internet (Gulli, Signorini 2005). More recent estimations, using similar methods, place this figure close to 60 billion pages (De Kunder 2008). With this volume of content it is clear that information contained will not be relevant to everyone and although the accuracy of these statements is questionable, the Internet remains a formidable corpus of documents that is regularly indexed and searched by millions of users each day using many of the popular search engines.

Retrieving the correct results and presenting them in a way that allows users to discover the information that is relevant to them is not a simple task, even with all of the work that has gone into search algorithms. Many users are still left unable to discover the documents that they need to do their job. "The accelerated growth of the World Wide Web has turned the Internet into an immense information space with diverse and often poorly organized content. Online employees are confronted with rapidly increasing amounts of information as epitomized by the buzzword 'information overload.'" (Hölscher, Strube 2000). Although the core mainstream search engines such as Google, Yahoo and Microsoft Windows Live Search have added small improvements, such as the ability to search within a site or to find documents that are related to the shown, the representation of search results has barely changed since their conception years ago (Wiza, Walczak & Cellary 2004). "Considerable research effort has been invested in the development of efficient methods of collecting and indexing data, algorithms for query processing, as well as data caching mechanisms. The element that has remained almost untouched since the very beginning of the search engines is the presentation interface." (Wiza, Walczak & Cellary 2004). In the majority of search engines the user is presented with the title of the web page followed by a short summary of the web page. Traditionally this would be the first few lines of the document although more recently query based summaries have become popular, showing text that contains or is related to the terms found within the document (Paek, Dumais & Logan 2004).

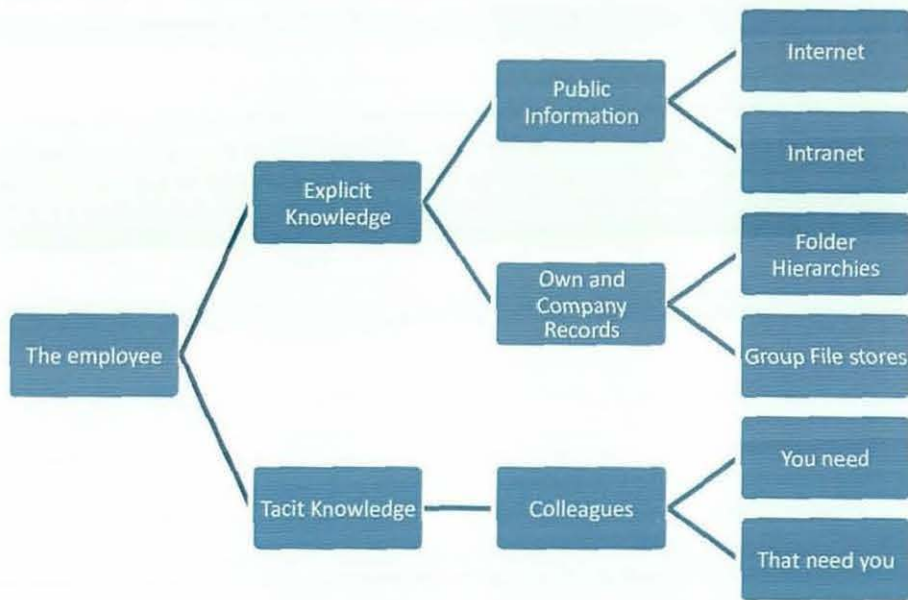
The results are then shown as a long list that the user must then look through on an individual basis in order to find the document or documents that they require. This often results in the user opening a large number of documents in an attempt



to find the information that they feel is relevant to them. In addition to this, research has also shown users often look at just the first two pages of search results (Spink et al. 2001).

“The notoriously low precision of Web search engines coupled with the ranked list presentation make it hard for users to find the information they looking for” (Zamir, Etzioni 1999). Although this quote comes from 1999, only a few years after search engines such as Google became mainstream, the issues still exist today. Over time the precision may have changed slightly yet the number of documents creating ‘noise’ has increased and the presentation of the results remains the same. This lack of change is especially interesting given the economic value associated with the major search engine companies. In February 2008 Microsoft offered 44.6 billion US dollars in an attempt to acquire Yahoo! (Microsoft). Having the advantage in the search engine industry is apparently worth a substantial amount of money and considerable effort is placed into the development of retrieval. Whilst it is clear that the list-based approach to displaying results has limitations, each of the major search engines still adopts this and documents are still difficult to find. Even if the correct result is within a list of search results, it may even be difficult to find the correct document from the list.

The Internet or intranet is just one way employees can find and retrieve information. If working within an organisation, it will also have records that are stored electronically in either hierarchical folders or group file stores. Trying to locate information from these stores can also be difficult due to the different naming schemes given by different employees. Although the information might be there, searching to find it could be both time consuming and fruitless. Seeking tacit knowledge within an organisation can also be a complex and time consuming task. It revolves around who knows how to locate the right person to talk to to locate the crucial information required. For the purpose of the research detailed within this thesis, the author has created a conceptual map of the sources of information that an employee can use to find information, shown by Figure 1. As part of the research, these sources will be explored to determine their impact on employees information effectiveness and efficiency.



**Figure 1 - An overview of the sources of information investigated by the thesis**

### 1.3 Aim and Objectives

The overall aim of this research is to establish the cause and effect of information overload upon information workers, utilising different information sources and to provide potential solutions through an information overload framework.

#### 1.3.1 Objectives

In order to fulfil the aim, the following objectives have been identified

1. Critically review literature on information overload and other information defects and the effect it has upon information workers.
2. To establish through the use of a questionnaire the extent that multi-faceted barriers hinder information and knowledge sharing.
3. To determine how information overload can be reduced through the investigation and development of summarisation techniques.
4. To develop and assess alternative approaches to storing information to improve information retrieval and reduce information overload.
5. To establish the role ontologies can play in the retrieval of relevant information and reduction of information overload, the complexities of ontology development and the barriers to their use.
6. Investigate alternative approaches to traditional ontology development tools that may be used by subject experts rather than ontology specialists to aid in the creation of ontologies that can help the discovery of relevant information.
7. Establish an information overload framework to provide direction and solutions to the information overload problem experienced by information workers.

## 1.4 Research Environment

The research conducted for this thesis was undertaken within two organisations and practice driven. This was because the using information needed to be studied within its natural environment to explore how real organisations use information and the problems they encounter. The two organisations are (actual names of the two companies have been withheld to protect their identities):

**PharmaCo**, is one of the world's leading multinational pharmaceutical companies. They are active in over 100 countries with over 10,000 employees and research and development sites around the world. The key focus of the company is in the research and development of new pharmaceuticals. The company takes the drug development process from start to finish, from initial concepts to clinical trials and tests and further onto marketable medicines. The organisation prides itself on not just performing its core research but also creating a culture and environment in which people are valued and rewarded for their ideas and contributions.

**SoftwareCo**, is one of the largest software organisations in the world. SoftwareCo is within the top 10 of all of the major software rankings including the Forbes2000 and Research Foundation's top 100 with most indexes ranking the organisation in the top 5. The organisation employs over 50,000 people in over 50 countries. The company develops a range of software solutions in house and its products are used globally. The organisation has a number of research and development departments across a number of countries. The SoftwareCo department that participated in this work was one of the rapid development and value prototyping divisions. The department's employees are highly skilled within their respective field and the department has a very unique structure. The department has attempted to create an environment specifically suited to rapid application development, testing and deployment. The department aims to have only a limited hierarchical structure, with all members of the department seen as equals and interacting with each other to take advantage of their respective skills.

## 1.5 Thesis Outline

The thesis consists of eight chapters. The second chapter, the literature review, explores past and present research in the area of knowledge sharing, searching, information overload, tagging and semantic technologies that includes ontologies.

The third chapter consists of the methodology that discusses the research approaches and methods available to the author. The chosen approach and methods are then justified, explaining why they are to be used over other methods.

Chapter Four explores how an information and knowledge sharing environment can be assessed to identify the barriers and the good practice that takes place within organisations, in particular PharmaCo and SoftwareCo. The research in Chapter Four addresses Objective 2, "To establish through the use of a questionnaire the extent that multi-faceted barriers hinder information and knowledge sharing."

Chapter Five focuses upon the use of search engines to discover relevant information. The chapter highlights that although considerable effort has been placed into the search systems themselves, the presentation of results forms a barrier to information retrieval that has barely changed since conception. The research in Chapter Five addresses Objective 3, "To determine how information overload can be reduced through the investigation and development of summarisation techniques."

Chapter Six builds upon the work within chapters four and five and looks at the retrieval of information from a user's own and company records. The chapter investigates the potential of a tagging based file system for the retrieval of documents without having to use a search system and overcome the barriers that users experience. The chapter also investigates the barriers that may exist towards

the use of tagging. The research in this chapter addresses Objective 4, “To develop and assess alternative approaches to storing information to improve information retrieval and reduce information overload.”

The penultimate chapter investigates the final barrier addressed by this thesis. There are benefits that can be afforded by ontologies, especially when combined with tagging. Ontologies, however, can be incredibly resource intensive to create. The chapter proposes and examines a methodology that takes a semi-automated approach to the development of ontologies. The research in this chapter addresses Objectives 5 and 6 “To establish the role ontologies can play in the retrieval of relevant information and reduction of information overload, the complexities of ontology development and the barriers to their use.” and “Investigate alternative approaches to traditional ontology development tools that may be used by subject experts rather than ontology specialists to aid in the creation of ontologies that can help the discovery of relevant information.” respectively.

The final chapter, Chapter Eight, summarises the research contained within the thesis and relates the findings back to the aim and objectives in Chapter One and the literature. The chapter fulfils the final objective, Objective 7. “Establish an *information overload framework to provide direction and solutions to the information overload problem experienced by information workers.*” This chapter also provides recommendations for other organisations on how to reduce information inefficiencies within organisations. Recommendations and suggestions for further research in this area are also included.

## **1.6 Summary**

This chapter introduced the author’s research topic and provided a background of why such research is useful. The aim and objectives of the research were detailed, together with an explanation of the environment in which the research took place. The chapter concluded with an overview of how the thesis is structured, outlining the contents of each chapter.

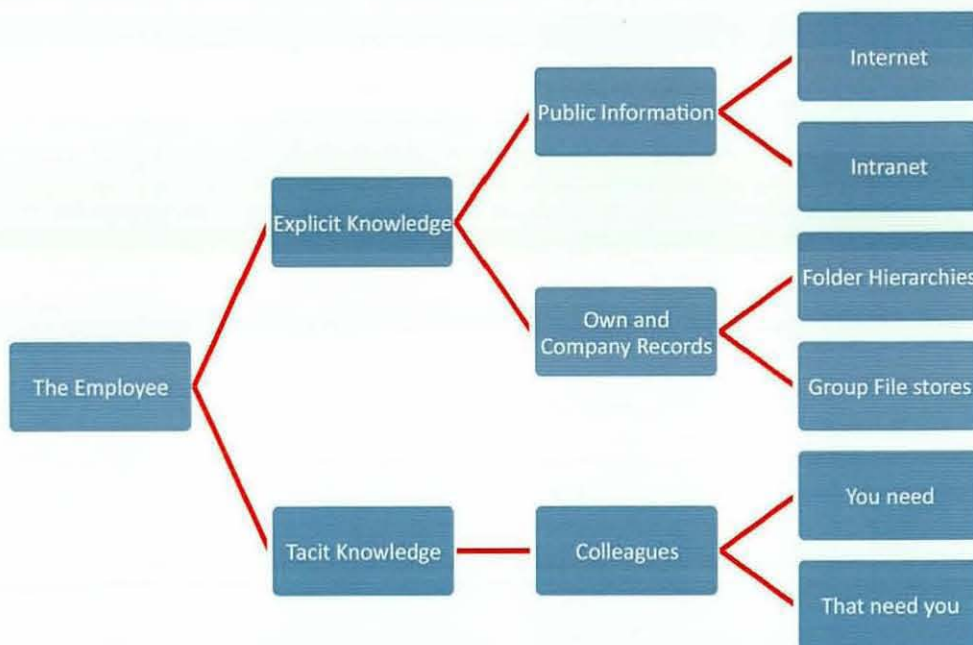
## **2 Literature Review**

### **2.1 Preface**

This chapter analyses the literature and studies relevant to the proposed research and meets Objective One. The chapter begins with an overview of information overload and the impact it can have upon employees' ability to work efficiently and effectively given any information source. The chapter goes on to explore the information defects spectrum that might contribute to information overload or lead to information deficiency such as barriers to sharing information and knowledge, limitations of current technology (visualisation, clustering, tag clouds and ontologies), processes and culture. The chapter concludes with a summary of the literature review and highlights the areas for research.

### **2.2 Information Overload**

This section looks at the role of information, its growth and how it might affect employees. Figure 2 shows the different information sources that might contribute to information overload as shown by the red lines.



**Figure 2 – Potential information sources that might cause overload**

### 2.2.1 Information and our working lives

As already mentioned in Chapter One, information is a fundamental part of an organisation's success. It forms the basis of an employee's knowledge and is vital to the success of the organisation. To illustrate the importance of information within an organisation, Nelson (Nelson 1994) highlights "In today's society, the success and survival of many companies and individuals hinges upon their ability to 'locate, analyze, and use information skilfully and appropriately'."

The amount of information available to organisations is growing rapidly (Nelson 1994). Huge volumes of information are generated every day in even the smallest of organisations. Murray (Nelson 1994) estimated that "In every 24-hour period approximately 20,000,000 words of technical information are being recorded. A reader capable of reading 1,000 words per minute would require 1.5 months, reading 8 hours every day, to get through 1 day's technical output, and at the end of that period, he would have fallen 5.5 years behind in his reading!"

In addition, harvesting and using information is not straightforward. Kirsh (2000) states "information is mediated by an ill understood array of technologies, at hand resources and shifting teams of people". Information comes from a wide variety of sources and with this, the office is no longer a straightforward and procedural place (Kirsh 2000). People are constantly interrupted, partake in a number of tasks

at once and are constantly creating and consuming information in one form of another.

With such an abundance of information in the workplace, it is little wonder that people may become overloaded. The information that is encountered by the average employee can be severely outweighed by their ability to process that information.

Nelson (1994) states "Our proficiency at generating information has exceeded our abilities to find, review and understand it." More recently Hemp (2009) stated a similar thing dating the problem back to the invention of movable type. "Since the invention of movable type led to a proliferation of printed matter that quickly exceeded what a single human mind could absorb in a lifetime".

With so much data available, the modern workplace is deemed overloaded with information and as such, the idea of information overload has evolved over the years and is becoming more relevant today. There are a number of different definitions for information overload, each one dependent on the subject of the overload itself.

Many academics, as shown below, have focused on an individual or organisational group's ability to process information, such as the definition given by Schick, Gordon & Haka (1990):

*"Information overload can occur when the information processing demand on an individual's time for performing interactions and internal calculations exceeds the supply or capacity of time available for such processing".* This definition comes from a literature review and attempt to determine a precise definition of information overload. This definition may however be too precise; many other academics are more generic, often including references to a system being a possible subject of information overload in addition to a human.

"Information overload is the state of an individual (or system) in which not all communication inputs can be processed and utilized, leading to breakdown" (Jones, Ravid & Rafaeli 2004)



“Information overload is the inability to extract needed knowledge from an immense quantity of information for one of many reasons.” (Nelson 1994)

In addition, some authors place less emphasis on the information and more on the individual that is experiencing information overload and symptoms of overload.

“At the personal level, we can define information overload as: a perception on the part of the individual or observers of that person, that the flow of information associated with work tasks is greater than can be managed effectively, and a perception that overload in this sense creates a degree of stress for which his or her coping strategies are ineffective.” (Wilson 2001). However this definition assumes that stress is the only problem caused by information overload at the individual level. This is investigated in more detail later within the literature review.

After reviewing the literature relating to information overload, it appears that although the subject and source of the overload change, based upon the domain in which the definitions come from, the basic principal is always the same. If the information entered into a system is greater than the information it is capable of processing, then an overload will occur.

The word ‘system’ was chosen carefully as it does not restrict the subject of the information overload. For this purpose the author shall use this proposed definition as a generic definition. However the idea of cognitive overload (discussed in the next section) and the fact that the overload is on a human process cannot be ignored.

In addition to the individual, Wilson also looks at how information overload is different at the organisational level.

“At the organizational level, information overload is defined as a situation in which the extent of perceived individual information overload is sufficiently widespread within the organization as to reduce the overall effectiveness of management operations” (Wilson 2001). This definition only focuses on the concept of management operations being affected, however it does provide a sufficient starting point to investigate information overload.

### 2.2.2 Cognitive Overload

One of the earliest pieces of literature referring to the problem of information overload was that of Ackoff (1967). Prior to the work of Ackoff, many managers believed that they were suffering from a lack of relevant information (Ackoff 1967). Whilst Ackoff did not contest that this is an important factor leading to deficiency, he denied that it was the most important factor. He considered it was inferior to another problem. The problem that Ackoff saw was that managers suffered from "an over abundance of irrelevant information". This clear differentiation highlighted a move in thinking from having a lack of relevant information to that of improving the ability to find this relevant information in what he called a "sea of misinformation" (Ackoff 1967).

This concept presented by Ackoff was further investigated by a number of authors (O'Reilly 1980)(Jones, Ravid & Rafaeli 2004) from the perspective of individual members of an organisation and their ability to cope with information.

In 1980, O'Reilly (1980) stated that little attention has been paid to the effect of information overload on employees, and that "Several reasons for this can be offered. First, the intuitive dangers of overload seem obvious, requiring little or no empirical support for substantiation. A familiar complaint among managers and administrators is that of being 'overloaded'. Similarly, the dangers of 'not getting the word' or receiving too little information also are intuitively obvious." (O'Reilly III 1980).

Although there were not many experiments at this time that looked at the impact on individual employees, there were experiments that highlighted the human brains ability to store and process information. O'Reilly appears to acknowledge experiments of this type stating that "Information does however exist in laboratory studies of information and decision making." (O'Reilly III 1980).

Jones, Ravid & Rafaeli (2004) highlight a number of experiments that demonstrate how psychologists have studied the concept of memory and the brains finite storage ability. "Psychologists have long appreciated the limited capacity of people to store current information in memory (e.g., William James in the nineteenth century)." (Jones, Ravid & Rafaeli 2004).

Jones then gives some of the most prominent examples from the field of cognitive psychology that were particularly influential to the concept of information overload and the limitations it creates. "Particularly influential in this regard were Miller's (1956) idea that we can process seven chunks of information (plus or minus two) and Broadbent's (1958) filter model of attention. In the physical or so called 'real world,' the maximum density and geographic spread of a culture's settlements are also linked to the management of information overload (Fletcher 1995)." (Jones, Ravid & Rafaeli 2004).

Before O'Reilly's work, two things were clear. Firstly, increased amounts of relevant information lead to better performance with regards to decision making. Secondly, irrelevant information would decrease the ability to make decisions (O'Reilly III 1980). Given this research, it might be inferred that the information overload problem stems from the inability of the human brain to cope with too much information at once. Although having access to more information appears to help, there does become a point where too much becomes of no benefit. In addition to this, irrelevant information has a direct, negative effect on the ability of an individual to make decisions.

The ability of employees to make decisions appears to be the focus of much research relating to information overload (Janssen, de Poot 2006)(Farhoomand, Drury 2002)(Hemp 2009). The work relating to concept of cognitive overload focuses primarily on the ability to make decisions whilst experiencing cognitive overload. This is illustrated in the following quote: "The main focus of these disciplines is the question of how the performance (in terms of adequate decision making) of an individual varies with the amount of information he or she is exposed to" (Eppler, Mengis 2003). Hemp further highlights this point stating directly that an increase in irrelevant information adversely affects not only the person themselves but also their ability to make decisions (Hemp 2009).

### **2.2.3 The cause and growth of information overload**

The problem of information overload has become widely recognized within today's information-intensive society (Janssen, de Poot 2006). The word 'society' chosen by Janssen and de Poot appears to be an interesting choice. Although

fundamentally information overload manifests itself as cognitive overload, it is the interaction with others that causes the overload. "For communities to function, individuals have to interact with each other. However, interaction involves the strain of dealing with other people, the effort of coping with the products of group activity such as noise and trash, and the effort we must expend to make communication possible" (Jones, Ravid & Rafaeli 2004).

The most common source of information overload described in literature is the Internet. The Internet represents a huge body of information that is increasing by the second. In 2005, Gulli and Signorini estimated that there were over 11.5 billion indexable pages on the World Wide Web (Gulli, Signorini 2005). Using similar methods, a more recent estimate by De Kunder (2008) places this number to be closer to 60 billion pages. With this in mind it seems appropriate that the Internet is often known as the "Information-Superhighway".

In addition to the World Wide Web, which was the focus of many authors research, such as Wiza, Walczak & Cellary (2004), Jones, Ravid & Rafaeli (2004) and Nelson (1994), a number of authors emphasised the concept of communications overload from sources such as email (Hemp 2009)(Janssen, de Poot 2006). Although communications overload is a valid problem, adding to the problem of information overload, it falls outside the scope of this work.

Information overload is now accepted as a known element of working life in many organisations. In a survey of a number of managers conducted by Faroomand in 2002 (Farhoomand, Drury 2002) over 50% of the study's respondents stated that they encounter information overload regularly. "Just over 33% reported experiencing information overload every day. We found the frequency of information overload to be statistically independent of subject gender, age, organizational level, or years of experience." (Farhoomand, Drury 2002).

One of the primary causes stated is that the volume of information available is too high (Nelson 1994). Although as Ackoff (1967) stated, and the literature already confirmed, it is the irrelevant information that causes the problem and not simply that there is too much information. However as the volume of information available expands, it becomes more and more difficult to find the relevant

information amongst the irrelevant. In modern organisations, we are forced to search through larger and larger quantities of data to locate the small piece that we might actually need (Nelson 1994). This point is reaffirmed by Hemp who states that the mass of information, received daily, appears to produce more harm than gain (Hemp 2009) highlighting that the problem is not getting any better.

Although technology may have accentuated the information overload problem, it is not merely the technology that is at fault. There are underlying factors, which are discussed in detail in this chapter, and technology merely provides the channels and mechanisms through which information is distributed or accessed (Wilson 2001).

Nelson argues that greater issues are created as the volume of information gets larger. The information may be full of inaccuracies and inconsistencies between the data, with some data contradicting the other. To illustrate the severity of this problem, Nelson references Naisbitt (Nelson 1994) arguing that "Inundated with technical data, some scientists claim it takes less time to do an experiment than to find out whether or not it has been done before."

In addition to the idea that volume of information is the cause of information overload, Wurman (1989) offers a number of other suggestions as to why information overload may occur. Wurman states that information overload can occur when a person:

- does not understand available information.
- feels overwhelmed by the amount of information to be understood.
- does not know if certain information exists.
- does not know where to find information.
- knows where to find information, but does not have the key to access it.

Taking a different approach to understanding information overload, Farhoomand and Drury (2002) asked a number of managers for their perspective and what information overload meant to them. Farhoomand and Drury found that "the most frequently cited meanings for information overload were an excessive volume of information (79%), difficulty or impossibility of managing it (62%), irrelevance or unimportance of most of it (53%), lack of time to understand it (32%), and multiple sources of it (16%)" (Farhoomand, Drury 2002).

An important observation made by Nelson (1994) was that information does not have the same value to each person. If all information has equal value to everyone then controlling the volume of information may be far easier, however this is not the case and thus users must search for the information that is relevant to them. Improving a user's ability to search for and find the relevant information to them could help to overcome the information overload problem.

Although there are many causes of information overload, the primary cause that academics focus upon is the volume of information. Ultimately, as the volume of information increases so do the other causes of information overload. As such, the volume of information itself can be seen as the key factor contributing to information overload. This is made worse by the growing mass of information available to organisations. Often this is the same information that these organisations see as being key to their success.

In the past decade, there has been an explosion of information. A number of authors make reference to the drastic increase in the amount of information that has become available in recent years. To illustrate this point, Wurman (1989) gives the example that in the present time, a single newspaper contains more information that someone in seventeenth century England might come across in an entire lifetime. In addition to this, Wurman also states that more information has been produced in the last three decades than in the last five millennia with almost 1,000 books being published daily and 9,000 periodicals are published in the United States per year (Wurman 1989).

Farhoomand and Drury (2002) actually state that as more and more technology develops, the channels of information increase, making the problem of information overload worse. In many cases previously, new technology would simply replace older technology making processes more efficient, recently however the growth of technology has lead to a whole new era of communications technologies that add to the media channels available. *Information can come from a wide variety of sources simultaneously (Farhoomand, Drury 2002).*

Although there has been a general growth in the information available and a great increase in communication, after reviewing the literature in this field, it appears

that the growth in the volume of information is often attributed to technology and more specifically the Internet (Nelson 1994).

#### 2.2.4 Push and Pull

In relation to the Internet, the idea of 'push technology' has a specific meaning and is used to facilitate distribution of information (Wilson 2001). This delivery is often instantaneous and seen as a good way to provide the most up to date and relevant information however it strongly contributes to the information overload problem.

Wilson's research showed that push technology generally works the best when information is accessed and acted upon immediately. Significant research exists within the field of push information overload and much of the research relating to information overload actually relates to the idea of pushed information and often a communications overload. Solutions often focus on training and locating the largest disseminators of irrelevant information (Wilson 2001). A key issue is that pushed information is information that the recipient has little control over. Pulled information is information that is on demand and can be searched for and used when required by the recipient (Kirsh 2000). With this in mind, it seems strange that relatively little work exists relating to information overload and pulled information. Many authors ignore the differentiation entirely and even those that do, rarely propose little in the way of a solution. Wilson however highlights the differences between push and pull information and states that pulled information is related to what researchers have called the "need for cognition". The need for cognition is the extent to which people feel they need information and therefore seek that information in order to understand something more fully. Many organisations have withdrawn push technology in favour of information pull.

However, this does not prevent people acting as pushers of information, whenever someone sends a document, whether electronically or otherwise, they are pushing information. This may not form a problem when performed on a need to know basis, however, there is of course the potential to cause disruption. (Wilson 2001). Wilson (2001) also states that it can be hypothesized that managers will have a high need for cognition and therefore drive to find more information and increase their understanding. "A key point is that, the more uncertain their life-world, the

more they will be driven to do this". This need for understanding is not however unique to managers and it could also be hypothesised that the desire would be especially high within knowledge workers such as those within PharmaCo and SoftwareCo, the two organisations that will be used within the authors research.

The interesting debate is determining which of the technologies should be used, push or pull, and when. Research by Kirsh (2000) showed that both pushing and pulling of information is a problem that needs to be addressed, as employees find it difficult to decide how and when to use such systems to gain information.

### **2.2.5 The effect of information overload**

The information overload problem does impact upon a user's ability to find the information that is relevant to them. In addition to not being able to find information, there are more factors related to information overload. Many arguably more severe symptoms can manifest such as decreased job satisfaction, stress, and performance loss (Janssen, de Poot 2006). Farhoomand and Drury (2002) identified a number of factors caused by information overload by surveying managers about the effects of information overload. One respondent wrote, "Information overload causes delays, mistakes, and nonperformance. Eventually it erodes the quality of work. My efficiency is decreased, and I find it hard to prioritize my tasks." The main comments given by each of the studies' participants were associated with the loss of time or that they became frustrated, tired, stressed and even panicked by the overload. A minority even felt that information overload damaged their personal lives. One participant stated "It leads to frustration and confusion. It can make me feel restless, anxious, and sometimes panicky. The worst is the discouraging effect on my commitment to my job." Again there was no direct association between the effects of information overload and the gender, age, organisational level or experience of the respondent. Farhoomand and Drury's (200) research would suggest that information overload can affect anyone, but the research only questioned managers of organisations and does not deal with all employees. A study with a wider ranging of managerial roles conducted by Waddington in 1996 is more frequently referred to in literature (Kirsh 2000)(Wilson 2001). The study found a number of issues created by



information overload. The study included 1,313 participants from the UK, U.S. Australia, Hong Kong and Singapore. The study found that

- Two thirds of managers report tension with work colleagues, and loss of job satisfaction because of stress associated with information overload.
- One third of managers suffer from ill health, as a direct consequence of stress associated with information overload. This figure increases to 43% among senior managers.
- Almost two thirds (62%) of managers testify their personal relationships suffer as a direct result of information overload.
- 43% of managers think important decisions are delayed, and the ability to make decisions is affected as a result of having too much information.
- 44% believe the cost of collating information exceeds its value to business.

The findings of this work actually contradict those of Farhoomand and Drury and show that personal relationships are harmed far more than the pair suggest. This may be because within the Farhoomand and Drury study, employees placed the issue of loss of time above that of strained relationships. However, the findings highlight the severity of the information overload problem. In addition Waddington's (1996) work also only appears to focus upon managers, although high, middle and junior level managers were included. There appears to have been very little research that examines the effects of information overload upon varying job roles.

### **2.2.6 Summary**

This section on information overload has identified the factors associated with it and has reviewed and presented a number of definitions. The literature suggests that the root of information overload appears to lie within the cognitive overload domain and that employees can only process a finite amount of information. An increase in relative information will aid a user in the decision process, but when irrelevant information is present, it is extremely detrimental to the decision process.

The literature review found that information overload has existed for a number of years and advances in technology have added new sources of information, both relevant and irrelevant, which rapidly add to the information overload problem. For example, the Internet in its entirety has created an information super highway with large amounts of information being created every day. It is becoming more and more difficult to find the relevant information amongst the irrelevant. Although no literature could be found relating to the effect of information overload on all employee roles, the effects of overload upon managers have been determined. Information overload has some serious consequences for modern managers, with information overload causing a lack of time, delaying decisions and even in many cases affecting the personal lives and relationships of those managers.

Two states of information overload were identified and these were driven by the concepts of pulling and pushing information. A number of studies have focused on the concept of pushing information and overload, but relatively few discuss the domain of pulling information and this is where the recipient of the information would have the most control.

What has not been discovered within the literature is how information overload can be tackled as a whole. The multifaceted approach to reducing information overload caused by technology and improving the interactions of colleagues could present possible benefit to organisations and would represent new research within the field of information overload. Information overload has serious impacts upon an organisation and it appears that if the information overload problem can be tackled then there will be significant advantages for both employees and the organisation and that work within this area might be relevant to current research.

### **2.3 Knowledge**

In order to understand how knowledge and information relate to each other and to better understand the information overload problem, it is important to define what is meant by knowledge, information and other related terms such as data, understanding and wisdom.

One of the definitions relating to information overload actually focuses on an inability to turn information into knowledge. "Information overload is the inability to extract needed knowledge from an immense quantity of information for one of many reasons." (Nelson 1994). Without a fuller understanding of information and knowledge, it might be difficult to fully understand what this means. For this reason, a literature review relating to information and its associated terms was performed and is detailed in the following section.

### **2.3.1 Wisdom, Understanding, Knowledge, Information and Data**

It is important to understand what is meant when talking about knowledge or information and the differentiation between words such as knowledge, information, data and even wisdom. There are various interpretations relating to these concepts but it is important to identify what is meant by each in the context of this work and where some of the definitions used have come from. There has never been a single agreed definition of knowledge but an overview of some of the more accepted definitions is discussed below.

Ackoff (1989) described five stages or categories that content within the human mind could be classified into: data, information, knowledge, understanding and wisdom. These are often referred to as DIKW, with understanding removed from the hierarchy. The origins of this differentiation are actually attributed to a poem called 'The Rock' by T.S. Eliot (Sharma 2004).

"Where is the Life we have lost in living?  
Where is the wisdom we have lost in knowledge?  
Where is the knowledge we have lost in information?"  
(Eliot 1934)

Eliot not only clearly differentiates between wisdom, knowledge and information but also suggests a hierarchy, even giving the impression that one may be directly influenced or changed into another (Hey 2004). The poem has since become the starting point for any discussions relating to the differentiation between information (Hey 2004)(Sharma 2004).

Bellinger, Castro & Mills (2004) further expanded the definitions given by Ackoff (1989). Using both Bellinger and Ackoff's work, the following definitions are presented for Ackoff's terms:

- **Data** – The raw form of content or the symbols that we interact with. “It simply exists and has no significance beyond its existence (in and of itself). It can exist in any form, usable or not. It does not have meaning of itself. An example of data may be the content of a spreadsheet. By itself it has no meaning.” (Bellinger, Castro & Mills 2004).
- **Information** – Data that is or can be processed and may lead to something that is useful. The word may is important here. “Information is data that has been given meaning by way of relational connection. This ‘meaning’ can be useful, but does not have to be” (Bellinger, Castro & Mills 2004).
- **Knowledge** – Knowledge occurs when the information is actually applied and interpreted to use the information in some form. “Knowledge is the appropriate collection of information, such that its intent is to be useful. Knowledge is a deterministic process. When someone “memorizes” information, then they have amassed knowledge. This knowledge has useful meaning to them, but it does not provide for, in and of itself, an integration such as would infer further knowledge” (Bellinger, Castro & Mills 2004).
- **Understanding** – Understanding is when the knowledge is taken a stage further. The subject understands it and why it may happen. “Understanding is an interpolative and probabilistic process. It is cognitive and analytical. It is the process by which I can take knowledge and synthesize new knowledge from the previously held knowledge. The difference between understanding and knowledge is the difference between ‘learning’ and ‘memorising’. People who have understanding can undertake useful actions because they can synthesize new knowledge, or in some cases, at least new information, from what is previously known (and understood). That is, understanding can build upon currently held information, knowledge and understanding itself.” (Bellinger, Castro & Mills 2004).

- Wisdom – Wisdom is at the top of the chain. Wisdom is an evaluated understanding allowing the subject to use, interpret and possibly derive new knowledge from existing knowledge. “Wisdom is an extrapolative and non-deterministic, non-probabilistic process. It calls upon all the previous levels of consciousness, and specifically upon special types of human programming (moral, ethical codes, etc.). It beckons to give us understanding about which there has previously been no understanding, and in doing so, goes far beyond understanding itself. It is the essence of philosophical probing. Unlike the previous four levels, it asks questions to which there is no (easily-achievable) answer, and in some cases, to which there can be no humanly known answer period. Wisdom is therefore, the process by which we also discern, or judge, between right and wrong, good and bad.” (Bellinger, Castro & Mills 2004).

Understanding relates to a human process required to utilise the knowledge that may be provided and as such, it is clear to see why it may be ignored when discussing information, knowledge management and associated concepts.

However, wisdom is an important factor. For this reason the hierarchy is often abbreviated to DIKW in literature and not DIKUW. Ackoff (1989) indicated that the first four of these definitions also really referred to the past, whereas wisdom refers to the future. It is important that we look towards creating wisdom through enabling people to discover and make use of information to the best of their ability and ultimately gather knowledge in an effective manner. Interestingly often literature only considers the first three of Ackoff's (1989) terms: data, information and knowledge. Although these definitions were given in 1989, they are still relevant today and are often discussed and cited in more recent articles (Hey 2004),(Sharma 2004). The acceptance of Ackoff's (1989) hierarchy is not universal however. There are authors such as Fricke (2009) who disagree with the relevance of the hierarchy.

### **2.3.2 Data and Information definitions**

In 1998, Davenport and Prusak gave similar definitions to those given by Bellinger (Bellinger, Castro & Mills 2004). Their definitions however were focused more towards the business perspective. Data is described by Davenport and Prusak

(1998) as “a set of discrete, objective facts about events. In an organisational context data is more usefully described as structured records of transactions”. Data has no further meaning associated with it; it is simply a fact that has been recorded. Data is often thought of as having little relevance or purpose by itself as there is no context associated with it.

Information was once said to be “data endowed with relevance and purpose” and has been described as “data that makes a difference (Davenport, Prusak 1998). “Information is meant to shape the person who gets it, to make some difference in his outlook or insight”. This also gives rise to the concept that it is the receiver of the information that decides whether or not this is information or simply data.

Information can be transported in a number of ways including ‘hard and soft networks’ (Davenport, Prusak 1998). Formal methods tend to make up hard networks. They are networks with a definite infrastructure such as wires, satellite dishes and even delivery vans. Email, normal postal mail and Internet transmissions are all examples of information sent using hard networks.

Soft networks are far less formal, for example, a conversation at a coffee machine although it could be argued that the coffee machine was placed there to facilitate message transmission. A note marked ‘for your information’ is also an example of information transmitted via soft networks.

For the purposes of the research presented in this thesis, the above definitions for data and information by Davenport and Prusak will be used as the reference point, because they present enough depth to proceed and allow sufficient detail to differentiate. The definition of knowledge is tackled in the following chapter.

### ***2.3.2.1 Knowledge definitions***

The definition of knowledge is perhaps a more difficult issue to agree on given the large number of definitions available. In 2001, Firestone argued that there were a number of different perspectives towards knowledge, each depending on the ‘world’ from which they are viewed (Firestone 2001). Within this, he also presents a vast number of references to definitions of knowledge and provides a very in depth discussion. One of the most widely known definitions of knowledge comes

from Nonaka and Takeuchi and states "Western philosophers have generally agreed that knowledge is 'justified true belief' a concept that was first introduced by Plato in his *Meno, Phaedo and Theaetetus*". (Nonaka, Takeuchi 1995b). This definition however is quite ambiguous by itself, because it gives rise to the question of what is true and justified, which in turn are difficult questions.

A further working definition of knowledge, based upon this definition and others, was presented by Alavi and Leidner (1999) "Knowledge is a justified personal belief that increases an individual's capacity to take effective action". Another interesting statement made by Alavi and Leidner is that "knowledge is not a radically different concept than information, but rather that information becomes knowledge once it is processed in the mind of an individual". The definitions given by Nonaka and Takeuchi (1995b) or Alavi and Leidner (1999) do agree to a certain extent with those of Ackoff (1989) but take the definition further stating that knowledge is in the mind of the individual.

This concept is further clarified by Wilson (2002) who states that "'Knowledge' is defined as what we know: knowledge involves the mental processes of comprehension, understanding and learning that go on in the mind and only in the mind, however much they involve interaction with the world outside the mind, and interaction with others. Whenever we wish to express what we know, we can only do so by uttering messages of one kind or another - oral, written, graphic, and gestural or even through 'body language'. Such messages do not carry 'knowledge', they constitute 'information', which a knowing mind may assimilate, understand, comprehend and incorporate into its own knowledge structures."

In addition a crucial part of this definition is that the knowledge structures and previous knowledge known by an individual will always be different to those of another individual. As such the knowledge built from these 'messages' will not be the same for the receiver, as the knowledge that represents the source (Wilson 2002).

This definition is similar to that of Davenport and Prusak (1998), however there is a fundamental difference between the two. Davenport and Prusak give the following definition for knowledge. "Knowledge is a fluid mix of framed

experience, values, contextual information, and expert insight that provides a framework for evaluating and incorporating new experiences and information. It originates and is applied in the minds of knowers. In organizations, it often becomes embedded not only in documents or repositories but also in organisations routines, processes, practices and norms." (Davenport, Prusak 1998).

The difference in this definition does not actually apply so much to the definition of knowledge but rather the acknowledgement by Wilson that the 'messages' which carry knowledge are simply information. It could be inferred that Davenport and Prusak (1998) believe that the 'messages' contain knowledge. Since the definition given by Wilson is similar to that of Davenport and Prusak, although less ambiguous, the definition given by Wilson (2002) shall be used for the purposes of research in this thesis.

### 2.3.3 Types of Knowledge: Explicit and Tacit

Perhaps one of the most frequently occurring ways to categorise knowledge is credited to Polanyi (1967). Polanyi separated knowledge into two forms: tacit knowledge and explicit knowledge. Polanyi stated that "we can know more than we can tell" (Polanyi 1967). He then gives a thought provoking example "We know a person's face, and can recognize it among a thousand, indeed among a million. Yet we usually cannot tell how we recognize a face we know. So most of this knowledge cannot be put into words."

A good example of this is provided by Polanyi's research. Through the use of a large collection of images, containing mouths, noses and other features, a police witness can select pieces of a face that they know. These pieces can be put together to give a reasonable likeness of the face being described. Thus, the knowledge that could not be communicated becomes communicable. "But the application of the police method does not change the fact that, previous to it, we did know more than we could tell at the time. Moreover, we can use the police method only by knowing how to match the features we remember with those in the collection, and we cannot tell how we do this." This is a prime example of knowledge that we have but cannot tell. This type of knowledge is called tacit knowledge. Polanyi (1967)



subsequently describes this as the “knowing what” of Ryle (1946). Tacit knowledge could be described as something that is held in the head of an individual and is more difficult to articulate and express. Tacit knowledge is therefore much harder to capture, re-use and share. Tacit knowledge can also be termed implicit knowledge. Nonaka and Takeuchi (1995b) describe tacit knowledge as ‘Something that is not easily visible and expressible’ and is very difficult to formulate and communicate with others. As it is stored in the head of the individual, it is “deeply rooted in their experience, as well as in the ideas, values, or emotions he or she embraces”.

Explicit knowledge is the other form of knowledge and can be captured and is communicable. This is aligned to the “knowing how” of Ryle. Explicit knowledge is the simplest form of knowledge, it is easy to express, can be stored and easily communicated. It is knowledge that can be captured in some way and expressed in any manner of ways including words and numbers. For example, the knowledge found within an encyclopaedia or stored on a corporate intranet or the Internet is by definition explicit knowledge. Going back to the definition given by Wilson (2002), although the medium that this knowledge is transferred may be information, it is still explicit knowledge, knowledge that is easy to express, that represents the source of this information.

In their book ‘The Knowledge Creating Company’, Nonaka and Takeuchi (1995b) state the Japanese companies have a very different understanding of knowledge (to those of the Western world) and that Japanese companies recognise that the knowledge expressed in words and numbers is only the beginning. They then go on to state that knowledge is primarily tacit and that due to its very nature, it is significantly different to the explicit knowledge “Explicit knowledge can be easily processed by a computer, transmitted electronically, or stored in databases”. This makes it very easy to work with. An important point to clarify is that again the knowledge may be represented as information or even data in the computer but it was, however, knowledge in the head of the originator.

Although some may argue that tacit knowledge is actually information and not knowledge at all, Polanyi argued to the contrary, stating that in both “knowing

what” and “knowing how”, we always speak of knowing and that “knowing” would be used to cover both the practical and theoretical knowledge. This is re-affirmed by Nonaka and Takeuchi when they state that “he [Polanyi] observes that science is operated by the skill of the scientist and it is through the exercise of this skill that he shapes his scientific knowledge. This suggests both a view of knowledge as an object and of knowing as an action of enactment in which progress is made through active engagement with the world on the basis of a systematic approach to knowing.” (Nonaka, Takeuchi 1995b).

Nelson and Winter (1982) referred to tacit knowledge not as a wholly incommunicable thing, but that it may have a degree of 'tacitness' that is dependent on its ability to be codified and abstracted. This leads to a very important concept. In order for tacit knowledge to be useful to more than just the person who holds it, it must be converted into explicit knowledge, which may be communicated as information. In order to be fully useful, it must then be converted back into tacit knowledge for use by the second person. This is the key to knowledge transfer and only when this process occurs, can tacit knowledge truly become useful. The concept of converting one form of knowledge to another such as from tacit knowledge to explicit is not a new one and has existed for years. It is highlighted with the example of a master and apprentice. This classical example shows the conversion of knowledge to allow transfer from one employee to another. The conversion of knowledge and the process of codification are seen by many as the key to knowledge transfer. However, the discussion on codification and knowledge transfer will take place later in this chapter. Conversion of tacit and explicit knowledge can take four forms according to Nonaka and Takeuchi (1995b).

- **Socialisation** – from **Tacit** to **Tacit** – Sympathised Knowledge
- **Externalisation** – from **Tacit** to **Explicit** – Conceptual Knowledge
- **Combination** – from **Explicit** to **Explicit** – Operational Knowledge
- **Internalisation** – from **Explicit** to **Tacit** – Systemic Knowledge

Table 1 shows a grid of the conversions described by Nonaka and Takeuchi. The four key concepts are outlined with the conversions they represent.

**Table 1 – Four Modes of Knowledge Conversion** (Nonaka, Takeuchi 1995b)

|                         | To Tacit Knowledge     | To Explicit Knowledge  |
|-------------------------|------------------------|------------------------|
| From Tacit Knowledge    | <b>Socialisation</b>   | <b>Externalisation</b> |
| From Explicit Knowledge | <b>Internalisation</b> | <b>Combination</b>     |

### 2.3.3.1 Socialisation – from Tacit to Tacit

“Socialisation is the process of sharing experiences and thereby creating tacit knowledge such as shared mental models and technical skills” (Nonaka, Takeuchi 1995b). It is also possible to learn from a person without using language or formal communication, in the way that an apprentice learns by observation, imitation and practice. “The key to acquiring Tacit Knowledge is experience” (Nonaka, Takeuchi 1995b). Simply transferring tacit information from one individual to another would make little sense without the context in which it comes from. It is the ability to share in that experience and see it from the other person’s point of view or perspective, which gives value to the information.

Taking another person’s perspective on a project or decision is often extremely important. Their previous experience may be able to show them an important aspect that they had ignored, or simply not even thought of. Taking into account the view of a customer is also a form of socialisation. In order to understand the product that they require one must learn to identify with them and their needs.

### 2.3.3.2 Externalisation – from Tacit to Explicit

Externalisation is the process of forming explicit and therefore describable and tangible assets from tacit knowledge. Externalisation is thought by many to be the key of many knowledge management programmes as it is the act of making something useful and sharable, something explicit, from tacit knowledge. It is often

centred on the concepts of metaphors, analogies, concepts, hypotheses, or models anything that will enable the expression of the tacit knowledge in a formal way or will provide the ability to capture that knowledge.

Often it can be difficult to capture the tacit knowledge and “expressions are often inadequate, inconsistent, and insufficient” (Nonaka, Takeuchi 1995a) .

When expression becomes difficult, it is often the case that metaphors and analogies become useful to aid the expression and allow a clearer picture to be perceived. Metaphors do however, have obvious differences to the real world situation being described and make no effort to show the differences. This is where analogies are useful to point out these differences.

Nonaka and Takeuchi describe metaphors as the association of two things driven mostly by intuition and holistic imagery and does not aim to find the differences between them. Association by analogy focuses on rational thinking and the structural and functional similarities between the two and their differences (Nonaka, Takeuchi 1995a).

#### ***2.3.3.3 Combination – From Explicit to Explicit***

Combination is perhaps one of the simpler methods of knowledge conversion. It is merely the amalgamation of existing data sources through any number of media. It can involve “sorting, adding, combining and categorising of explicit knowledge and may lead to the creation of new Knowledge” (Nonaka, Takeuchi 1995b).

In modern business, databases and document stores are often used to allow the collaboration and combination of documents. They often also provide great search facilities to allow ease of access to this information.

#### ***2.3.3.4 Internalisation – From Explicit to Tacit***

Internalisation is the act of taking explicit knowledge in any media and transferring that into tacit knowledge. “It is closely related to ‘learning by doing’” (Nonaka, Takeuchi 1995a).

“When experiences through socialisation, externalisation, and combination are internalised into the individuals’ Tacit Knowledge bases in the form of shared

mental models or technical know-how, they become valuable assets" (Nonaka, Takeuchi 1995a).

*This is one of the most important stages of knowledge conversion as it leads to knowledge that is easily accessible by the employee and can be used to allow them to complete their job successfully. Documentation of knowledge can also be useful to allow users to recall tacit knowledge or store new tacit knowledge. Documents and manuals are prime examples of how transfer of knowledge can be facilitated.*

An excellent example is that GE (General Electric) stores all customer complaints and inquiries in a database at its answer centre in Louisville. This then allows employees to re-experience what the telephone operators experienced and learn from that occurrence (Nonaka, Takeuchi 1995a).

#### **2.3.4 Summary**

This section has identified the differences between data, information and knowledge, which are essential to any employee in any organisation. It has provided a deeper and richer theoretical understanding of knowledge and in particular explicit and tacit knowledge. It has identified explicit knowledge and tacit knowledge as both communicable and incommunicable knowledge respectively. However, it has also been argued that knowledge may be communicable if information is the medium of transfer. Although the theoretical knowledge foundations have been established, the factors that could affect information and knowledge sharing and the sources still need to be identified. This will be covered in the next section.

#### **2.4 Knowledge Sharing**

This section continues to investigate the sharing of knowledge and information and the requirements to successfully share information. However, what is meant by knowledge sharing, what does it entail and why is it important? Knowledge sharing "refers to activities associated with the flow of knowledge from one party to another. This includes communication, translation, conversion, filtering and rendering." (Newman, Conrad 2000). Knowledge sharing often forms a key part of knowledge management initiatives and the benefits of sharing knowledge are

widely known (Alavi, Leidner 2005)(Nonaka, Takeuchi 1995b). Riege (2005) states that "The principle equation is: better and purposeful sharing of useful knowledge translates into accelerated individual and organisational learning and innovation through the development of better products that are brought faster to a target market, thus enhancing market performance".

The literature suggests (Riege 2005) that knowledge sharing and having associated goals towards the sharing of knowledge between employees is often a forgotten part of a business approach. There are a number of reasons often attributed to this such as the inability to measure knowledge sharing or even that the barriers to sharing knowledge are not sufficiently identified within an organisation (Riege 2005). Riege's research undertook a comprehensive review of the body of literature around knowledge sharing and identified a number of barriers to the knowledge sharing activities.

The value assigned to knowledge sharing is rapidly growing. Nahapiet and Ghoshal (2005) state that they "have noted the significant and growing body of work that indicates organizations have some particular capabilities for creating and sharing knowledge, giving them their distinctive advantage over other institutional arrangements" and argue that it is actually the combination of social and intellectual capital that underpins organisational advantage. The work suggests that those organisations that encourage and promote knowledge sharing will gain a competitive advantage. What is not clear is the best way of encouraging and promoting different types of knowledge sharing.

Maximising the value obtained and derived from the knowledge held by employees is often acknowledged as one of the key challenges that companies have in regards to knowledge sharing. Most importantly it is the tacit knowledge held by employees that is said to hold the key to success (Riege 2005). Building on the work previously shown relating to knowledge conversion methods by Nonaka and Takeuchi (1995b) and Spender (1996) further differentiated the concepts identified by Nonaka and Takeuchi and added another layer. The layer added by Spender took the concepts of both social and individual knowledge into consideration to form a matrix containing four types of knowledge.

**Table 2 - Spender's Individual and Social Knowledge (Spender 1996)**

|          | Individual   | Social   |
|----------|--|--|
| Tacit    | Individual Tacit Knowledge<br>(Automatic Knowledge)    | Social Tacit Knowledge<br>(Collective Knowledge)     |
| Explicit | Individual Explicit Knowledge<br>(Conscious Knowledge) | Social Explicit Knowledge<br>(Objectified Knowledge) |

Table 2 shows Spenders four types of knowledge and they describe the knowledge that can be found in any organisation. The first type is Conscious Knowledge, which is individual explicit knowledge, personal records or memory that is easily storable and retrievable. The second type is Automatic Knowledge, or individual tacit knowledge; this is knowledge based on personal experiences. The third type is Objectified Knowledge, or social explicit knowledge. This is information that is generally available and well documented knowledge available to a collective. The fourth type is Collective Knowledge, or social tacit knowledge. Social tacit knowledge is knowledge embedded into the organisations culture and way that it works.

This social tacit or collective knowledge was argued by Spender(1996) to be the "most secure and strategically significant kind of organisational knowledge". The differentiation between individual knowledge and social or group knowledge here is an interesting and necessary differentiation. However, for knowledge sharing to succeed it will be necessary for employees to recognise the benefits of both. It is important for employees to use their own knowledge to make decisions and perform their job, but it is also important that team based work is recognised and performed.

This work was further developed by Dixon (2000). Dixon identified five different types of knowledge transfer and they are outlined in Table 3.

**Table 3 – Shortened version of Dixon's knowledge transfer types (Dixon 2000).**

| Title                          | Serial Transfer   | Near Transfer   | Far Transfer  | Strategic Transfer   | Expert Transfer  |
|--------------------------------|---|---|---|--|--|
| Definition                     | The knowledge a team has gained from doing its task in one setting is transferred to the next time that team does the task in a different setting | Explicit knowledge a team has gained from doing a frequent and repeated task is reused by other teams doing very similar work | Tacit knowledge a team has gained from doing a non-routine task is made available by other teams doing similar work in another part of the organisation | The collective knowledge of the organisation is needed to accomplish a strategic task that occurs infrequently but is critical to the whole organisation | A team facing a technical question beyond the scope of its own knowledge seeks the expertise of others in the organisation |
| Similarity of task and context | The receiving team (which is also the source team) does a similar task in a new context   | The receiving team does a task similar to that of the source team and in a similar context                                    | The receiving team does a task similar to that of the source team but in a different context  | The receiving team does a task that impacts the whole organisation in a context different to that of the source team                                     | The receiving team does a different task from that of the source team but in a similar context                             |
| Nature of the task             | Frequent, and non-routine   | Frequent and routine  | Frequent and non-routine  | Infrequent and non-routine   | Infrequent and routine   |
| Type of knowledge              | Tacit and Explicit  | Explicit  | Tacit   | Tacit and Explicit   | Explicit   |

The original table created by Dixon contained more rows with further information such as design guidelines and examples. The table shows how tasks performed by two different groups may relate and the transfer between them. In reality, this transfer need not to be restricted to groups and could be applied to individuals in many cases. The table itself was intended by Dixon to highlight two things. Firstly, that there are a number of different types of knowledge transfer methods. Secondly, that “knowledge is transferred most effectively when the transfer process fits the knowledge being transferred. Table 3 is really given as a guide to



show what the most effective method might be when transfer is necessary. But it is not exclusive and in many cases more than one method may be required simultaneously in order to be successful. The literature presents a rich and somewhat complex picture of knowledge transfer. The next section highlights further difficulties that organisations might face when it comes to knowledge transfer.

#### 2.4.1 Knowledge Sharing Difficulties

The key issue with knowledge sharing is that it is very difficult to assess knowledge sharing (Riege 2005) and provide a tangible value and therefore difficult to promote effective knowledge sharing and gain buy in from management. In addition to these problems, Argote and Ingram (2000) argue knowledge sharing is unique to individual organisations with varying levels of success, implying that a one size fits all model will have limited success within organisation. Despite the difficulty in determining the success of knowledge sharing, there are known barriers towards knowledge sharing. There are many systems available for sharing explicit knowledge, such as document repositories but there are often problems associated with them (Riege 2005). One of the key issues is that some of the most important knowledge is not stored explicitly but relies on the knowledge stored tacitly by the employees themselves. Nonaka and Takeuchi (1995b) noted that the sharing of tacit knowledge amongst different individuals becomes a critical step of the creation of new knowledge, however actually encouraging and facilitating this sharing is often more difficult than it first seems.

Although assessing how well knowledge sharing is performed within an organisation is extremely difficult, the barriers to knowledge sharing that exist within an organisation can be addressed. Indirectly, by addressing these issues an assessment of how well an organisation shares knowledge might become apparent. Riege presented a number of barriers in his paper "Three-dozen knowledge-sharing barriers managers must consider" (Riege 2005). The barriers were divided into three different sections, each consisting of between 8 and 17 different barriers, taken from literature. The three categories were as follows: potential individual barriers, potential organisational barriers and potential

technological barriers. Riege acknowledges that the barriers may have a different level of effect within different organisations and whilst one barrier may be of great interest to one organisation, it may be of no significance to another. However, Riege does not suggest which barriers may be of interest to an organisation and makes no real reference to any barrier having more of an influence than another. The barriers identified by Riege (2005) come from an extremely comprehensive literature review and capture the key issues identified from a large list of previous work in a concise format. The barriers are listed below:

#### **2.4.1.1 Potential individual barriers**

1. General lack of time to share knowledge and time to identify colleagues in need of specific knowledge;
2. Apprehension of fear that sharing may reduce or jeopardise people's job security;
3. Low awareness and realisation of the value and benefit of possessed knowledge to others;
4. Dominance in sharing explicit over tacit knowledge such as know-how and experience that requires hands-on learning, observation, dialogue and interactive problem solving;
5. Use of strong hierarchy, position-based status, and formal power ("pull rank");
6. Insufficient capture, evaluation, feedback, communication, and tolerance of past mistakes that would enhance individual and organisational learning effects;
7. Differences in experience levels;
8. Lack of contact time and interaction between knowledge sources and recipients;
9. Poor verbal/written communication and interpersonal skills;
10. Age differences;
11. Gender differences;
12. Lack of social network;
13. Differences in education levels;

14. Taking ownership of intellectual property due to fear of not receiving just recognition and accreditation from managers and colleagues;
15. Lack of trust in people because they may misuse knowledge or take unjust credit for it;
16. Lack of trust in the accuracy and credibility of knowledge due to the source; and
17. Differences in national culture or ethnic background and values and beliefs associated with it (language is part of this).

#### **2.4.1.2 Potential organizational barriers**

1. Integration of km strategy and sharing initiatives into the company's goals and strategic approach is missing or unclear
2. Lack of leadership and managerial direction in terms of clearly *communicating the benefits and values of knowledge sharing practices*;
3. Shortage of formal and informal spaces to share, reflect and generate (new) knowledge;
4. Lack of a transparent rewards and recognition systems that would motivate *people to share more of their knowledge*;
5. Existing corporate culture does not provide sufficient support for sharing practices;
6. Knowledge retention of highly skilled and experienced staff is not a high priority;
7. Shortage of appropriate infrastructure supporting sharing practices;
8. Deficiency of company resources that would provide adequate sharing opportunities;
9. External competitiveness within business units or functional areas and between subsidiaries can be high (e.g. not invented here syndrome);
10. Communication and knowledge flows are restricted into certain directions (e.g. top-down);
11. Physical work environment and layout of work areas restrict effective sharing practices;
12. Internal competitiveness within business units, functional areas, and subsidiaries can be high;

13. Hierarchical organization structure inhibits or slows down most sharing practices; and
14. Size of business units often is not small enough and unmanageable to enhance contact and facilitate ease of sharing.

#### **2.4.1.3 Potential technological barriers**

1. Lack of integration of IT systems and processes impedes on the way people do things;
2. Lack of technical support (internal or external) and immediate maintenance of integrated IT systems obstructs work routines and communication flows;
3. Unrealistic expectations of employees as to what technology can do and cannot do;
4. Lack of compatibility between diverse IT systems and processes;
5. Mismatch between individuals' need requirements and integrated IT systems and processes restricts sharing practices;
6. Reluctance to use IT systems due to lack of familiarity and experience with them;
7. Lack of training regarding employee familiarisation of new IT systems and processes; and
8. Lack of communication and demonstration of all advantages of any new systems over existing ones.

#### **2.4.2 Summary**

This section has shown a number of factors related to knowledge sharing. Firstly, it has shown the potential that exists in knowledge sharing, in particular how effective knowledge sharing can be attributed to commercial competitive advantage, especially within knowledge intensive companies. Secondly, it has identified a number of different methods of knowledge transfer and the difficulties involved, in particular the barriers that exist that may prevent the sharing of knowledge. The barriers fell into one of three key categories:

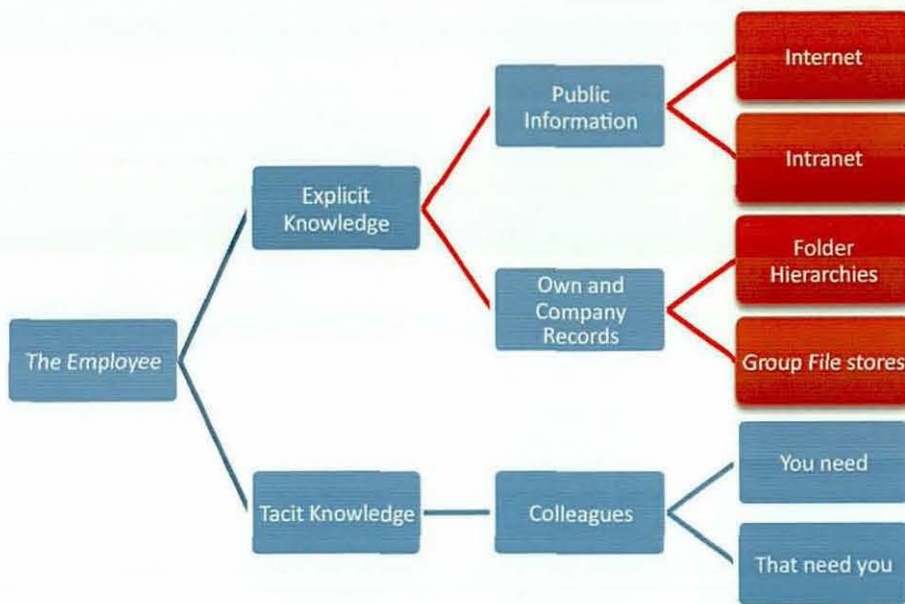
- Potential technological issues
- Potential individual issues
- Potential organisational issues

One of the important points raised about these barriers is that they might affect different organisations in different ways, hence why they are deemed potential barriers. If the transfer of knowledge can be increased via information then it is feasible that there could be an increase in information overload, as more information will be shared in the attempt to transfer knowledge. However, if knowledge sharing barriers are not identified and addressed then they could hinder the commercial competitive advantage. A research question could be framed as 'could knowledge transfer be increased to aid organisations, but at the same time not exponentially increase the amount of irrelevant information that might cause information overload'. The remaining sections in this chapter look at the role of technology in increasing knowledge transfer through information and the impact it might have on information overload.

## 2.5 Retrieving and Navigating Information

The following three main sections focus upon three key areas that could be of use to improve the discovery and navigation of relevant **explicit information and knowledge** from a number of information sources identified in Figure 5. These are namely:

- **Search engines and user performance**
  - Types of search
  - Mainstream search improvements
  - Visualisation to Improve Search Performance
  - Clustering
  - Tag Clouds
- **Tagging**
  - Collaborative Tagging
  - Issues with Tagging
- **Semantic Technology and Ontologies**



**Figure 3 – The interface to information sources**

The literature review has identified the concept of pulled information and that one of the key causes of information overload comes from the inability to find relevant information within large volumes of information. The Internet is one of the primary sources that has a vast amount of information available to an employee, but the Internet is not the sole contributor of information within an organisation. Documents in file stores and those found on corporate intranets also add to the vast sea of information. Although the literature review found that technology was not the only factor at fault, it did show that technology provides the channels and mechanisms through which information is distributed or accessed (Wilson 2001). The next sections identify how various methods may be of benefit to information retrieval, either within corporate document stores or on the Internet and intranets. The risks and barriers associated with these various methods are reviewed.

## 2.6 Search Engines and User Performance

Many of the traditional sources of information require a pull approach which usually involves the use of a search interface. The results are then normally presented to the user via a graphical user interface. This section addresses the information retrieval challenge faced by users trying to find information relevant to their daily tasks. It reviews the current visualisation techniques and discusses

the topic of clustering, looking at the benefits and disadvantages of the various methods deployed to help navigate through the information minefield.

### **2.6.1 Types of search**

There are several different types of search and not all are of interest when looking at improving mainstream searches. Three different types of searches are identified in this section: closed searching, open searching, and transactional queries.

The first search, closed searching, involves a user searching for a single document related to a field. For example, a user wishes to find information about a single article. In the context of one of the case study organisations in this thesis, PharmaCo, this may be a document from a specific medical study that occurred in the past. These queries are also known as navigational queries “the immediate intent is to reach a particular site” (Broder 2002)

The second type of search, open searching, occurs when an end-user wants to find information that may exist in one or more places. For example, to find all medical studies that have occurred relating to a certain chemical compound or past examples of a drug being submitted for review by an Approvals Board. This type of search is also known as informational as “the intent is to acquire some information assumed to be present on one of more web pages” (Broder 2002).

The third option and which is less frequently found in literature is called transactional queries (Broder 2002). “The purpose of such queries is to reach a site where further interaction will happen” (Broder 2002). In a transactional query, the search engine simply provides another location for the search to continue. The discovery of such sites and the results of transactional queries are extremely difficult to evaluate (Broder 2002). Both navigational (closed search) and informational queries (open search) will be investigated further in the following sub-sections as they are the most common type of searches and potentially could have the most impact upon employees.

### **2.6.2 Mainstream search improvement**

In our modern information rich era search engines are heavily relied upon to help users discover the content they are looking for. In 2005, Gulli and Signorini

estimated that there were over 11.5 billion indexable pages on the Internet (Gulli, Signorini 2005). This estimation was made by performing searches with a number of different search engines and comparing similar links to attempt to discover the overlap between them. It involved taking pages at random and then querying certain terms to see how many pages were returned to provide the estimation. However, each stage provides a considerable amount of opportunity for inaccuracy to be introduced. Although the authors provide a number of statistical tables and probabilistic equations, they never elude to their work being anything more than estimation.

More recent estimations, using similar methods, place this figure close to 60 billion pages (De Kunder 2008). However, the accuracy of these results is again somewhat questionable. Even if this work is inaccurate by a factor of 10 or even 1000 the Internet still represents a formidable corpus of documents that is regularly indexed and searched by millions of users each day using many of the popular search engines. Retrieving the correct results and presenting them to the user is no simple task and even with all of the work that has gone into search algorithms, many users are still left unable to locate the documents that they need.

There has been a considerable amount of effort invested in the back-end processes of search engines (Wiza, Walczak & Cellary 2004). This investment has increased the search engines accuracy and the ability for users to find relevant information. Many of the techniques and technologies developed have been applied to the major search engines and corporate search systems alike (Brin, Page 1998) (Joachims 2002) (Zhang, Dong 2004). In addition to this, a large area of research relating to the behaviour of search engine users also exists (Jansen, Spink 2006) (Teevan et al. 2004) (Madden et al. 2006).

Although the core mainstream search engines such as Google, Yahoo and Microsoft Windows Live Search have added improvements to the core search ability, such as the ability to search within a site, or to find documents that are related to those shown, the representation of search results has barely changed since their conception years ago (Wiza, Walczak & Cellary 2004). To date, the investment has been in collecting and indexing data, algorithms for query processing, as well as



data caching mechanisms. According to Wiza, Walczak & Cellary (2004) the element that has remained almost untouched since the very beginning of the inception of search engines is the presentation interface of the search results..

In the majority of search engines, the user is presented with the title of the web page followed by a quick summary of the web page's text. Traditionally this would be the first few lines of the document although more recently query based summaries have become popular. Query based summaries show text that contains or is related to the terms found within the document (Paek, Dumais & Logan 2004). The query relevant text that exists in modern search engines has provided significant steps towards helping the user determine how relevant a concept is and the query relevant text has allowed users to see parts of the page that might enable them to distinguish the relevance of the page to their query without having to open the page itself. However, this approach still requires the user to read the short summary and manually determine the benefit of each returned web site. Although better than the traditional system, it still results in the user opening a large number of documents in an attempt to find the information that they feel is relevant to them. "The notoriously low precision of Web search engines coupled with the ranked list presentation make it hard for users to find the information they looking for" (Zamir, Etzioni 1999). Although this quote comes from 1999, only a few years after search engines such as Google became mainstream, the principles still exist today. Although the precision of search engines may have changed and people's behaviour towards searching may have changed, the number of irrelevant documents creating 'noise' has increased and the presentation of the results remains the same.

The lack of change in the information retrieval domain is especially interesting given the economic value associated with the major search companies. In February 2008, Microsoft offered 44.6 billion US dollars in an attempt to acquire Yahoo!(Microsoft). Having the advantage in the search engine industry is apparently worth a substantial amount of money and considerable effort is placed into the development of retrieval. Whilst it is clear that the list-based approach to displaying results has limitations, each of the major search engines still adopts the approach. This still causes an issue as even if the correct result is within a list of

search results, it can still be difficult to find the correct document within the list of results.

Given that there is a large body of research relating to the improvement of search algorithms (Joachims 2002), (Zhang, Dong 2004) search behaviour and the processes involved in searching (Jansen, Spink 2006), an area for further investigation is the presentation of search results. Improvements in this area could improve the effectiveness and efficiency of the end-user. There have been several approaches at solving the issue of finding the appropriate document and presenting documents and search results in different formats, each with varying success, namely visualisation and clustering. These approaches are detailed in the following sections and will help to determine if there is a better way of presenting search results.

### **2.6.3 Visualisation to Improve Search Performance**

There have been a number of research studies that suggest visualisation can help the generic search process (Wiza, Walczak & Cellary 2004) (Sebrechts et al. 1999) (Xie, Poshyvanyk & Marcus 2006) and in this section they shall be reviewed and the benefits and disadvantages outlined. The aim is to determine the aspects of the visualisation process that could aid users in the quest for relevant documents. This will be achieved by reviewing techniques that will aid in performing navigational and informational searches as discussed in section 2.6.1.

A study conducted by Paek, Dumais & Logan (2004) looked at several options for improving the presentation of information from a search query to building upon the existing display method. Their system added additional query relevant text to the search results from the target document at the request of the user. The query relevant text was text related to the search terms that attempted to show the user why the document had been returned. A system was produced that revealed the query relevant text beneath the search result in one of two ways. The first method instantly revealed the text underneath the content. When this result was clicked on, the user could see the expanded text. The second method applied a 'fish eye lens' to the text so when a user moved their mouse over the text, the section of text

closest to the mouse would be magnified and text that was further from the mouse would reduce in size.

The system was motivated by the need to provide more information to the user whilst minimising the screen space that is taken up by each search result. Results showed that the method for instantly displaying the relevant text actually resulted in a quicker mean time for users to answer questions related to the web pages that they were searching for. This system has apparent benefits and enables users to control the content that they see whilst searching. There are some limitations to such a design including the system still requiring the readers to look through the large amounts of text related to the query. This issue is actually more exaggerated than in traditional systems as more information can be displayed on a single page of results using this method. Results also have to be manually scanned and carefully looked through to find the most relevant. Hiding some of the information makes the presentation of additional results possible and perhaps allows users to scan more quickly. The system will allow users to avoid obviously unrelated topics and can also be used in conjunction with other systems providing an interesting alternative to some systems. However, this method may even lead to a greater level of 'information overload' as more data is available.

A general feasibility study using 2D and 3D visualisations for queries was conducted in 1999 by Sebrechts et al. (1999). The study, "Visualisation of search results: A comparative evaluation of text, 2D and 3D interfaces", looked at using a text based approach, a 2D approach and a 3D approach to display search results. They measured the time it took a user to locate a target over six different sessions. In the study, the users were presented with a number of different task types. In all task types, the 2D representation outperformed the 3D representation and in almost all tasks when using the text based system. However, the scenarios where the 2D system performed better than or equivalent to the text based system were when users were asked to "Recover a document and locate a new document given its title or content" and "Recover and compare contents of documents" (Sebrechts et al. 1999). This shows that there may be benefits to using visualisation, especially two-dimensional visualisation, for these types of search tasks. The 2D representation in the study appears to simply be a squashed version of the 3D

representation, taking the 3D texture and wrapping it onto a 2D plane. The three-dimensional visualisation may not have provided an adequately designed visualisation for its purpose and this in turn may have led to it being insufficient rather than leading to the conclusion that any three-dimensional visualisation will *prove ineffective*.

Overall, the text based system gave the fastest response times. Both 3D and 2D systems showed an increase in response rate over time, as the users became more used to using the novel system. This demonstrated that there was time spent learning and adapting to the new approach of search representation, as could be expected for any new system that is introduced. The results also showed that 3D and 2D systems had a large adaptation time which may be explained by the fact the study was conducted in 1999, when the Internet was not as graphically rich as it is today.

In 2004, Wiza, Walczak & Cellary (2004) suggested an interesting approach to tackle the displaying of multiple documents relating to a certain topic. Their approach used a 3D interface to allow search results to be displayed in a holistic approach and to allow the user to visualise not only the concepts within documents but also their relationships. The approach is not too dissimilar to that of Mukherjea and Hara in 1999 (Mukherjea, Hara 1999), however significantly more visualisation approaches are explored. They concluded that although the system response time was higher than that of traditional result displays, “the informational completeness of the results and understandable form of presentation proved to be worth the short delay” (Wiza, Walczak & Cellary 2004). What is interesting here is that not only did this interface appear to work but also the more subtle fact that users were prepared to wait a short amount of time for additional content to load as long as it provided sufficient benefit.

The two key issues with this method are that firstly it really only works for the retrieval of multiple documents and for showing the relationships between them. The authors stated that for finding one document the system would probably not be of great benefit. Secondly the system did not present results using standard html. This resulted in an extended loading time of the system. Having an increased

loading time could increase the risk associated with the system. If a user has to wait too long then they will begin to see the system in a different light claiming that it is *slow, or even badly programmed*. Users may even discard the system entirely in preference for something that provides quicker results. There is also the risk that plug-ins will not be loaded on some user's computers and whilst many users are unable to add new plug-ins to their browsers due to security restrictions, many will choose not to load them anyway because of the time constraints involved before they can even use a system.

The impact of representing search results as text, 2D and 3D on the end-user could be affected by which side of the brain they use. For several decades it has been understood that people may have a *tendency towards left or right brain* dominance. Nobel Prize winner Sperry, along with his student, Gazzaniga, worked for many years around the subject of the brain and its differing hemispheres (Gazzaniga, Sperry 1967) (Sperry 1984). Sperry found 'that the left half of the brain tends to function by processing information in an analytical, rational, logical, sequential way. The right half of the brain tends to function by recognizing relationships, integrating and synthesizing information, and arriving at intuitive insights.' (Dew 1996).

Although this research has been widely used, Hermann (1996) further extended this knowledge to develop a profile instrument to help assess those who are right-brain and those who are left-brain dominant. The research and the development of such a tool leads to the suggestion that some users are therefore better at *visualising results (the right brain thinkers)* and some are better with a more logical keyword oriented approach (*the left brain thinkers*). It may therefore be necessary to consider that some people may be more suited to the more visual approaches.

This information demonstrates the differences between users of search systems. In the early and embryonic stages of the development of the Internet and many of the search systems, the end user would have frequently been technically minded as they would have been working in the field of the Internet or IT in general. Over the years this has changed and nowadays the *end-user of a search system can come*

from any domain. For this reason, it is important that different types of user are considered and that visualisation does become part of the search process to aid the more 'right brained' users.

In summary, a number of existing visualisation techniques have been discussed. There have been a number of attempts to use both two and three-dimensional techniques with some success. Many of the systems discussed require browser plug-ins that could cause performance issues. It was shown that the system developed by Wiza, Walczak & Cellary (2004) required a longer processing time that resulted in a delay providing results to the end users. Users did not mind the load time as long as the result was worth the delay. Many of the approaches in the search systems were new to the end-user and required a period of adjustment for the user. In an attempt to speed up the adjustment time, training could be considered for any new system developed and the ease of adjustment to any system should also be considered. The literature has highlighted that there is a need for visual approaches, but some users will be suited to one method of presentation (visual) whilst others may suit a completely different method (textual). The next section takes visualisation a step further by reviewing the techniques associated with clustering search results to improve end-user efficiency and effectiveness.

#### **2.6.4 Clustering**

An area of investigation relating to the presentation of search results is clustering. Clustering documents is derived from the idea of bringing related documents or concepts together based upon the content of the documents. The concept of clustering has appeared within a number of the studies related to visualisation (Wiza, Walczak & Cellary 2004)(Sebrechts et al. 1999).

In 1999, in the qualitative analysis section of their paper, Sebrechts et al. (1999) noted that their systems clustering and grouping of concepts was liked and also that "participants used the grouping of concepts into clusters to narrow their search for particular documents, if a particular concept was not of interest, the participant knew which set of documents to avoid". They also mentioned that

clustering allowed users to see combinations of concepts that could be re-used in their search.

There have been numerous research studies on search clustering and improving clustering algorithms (Zamir, Etzioni 1999) (Zhang, Dong 2004) (Wen, Nie & Zhang 2001) (Leouski, Croft 1996). Whilst literature regarding visualisation improvements to search engines is relatively scarce, the concept of clustering has been the focus for many researchers for some time and its application to visualisation is still being actively researched today (Tvarozek, Bielikova 2008).

Clustering has been shown to provide benefits to users when searching for documents enabling faster and more efficient searches. Zeng et al (2004), when looking at a more efficient way of clustering results, noted that “organizing web search results into clusters facilitates users’ quick browsing through search results”. This improvement of quick browsing was also documented by Hearts and Pederson (1996). Hearts and Pederson showed that documents that are similar to each other often tend to be more relevant to each other. This also lends itself to the inference that clustering similar results together will help users find a number of relevant results once they find a relevant document or cluster.

There are a number of clustering systems available for document retrieval ranging from simple and manually created clusters to automated systems making use of extremely complicated algorithms, which are still being actively developed and improved today. One example of a manually constructed cluster is the open directory, project also known as the DMOZ (Open Directory Project), it was one of the first systems of this type and acts as a directory for sites on the Internet. Before the DMOZ, Yahoo, now known as a famous search engine and portal began its life as a directory (Jacso 2007)(Northedge 2007). The directory can still be seen today at <http://dir.yahoo.com>. Sites are placed into categories that can be browsed and searched by users in order to find sites within a particular domain. This manual system is extremely time consuming process of cluster creation, as each new item has to be added by an editor. The use of clustering highlights two areas for further exploration within the main body of the thesis. Firstly, the extent that users are capable of both categorising documents and effectively retrieving them, based

upon the content of the document and the key concepts described within the document. Secondly, the impact of clustering as an alternative presentation method.

Automated approaches for clustering research documents also exist. Kobayashi et al. (2006) described a method of presenting clustered concepts and categorising search results together using these concepts. This has become quite common when clustering is used on the Internet and there are many search engines now available that make use of clustering technologies. One of the best examples of a clustering search engine, freely available to use on the Internet, is the Vivisimo (2006) search engine. Figure 4 shows the system with the term ECG entered. ECG stands for Electro-Cardio-Gram or Electro-Cardio-Graph and is a system commonly used for measuring electric activity of the heart.

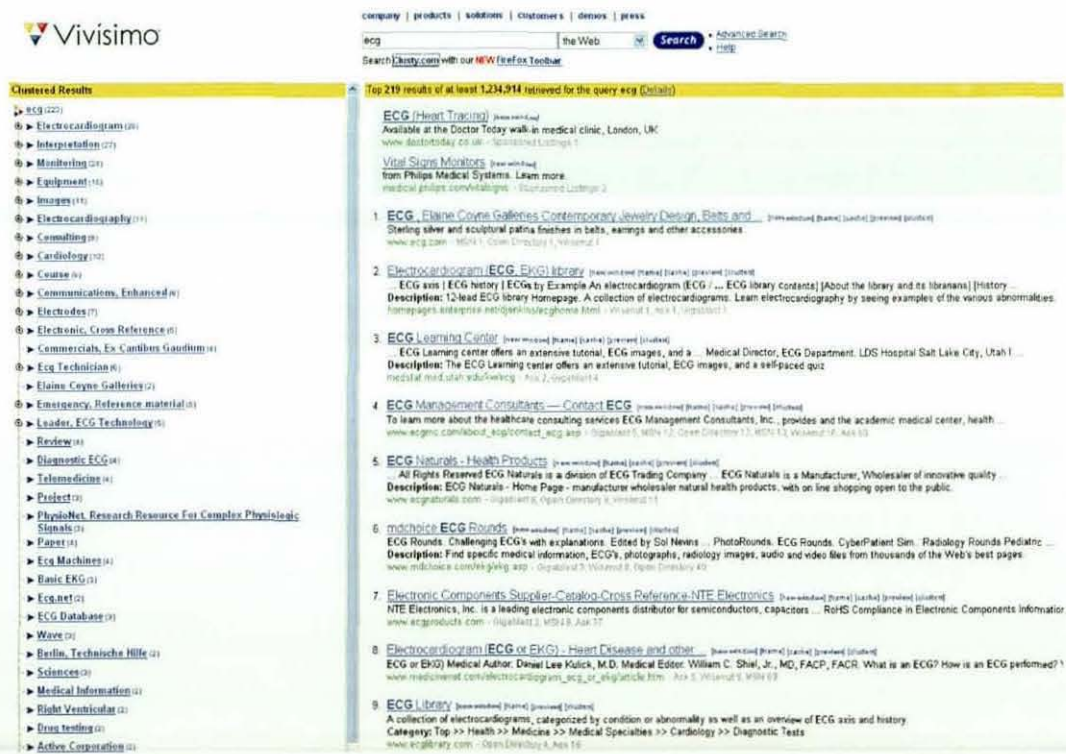


Figure 4 Vivisimo Search Engine – Search Term ECG

Figure 5 shows a zoomed view of some of the search clusters that have been made available in Figure 4 including: electrocardiogram, electrocardiography, cardiology, drug testing, but also a number of topics completely unrelated to the heart that happen to have relevance to the term or letters ECG.





**Figure 5 Clustered Terms in Vivisimo**

One of the benefits of automated clustering is that it does not require pre-defined categories for items as classification methods do (Zeng et al. 2004). This means that as new concepts arise, the search system adapts to display these new concepts without user interaction or an administrator having to create categories and concepts that may relate to this (Zeng et al. 2004).

In early clustering systems, an entire document had to be downloaded and the full text of that document had to be analysed in order to correctly place the document (Zeng et al. 2004). Using the full document can be extremely resource intensive, as categories are created at search time to combat the issues. Authors such as Zeng et al (Zeng et al. 2004) recommend using snippets of the document instead to create a more efficient way of cluster generation. This concept may be of use when generating a visualisation system as visualisations can be extremely resource intensive if not designed with performance attributes in mind (Sebrechts et al. 1999).

Clustering appears to be of benefit and many of the concepts associated with clustering may have implications for visualisation systems (Zeng et al. 2004). The idea that documents can be categorised by their content and that documents can be related to each other could be used to enhance the search process. Taking

important concepts has also been shown to be of use when narrowing down (Zeng et al. 2004) a search and may have interesting consequences when applied to visualisation. Although the clustering helps to show where a document belongs, it does not appear to help to describe all of the concepts that may exist within a document (Zeng et al. 2004). Another potential drawback is that clustering requires significant processing power and therefore it could be an expensive option when it comes to acceptable performance.

In recent years, there has been further development in the area of clustering (Hassan-Montero, Herrero-Solana 2006) to try and overcome these issues. Tag clouds have emerged enabling users to view key concepts that are being described by a web site. This area of visualisation is discussed in detail in the next section.

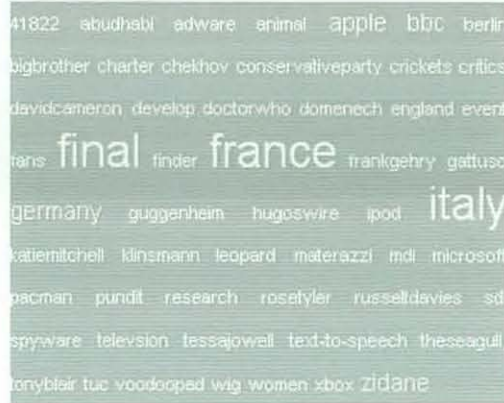
#### 2.6.5 Tag clouds

The concept of tagging requires users to add tags, freely chosen keywords, to web resources in an attempt to categorise these resources and identify what the content represents. These concepts are then weighted across a collection of these resources in order to see which occur most frequently. They are then represented as an alphabetically ordered list of the most popular tags.

Tag clouds have been used by sites such as Flickr (2006) and the Guardian newspaper (Guardian Newspapers Limited 2006) and they have also been extremely popular within 'web logs' or 'blogs' to identify what a 'blogger' has been talking about or the different sections of the 'blog'. Tag clouds have become an integral part of web-based systems within the concept of Web 2.0. They provide a lightweight and quite informative overview of the content of a site.

In many systems they show the relative popularity of certain concepts across the entire site. Figure 6 is an example of a tag cloud from the Guardian newspaper. It shows the Guardian newspapers 'folksonomic zeitgeist' taken shortly after the football World Cup final between Italy and France in 2006.

### The Folksonomic Zeitgeist <sup>2</sup>



**Figure 6 The Guardian Newspaper's Folksonomic Zeitgeist**

Several academic papers have also recently emerged providing suggestions on how to help the tagging and visualisation process (Hassan-Montero, Herrero-Solana 2006)(Begelman, Keller & Smadja 2006). Hassan-Montero (Hassan-Montero, Herrero-Solana 2006) offers suggestions that this 'square' alphabetically based layout seen in Figure 6 could be improved, based upon the works of a number of other authors who present it as a 'cluster based layout'. The study looked at both the content of the tag clouds and at the layout. The cluster based layout attempted to reduce issues caused by synonyms and related concepts and borrowed concepts from those of clustering. The idea behind the study was that a more coherent visual distribution of tags than traditional alphabetical arrangements would enable users to differentiate amongst main topics in Tag-Cloud, as well as infer semantic knowledge from the neighbours' relationships (Hassan-Montero, Herrero-Solana 2006). However, the study does not provide evidence to suggest that this type of visualisation aids the user in anyway. Figure 7 shows Hassan-Montero's improvements to tag clouds.



**Figure 7 - The tag cloud after Hassan-Montero's improvements**

Research undertaken at similar time to Hassan-Montero by Begelman, Keller & Smadja (2006) suggests that by relating tags to their hyponym and hyponym terms, more specific or general words describing these tags could be found. Again the research did not provide sufficient evidence to back the claim.

Whilst tag clouds present a very interesting and resource friendly option for adding visualisation to a web page, they also present potential benefits when summarising the content of a site. The drawback is that they do require manual tagging for the entire site's content. They also only focus on the contents of a selection of resources rather than the individual resources preventing tag clouds from being a viable option when presenting a visualisation for a single document rather than a collection of documents.

### 2.6.6 Search Summary

Search engines are relied upon by users to help them search and discover the relevant information amongst a mass of irrelevant information contained on both the Internet and corporate intranets. As such they represent a key area that can be focused upon to help discover relevant information and reduce the problem of information overload. The literature has shown that although substantial literature exists within the field of search engines, much of this literature has focused on the improvements to the algorithms that power the search systems themselves and the behaviour of how users search. Relatively little research has been focused on the presentation of search results. The little research that was found has focused on adding visualisation to search results, mainly in the field of web searching,

rather than intranet based search systems. Although it could be argued that the same visualisation techniques could be applied to both, many of these attempts have said to have been successful in providing visual additions to search results, but the extent to the success has not been provided within the literature.

The literature has also provided some disadvantages to visualisation, such as requiring a significant amount of processing time or even that the system is only of benefit when comparing documents rather than performing a navigational search. These disadvantages in many cases may actually have created barriers preventing the visualisations from being as effective as they could.

One of the most frequent researched areas for visualisation improvement has been around clustering. The work relating to clustering has shown that alternative visualisation approaches might work and that it is possible to identify documents by key words relating to their content. The main advantage of clustering is that it does not require a significant computational overhead. Table 4 summarises three of the key research papers on visualisation. It provides an overview of the advantages and disadvantages of the various approaches and helps to summarise this section.

**Table 4 - A Summary of Visualisations Discussed**

| Title  | Author(s)               | Year | Advantages  | Disadvantages   |
|--|-------------------------|------|---|---|
| WaveLens: A new view onto Internet search results  | Paek, T. et al.         | 2004 | 'Fish eye' view allowed easy expansion of content and user to narrow down to see more information. Page space was better utilised with more detail when required. Could lead to increased scanning speed. | Reader still has to sift through large volumes of information, even more than previously except it is now all on one page. Could lead to increased information overload, the problem that ultimately the system is attempting to address. The system requires at the least JavaScript to be enabled in the browser. |
| Visualization of search results: a comparative evaluation of text, 2D, and 3D interfaces | Seberechts, M.M. et al. | 1999 | Comparison of 2D, 3D and text-based displays. Users presented with a number of different task types. Shows there may be benefits with Visualisation especially Two-Dimensional.                           | Systems required browser plug-ins or enhancements. Two Dimensional system was really only three-dimensional image wrapped around two-dimensional plane. Study performed in 1999 before Internet was as graphically rich as it is today.   |
| Periscope: a system for adaptive 3D visualization of search results                      | Wiza, W. et al.         | 2004 | Holistic Information Display. Visualise relationships between concepts.   | Delay of up to 15 seconds loading. Really only of benefit when looking at multiple documents not searching for one in particular. Required browser plug-ins or enhancements.  |

The literature review of searching and visualisation has shown that further investigation within the field of visualisation could aid information retrieval. The current research falls short in providing the quantifiable benefits of using such techniques, but does allude to their potential benefits. It also provides limited information about the potential barriers to using visualisation techniques, for example, significant processing power. Further research is required to understand

both the barriers to using visualisation techniques and quantifiable benefits, in order to aid the user in retrieving information.

## 2.7 Tagging

Tagging refers to adding freely chosen keywords to a resource, with a view to making the resource retrievable using those keywords. The resources may be anything from photographs to URLs to music or even documents. These freely chosen keywords are often restricted to single words, although some systems allow multiple words and others simply ask users to combine words without spaces. In this chapter a form of tagging has already been reviewed - tag clouds (section 2.6.5). In this section a much broader and deeper review of tagging will be provided to gauge the impact it might have on information retrieval.

The concept of tagging resources has grown dramatically in recent years. Sites such as Flickr and Del.icio.us have brought the concept of tagging information to the masses and have attracted so much attention that companies such as Yahoo have deemed them financially viable for acquisition (Hu 2005). Del.icio.us, for example, received around 150,000 posts per day in June 2008 (Keller 2008) and, according to Rainie (2007), 28% of Online Americans have used the Internet to tag content.

Tagging has been seen as one of the key aspects of the phenomenon known as Web 2.0. Web 2.0 is the term given to a wide range of aspects as the Internet moves towards a more graphical and interactive medium and a more collaborative and useful environment (O'Reilly 2005) (Giustini 2006). The exact turning point from Web 1.0 to Web 2.0 is unclear however it is O'Reilly media who are attributed with coining the term in September 2005 (O'Reilly 2005). The exact meaning of Web 2.0 is also unclear; one thing that is known is that it was the rejuvenation of the Internet after the "dot-com crash" of the early 2000s (O'Reilly 2005). O'Reilly attempted to clarify what Web 2.0 actually is in comparison to Web 1.0. Tagging was one of the key elements of Web 2.0, forming a strong contrast to the structured storage in pre-defined folders in Web 1.0.

Tagging systems form a strong contrast to many traditional systems of classifying resources that “relied on well-defined and pre-declared schemas ranging from simple controlled vocabularies to taxonomies to thesauri to full-blown ontologies” (Hammond et al. 2005). Taxonomic and ontological systems are often very expensive in terms of creation and maintenance in comparison to lighter weight systems such as tagging (Christiaens 2006).

Some authors argue that the barriers to the entry of using a tagging system are far lower than those of taxonomies (Mathes 2004) (Schmitz 2006) (Gruber 2007), with Mathes (2004) arguing that getting the everyday user to use a complex hierarchical controller vocabulary would be too difficult. Not relying on a hierarchical or taxonomic structure often has advantages when retrieving resources. Golder and Huberman (2006) offer the following example to affirm this point.

“Consider a hypothetical researcher who downloads an article about cat species native to Africa. If the researcher wanted to organize all her downloaded articles in a hierarchy of folders, there are several hypothetical options, of which we consider four:

1. articles\cats - all articles on cats
2. articles\afrika - all articles on Africa
3. articles\afrika\cats - all articles on African cats
4. articles\cats\afrika - all articles on cats from Africa

Each choice reflects a decision about the relative importance of each characteristic. Folder names and levels are in themselves informative, in that, like tags, they describe the information held within them.

Folders like one and two make central the fact that the folders are about ‘cats’ and ‘Africa’ respectively, but exclude all information about the other category. Three and four organise the files by both categories, but establish the first as primary or more salient, and the second as secondary or more specific. However, looking in three for a file in four will be fruitless and so checking multiple locations becomes necessary.” (Golder, Huberman 2006).



Tagging is not a new concept, as recognised by Hayman and Lothian (2007), as librarians, indexers and professionals have been tagging content for years. The difference with modern systems is that users are empowered to create their own tags rather than relying on those created by experts. Tagging has also been referred to as “subject indexing without a controlled vocabulary” (Hayman, Lothian 2007).

The word folksonomy often synonymous with tagging, is often cited as being invented by Vander Wal (Smith 2004), (Christiaens 2006), (Peters, Stock 2007) and is used to refer to a collection of tags. The term takes its name from the combination of the words folk and taxonomy, literally meaning a taxonomy created by people. This term is actually quite misleading as tags are not taxonomic in structure at all; their purpose is to lack any structure whatsoever. The replacement of the taxonomy with these un-structured tags is symbolised quite well by this name. One of the key benefits identified by the literature of the folksonomy, over a system that makes use of a controlled vocabulary, is that the tags automatically reflect the vocabulary of the users (Mathes 2004). While this is a subtle point, it does make a large difference when it comes to document retrieval. The vocabulary is free to adjust to the latest developments, buzzwords and colloquialisms for expressing terms.

Tagging also has the benefit of instant feedback, which is rarely mentioned in the literature but a very desirable feature for a search system (Mathes 2004). Articles that are related to each other can be found using the same tags and thus when a user tags an item, you can see all of the documents that have received the same tags and are therefore likely to be related (Mathes 2004). This comes with the very nature of the tagging system and requires no extra effort from the user. This can also lead to a tight feedback loop with users seeing tags that others have used to describe an article and thus leading to a more focused choice of tags (Mathes 2004).

### 2.7.1 Collaborative Tagging

Hammond et al. (2005) highlighted that tagging related sites may be classified by both their audience and tag author. Many existing sites that use tagging and their reasons for doing so were presented in the following table:

|                 |               |  |   |
|-----------------|---------------|--|---|
| <b>Tag User</b> | <b>Others</b> | <i>Technorati</i><br><i>HTML Meta Tags</i> | <i>(Wikipedia)</i>  |
|                 | <b>Self</b>   | Flickr                                     | CiteULike<br>Connotea<br>del.icio.us<br>Frassle<br>Furl<br>Simpy<br>Spurl<br>unalog |
|                 |               | <b>Self</b>                                | <b>Others</b>   |
|                 |               | <b>Content Creator</b>                     |   |

Figure 8 - Reasons for Tagging Content (Hammond et al. 2005)

Tags are described as either being created by the user or by another user and the content being created by both the user and others. There are a number of entries in the bottom right corner of Figure 8. This concept of tagging a document for use by others has recently become a strong feature of tagging, often referred to as collaborative tagging (Hammond et al. 2005)(Golder, Huberman 2006)(Macgregor, McCulloch 2006) (Halpin, Robu & Shepherd 2007).

Collaborative tagging has grown “so that users can not only categorize information for themselves, they can also browse the information categorized by others. There are therefore at once both personal and public aspects to collaborative tagging systems” (Golder, Huberman 2006).

Golder and Huberman (2006) looked into the dynamics of the popular social bookmarking tool Del.icio.us and based the research on two datasets. The first contained a list of locations that were marked as popular within a certain

timeframe. This list was used to see which users had bookmarked particular sites. The second dataset took a sample of users that posted links within a certain time and then obtained all of the sites that they had ever bookmarked. They discovered a number of different classifications of tags:

- Identifying what a bookmark is about – here the user lists topics to identify the content of the bookmark.
- Identifying what a bookmark is – for example a blog post, an article or a book.
- Identifying who owns the bookmark
- Refining categories – interestingly they discovered that some tags do not stand in their own right but rather restrict the set of bookmarks and allow a user to narrow down their search.
- Characterise a bookmark – keywords such as funny, scary and inspirational are referenced to highlight the taggers opinion of the article.
- Self reference – Tags can help the tagger identify with the content or place it in a category of its own.
- Task organisation – Some people added tags in relation to a task they may be performing. Examples here included ‘toread’ or ‘jobsearch’.

An interesting conclusion can be drawn here. Although it does not appear to be mentioned within the literature, some users tagging for general use also include tags that are references for themselves. The tags used in categories such as Self Reference and Task Organisation do not differ in any way from those for public use and will therefore be mixed into the public ones creating noise. These additional tags will continue to mix tags that are irrelevant to other users with relevant ones making information harder to find. Additional drawbacks to tagging are discussed in the next section.

### 2.7.2 Issues with Tagging

A key issue with tagging is that users freely choose the tags and as already mentioned, tagging can be seen as “subject indexing without a controlled vocabulary” (Hayman, Lothian 2007) and this is exactly the issue, there is no

controlled vocabulary. With users free to choose their own tags, even the same user may choose different tags on different occasions.

When items are tagged by multiple parties, during collaborative tagging there is a higher likelihood that different tags may be chosen (Golder, Huberman 2006). Since there is no restriction in the words that can be chosen, this can often lead to users choosing different words to describe the same resource. When a different user attempts to find a resource, they may not be able to because of the difference in tags given by the tag author and the user performing retrieval.

Research has shown (Golder, Huberman 2006) that as more authors tag content, the tags chosen by authors do converge to give a consistent number of appearances relative to the number of authors tagging content. This does show that there is a consistency in the way that an article is tagged. However it is still evident that many authors use tags that are not used by others. In many cases the tags given by a user can be too general or too specific compared to those given by another. The words used may also be synonyms of one another, for example, whilst one person describes a resource with the word person, another may use the word human (Golder, Huberman 2006).

The concept of collaborative tagging thus far has only referred to a collaborative tagging environment with multiple users tagging content. This is often referred to as creating a broad folksonomy. In many situations, solely the author of that document tags a document, but the tags are seen and consumed by many users. This is often referred to as a narrow folksonomy (Christiaens 2006). As many people tag content differently, in this case the problem is only going to cause greater issues if only one person, the author, tags the content. There are also further issues with tagging. Simple distinctions can also cause issues with tagging. The unstructured nature of tagging can lead to issues when items are tagged with either singular or plural versions of a tag (Mathes 2004). In some cases a user may use the tag 'apple' and in others 'apples'. It is clear that whilst 'apple' could describe both the fruit and the computer manufacturer the word 'apples' would not describe the company. An even simpler issue that may arise is that of spelling; if a user misspells a word then it will appear as a different tag.

Stemming is rarely mentioned in conjunction with tagging but it shall be mentioned here in order to dismiss the idea. As described by Sood et al. (Sood et al. 2007) stemming can lead to an over-normalisation of tags with tags sharing the same root such as "While 'production', 'product', and 'producers' share the same morphological root 'product', they each have distinct meanings." (Sood et al. 2007). Interestingly, even this example suffers from the issue of pluralisation and highlights the issue again. Sometimes a document could describe a single producer and sometimes it may describe a number of producers. It is a difficult question to decide whether the singular or plural form should be enforced or if users should be free to choose.

Further to the issues mentioned already, synonyms also cause issues like "mac," "macintosh," and "apple" all being used to describe materials related to Apple Macintosh computers (Mathes 2004). Whilst some systems such as Del.icio.us show a "related tags" feature, in many cases this is not possible. The related tags are created using a database of tags that are used to describe the same URL on a regular basis. This in itself requires additional information to be able to recommend the tags and this is not always available, especially in non-collaborative, narrow folksonomies. These issues highlight the reason why controlled vocabularies have been used and enforced in traditional settings (Mathes 2004). One of the main reasons why hierarchical systems are chosen over tag based systems is that they provide a much higher degree of meaning (Christiaens 2006). Information annotated by means of an ontology provides a far more advanced method of querying. Christiaens (Christiaens 2006) argues that people will not learn to use a query language, as much of the Internet is currently based on keyword searches, but there is no evidence to support this. However, there is some literature (Barbosa 2008) to suggest a hybrid approach may be taken, using as much of the taxonomy or as much as the folksonomy based technology as appropriate for the organisation in question.

### **2.7.3 Summary**

This section has investigated the concept of tagging in order to improve a user's ability to retrieval accurate information. The literature has shown that tagging can increase a user's ability to discover relevant documents that have been tagged

previously. It has also highlighted that tagging can be of benefit to discover relevant information that might not lend itself to full text search used by search engines. Documents such as images and movies can be found with tagging based systems. One of the key difficulties in finding relevant information comes from the idea that information does not have the same value to each user (Nelson 1994). The concept of collaborative tagging has been reviewed and it has been shown that over time, a certain degree of consensus shall be reached with the tags chosen to represent content. The literature also points to tagging as a lightweight alternative to heavier weight alternatives, such as ontology and taxonomies. It has been stated that the barriers to the entry of using a tagging system are far lower than those of taxonomies (Mathes 2004), however a number of barriers to tagging were identified. The effect of these barriers however was not assessed in detail and the extent to which these barriers are present does not appear to have been determined in the literature.

The literature has highlighted a number of benefits to tagging and that tagging may provide a way to reduce the information overload problem by enabling users to find more relevant information. This is especially the case for files such as images and video that might not appear through full text search. However, further research is required to quantify the benefits of using such tagging systems and the potential barriers that need to be overcome to ensure such systems are successful within organisations.

## 2.8 Semantic Technologies and Ontology

The word semantic has roots in the theory of meaning. To add semantics to something is said to add meaning to something and semantics is the study of meaning (Bozsak et al. 2002). Traditionally this has been associated with language but the term has also occurred within technological advancement and the desire for a computer to understand more about content rather than just being available for human interpretation (Bozsak et al. 2002) (Berners-Lee, Fischetti 1999).

Semantic technologies are often confused with the semantic web. Although semantic technology will be needed to realise the idealised semantic web, they

may also exist in their own right and form the basis for many Web 2.0 applications and even independent applications (Noy, McGuinness 2001).

### **2.8.1 The Semantic Web**

The concept of the semantic web is only a slight modification of the existing World Wide Web that we know today. The increase in the volume of information available and has already been discussed along with its contribution to information overload in section 2.2.

The literature review so far has shown the difficulty in retrieving relevant information and the systems that exist to aid the discovery of relevant information such as those in sections 2.6 and 2.7. The literature review has also shown, in section 2.2.3, "The cause and growth of information overload" how technology has been blamed as one of the causes for the increase in the information overload problem (Farhoomand, Drury 2002).

Although it is the volume of information that causes a problem, the variety of different mediums of information available on the Internet may also be to blame (Kirsh 2000) (Farhoomand, Drury 2002). In recent years, the Internet has progressed from its initial text only, 'hypertext', stages to a media rich frenzy of information. Suddenly the Internet has exploded with new technologies and ideas. This move from standard html web pages to the second generation of technologies was coined 'Web 2.0'.

In 2002, McGuinness suggested that there were two key reasons that users were not satisfied with their ability to find relevant information through search engines (McGuinness 2002). The first is to do with the presentation of results which was explored in section 2.6. The reason given by McGuinness is that web pages typically contain little markup about the contents of the page. Whilst the concept of 'Web 2.0' included the use of technologies like AJAX and flash to create rich user interfaces and a better experience, they also included integration technologies that attempt to make information more readily available. Websites, such as blogs, began to contain links to RSS which stands for RDF (Resource Description Framework) Site Summary (although colloquially Really Simple Syndication) feeds and other Extensible Mark-up Language (XML) based technologies. These

technologies enable the user to take advantage of the information available to them, accessing information from other applications and performing processing upon this information (O'Reilly 2005). One of the key technologies of 'Web 2.0' was the invention and wide-scale implementation of web services. Web services use XML and a description file to enable two computer programs to communicate.

Machine to machine communication is the key issue with the Internet today (Berners-Lee, Fischetti 1999). The majority of the web is human readable with no thought to the ability of a computer to read the information stored on the Internet (Berners-Lee, Fischetti 1999). If it were possible to make all of the information that is stored electronically readable by computer, then a huge number of possibilities are created. Not least the ability to search the information in ways that have never been possible before. For example, imagine being able to ask the search system to show all people who worked on project x and project y and also have knowledge of programming language z. The aim of a machine readable Internet is also the aim of the semantic web. The semantic web, now often referred to as 'Web 3.0' is a web that is as accessible by machines as the current World Wide Web is to humans. However, this is not simple task.

Perhaps the most notable dream for the semantic web is that of Tim Berners-Lee (Berners-Lee, Fischetti 1999). In his book "Weaving the Web," Berners-Lee discusses the dream of computers interacting with each other, understanding the terminology used by different systems and being able to understand and analyse the data available. Content, links and transactions will all become machine understandable. Section 2.7, "Tagging" has mentioned ontologies during the contrast between 'lightweight' tagging based systems and the 'heavier' hierarchical systems that make use of ontologies and taxonomies. Ontologies also lie at the heart of the semantic web and although they are more complex than some systems, their benefits are often greater. The next section reviews the literature on ontologies and taxonomies and how they could impact upon information retrieval.

### **2.8.2 Ontologies and Taxonomy**

The term ontology has been in use for a number of years (Gruber 1993). An ontology can be used to infer information around a specific concept. Simple



ontologies or taxonomies have become commonplace on the World-Wide Web (Noy, McGuinness 2001) and are one of the key methods for representing knowledge as *information* within knowledge management applications (Brewster 2002). Both ontologies and taxonomies have the potential to help map the knowledge around a particular domain in a machine understandable way (Brewster 2002). If the knowledge around a particular domain can be formalised then it can be searched in a more effective way. This can allow optimised discovery of relevant information and the retrieval of the information can be performed in a more structured way enabling users to locate more specific information.

The word *ontology* has its roots in philosophy, and describes the conceptual relationship between concepts, for example, "An ontology is an explicit specification of a conceptualization. For knowledge-based systems, what 'exists' is exactly that which can be represented." (Gruber 1993). Given this statement, an *ontology* is said to represent the notion of capturing a proportion and not necessarily something in its entirety, therefore, a selection of concepts and relationships that ties each together for a given domain. For example, "An ontology defines a common vocabulary for researchers who need to share information in a domain. It includes machine-interpretable definitions of basic concepts in the domain and relations among them" (Noy, McGuinness 2001). Ontologies encompass a wide variety of different manifestations and the distinction between an *ontology* and a *taxonomy* is often difficult. McGuinness (McGuinness 2002) outlines a number of properties that make an *ontology* what it is. In doing so, she refers to the concept of *taxonomies* as a subset of *ontologies*, using *taxonomy* to describe an *ontology* that simply shows a hierarchical structure. Although a *taxonomy* is actually a *form of ontology*, more complex *ontologies* have a far greater range of relationships between concepts than a single hierarchical one found in *taxonomies* and therefore not all *ontologies* are *taxonomies*.

Garshol (2004) states the term *taxonomy* has been so widely used that it may be used to refer to "just about anything" though it usually refers to a form of abstract structure. When going on to formally describe *ontologies* though, Garshol does agree with McGuinness that *taxonomies* usually represent a *single relationship* between concepts.

Although ontologies have existed for a number of years in the early 1990s *ontologies moved beyond the academic domain and into mainstream business* (Hepp 2008). Ontologies range from large taxonomies categorising Web sites, such as Yahoo, to categorisations of products for sale and their relationships as used by Amazon. They are usually created for an environment in which a common understanding of the structure of information amongst users or software agents is required (Ding et al. 2007). The use of ontologies is further defined by Horrocks, Patel-Schneider & van Harmelen (2003), who say "ontologies are expected to be used to provide structured vocabularies that explicate the relationships between different terms, allowing intelligent agents (and humans) to interpret their meaning flexibly yet unambiguously." In Noy and McGuinness's (2001) research they explicitly offer a number of suggestions to why someone would wish to create an ontology:

- To share common understanding of the structure of information amongst people or software agents
- To enable reuse of domain knowledge
- To make domain assumptions explicit
- To separate domain knowledge from the operational knowledge
- To analyse domain knowledge

A further benefit of ontologies is the ability for many systems to perform reasoning and make inferences based upon the data due to the way it is structured (Wang et al. 2004). Reasoning allows much more complex relationships between items to be explored and can even allow intelligent classification of concepts (Pan, Horrocks 2003). As a simple example, Pan and Horrocks (2003) state that reasoning might allow an ecommerce system to classify all items in a particular category that measure less than a certain height or width and weight less than a certain amount into a class of small items that may receive free delivery.

Ontologies have a wide variety of potential uses to increase a systems ability to understand and interact with information and hopefully present more relevant information when used to aid discovery of relevant information. Ontologies can provide structure and context to search domains to enable computer programs to

interrogate them. The aim of this context and understanding is to enable users to effectively find what they are looking for when they come to searching information (Pan, Horrocks 2003). In addition, the context allows the computer to make more informed decisions, increasing its ability to sift through the irrelevant information and only return the information that is relevant to the user in that context. With more relevant information and less irrelevant information the problem of information overload can be reduced.

One area of dispute in the literature surrounds around whether ontologies must have a controlled vocabulary. One of the key points made by McGuinness (2002) is that ontologies have a controlled vocabulary. This is also stated by Horrocks, Patel-Schneider & van Harmelen (2003) as one of the main potential benefits of ontologies. Garshol (2004), on the other hand, states that ontologies are more of a progression from subjects such as taxonomies and thesaurus towards an open subject description without controlled vocabularies. Although some implementations may allow for an open vocabulary, having a closed vocabulary to describe concepts can also be a benefit, allowing a single frame of reference. The choice of an open or closed vocabulary in the use of ontologies can be left to the implementation of that ontology and McGuinness' definition shall be used for the purposes of this research.

Although ontologies offer a wide array of potential benefits they have a number of barriers to their implementation, and these will be discussed further in the following sections.

### **2.8.3 Issues with ontology development**

A growing number of companies now offer support for ontologies and triple stores such as Oracle within in their Oracle 11g database system, incorporating ontologies into their commercial offerings. This offering has made ontology development more accessible, but there are still barriers that have to be overcome. *Common understanding of information is often referred to as one of the major drivers for ontology development (Noy, McGuinness 2001) (Gruber 1993).* However, achieving a common understanding is quite time consuming and is seen as a barrier for many organisations (Farquhar, Fikes & Rice 1997). Reducing

ambiguity is also an important driver but one that also creates a substantial challenge. To remove ambiguity is extremely difficult and one of the key challenges faced when creating an ontology. A large body of literature relating to ontologies relates to the back-end processes involved with ontologies (Bechhofer et al. 2004),(Horrocks, Patel-Schneider 2004). Things such as how the ontology is stored and how to efficiently reason through the data stored in ontologies form the bulk of literature surrounding ontologies. However one of the key difficulties with ontologies actually lies with the creation of the ontology itself.

The formation of an ontology is not a simple task. There are a large number of factors that must be considered when starting to create an ontology. Some authors even state that due to the vast amount of information that is necessary, ontology development has proven extremely expensive and possibly impracticably expensive (Farquhar, Fikes & Rice 1997),(Good et al. 2006). The cost of ontology development and maintenance is often quoted as a key concern of the semantic web (Shadbolt, Berners-Lee & Hall 2006). Although the algorithms and technology associated with ontologies continue to improve, the human side of ontology development is actually the expensive part.

Many authors refer to designing an ontology rather than simply constructing one (Gruber 1995) (Noy, McGuinness 2001), emphasising the large amount of work and consideration that needs to go into the creation of an ontology. One author who specifically remarks on the differentiation is Gruber (1995). Gruber states, "Formal ontologies are designed. When we choose how to represent something in an ontology, we are making design decisions." Gruber outlined several design criteria that should be considered when designing an ontology (Gruber 1995).

- Clarity - An ontology should effectively communicate the intended meaning of defined terms.
- Coherence - An ontology should be coherent: that is, it should sanction inferences that are consistent with the definitions.
- Extendibility - An ontology should be designed to anticipate the uses of the shared vocabulary.

- Minimal encoding bias - The conceptualisation should be specified at the knowledge level *without depending on a particular symbol-level encoding.*
- Minimal ontological commitment - An ontology should require the minimal ontological commitment sufficient to support the intended knowledge sharing activities. An ontology should make as few claims as possible about the world being modelled, allowing the parties committed to the ontology freedom to specialize and instantiate the ontology as needed.

Gruber also states that one of the key issues when designing an ontology is that it will require making tradeoffs among the criteria. His research shows that this is frequently overlooked in ontological development and it needs to be taken into account (Gruber 1995). Ontology development also requires a broad range of skills and knowledge. Obtaining the various skill and knowledge sets can lead to a bottleneck that not only slows down the initial development but also makes it difficult to keep the ontology up to date as new knowledge becomes available (Good et al. 2006). To overcome some of these issues, an iterative approach to ontology development can be adopted to reduce the overall risk as identified by Noy et al. (Noy, McGuinness 2001).

With such a complex development process and so much potential for error, ontology development comes with *great risk and great expense, even when executed correctly.* To demonstrate this point, the Gene Ontology, a well known ontology in the biomedical field, is known to have cost at least an estimated \$16 million at the end of 2006 (Good et al. 2006). However, researchers such as Shadbolt, Berners-Lee & Hall (2006) have shown that in some cases, the costs are recoverable from the overall benefits of developing the ontology.

The literature has shown that there are many issues associated with ontology development. One of the main issues for any organisation is the time and expense of the construction of the ontology. The next section reviews the literature relating to the construction of ontologies and identifies potential areas for improvement.

#### 2.8.4 Creating an Ontology and Automated Approaches

Ontology creation is usually a manual task that is quite time consuming. In recent years there has been some research undertaken into semi and fully automating the process (Hepp, Bachlechner & Siorpaes 2006), (Ruiz-Casado, Alfonseca & Castells 2006), (Ponzetto, Strube 2007). The research has been focused on harvesting information from the Internet to create the ontology. The research concept is to utilise existing content and structures used on the Internet to provide some or all of an ontology. The majority of the research to date, however, has only focused on automatically harvesting concepts from the Internet to create an ontology (Hepp, Bachlechner & Siorpaes 2006) (Ponzetto, Strube 2007). This approach only provides the concepts and in most cases, does not involve complex relationships between the concepts.

In 2006, Hepp, Bachlechner & Siorpaes (2006) researched and discussed the possibilities of taking concepts from Wikipedia and using them as ontological structure. Given that Wikipedia is a consensus driven system and has a great human involvement, the research questioned the validity of Wikipedia forming the basis of an ontology. The research gave some positive results, with the English version of Wikipedia containing over 850,000 entries. The study showed that not only could wikis form the basis of ontological creation but also that the URLs of Wikipedia itself could form an ontology.

A paper called "From Wikipedia to Semantic Relationships" (Ruiz-Casado, Alfonseca & Castells 2006) showed that following a significant amount of work or training, it is possible to take relationships that have been gained from a corpus of information and apply these rules to the information stored within Wikipedia. This allows the extraction of relationships from Wikipedia. Following the training, the system could then be asked to complete a task, for example, to find all people that were born in 1900. Although there were a significant number of anomalies, relationships were discovered and it was possible to extract relationships from Wikipedia to derive useful information (Ruiz-Casado, Alfonseca & Castells 2006).

Research by Ponzetto and Strube (2007) looked at deriving a large scale taxonomy from Wikipedia. The study took the categories of Wikipedia and then attempted to

use methods to search for “is a” and “not is a” relationships within the text. Given these relationships, the system then attempted to create a taxonomy of concepts with “is a” relationships. The research provided the first logical step in automatically creating an ontology from Wikipedia and was arguably competitive to ResearchCyc, an established ontology.

The results of these research studies have shown the potential that exists for extracting information from the Internet and in particular from Wikipedia, one of the largest knowledge bases on the Internet. The key issue however lies within the quality of the results extracted. Only seemingly simple relationships could be extracted from Wikipedia and the fully automated step also did not produce ontologies of a high enough quality for production use.

### 2.8.5 Summary

This section has presented semantic technologies and one of the fundamental concepts of the semantic web, the ontology. The literature has shown how ontologies can be used to harness information and provide machines with a richer and deeper understanding of the content available to them. This understanding can aid the end-user when trying to retrieve information, as the system has a wider understanding of the information the user is looking for and an increased ability to deliver the information that is relevant to the user. The delivery of more relevant information comes with the removal of irrelevant information, reducing the information overload experienced by this user.

However, there are a number of issues surrounding ontology development and this could be the reason why they are not widely used within many organisations, even though the benefits can be shown to outweigh the costs. The main problem is the time and cost it takes to construct one, as ontologies have been known to cost into the millions (Good et al. 2006).

To tackle this problem, research has been conducted into automatically generating an ontology to overcome the associated costs. However, the quality of these automatically generated ontologies has been deemed insufficient and the issues associated with creating high quality ontologies still exist.

## 2.9 Conclusion

The literature review has covered a wide range of topics related to improving information and knowledge retrieval for the end user. It started with an overview of information overload and the impact it can have upon employees' ability to work efficiently and effectively given any information source. It then explored the information defects spectrum that might contribute to information overload or lead to information deficiency, such as barriers to sharing information and knowledge, limitations of current technology (visualisation, clustering, tag clouds and ontologies), processes and culture.

The section on information overload suggested that the root of information overload appears to lie within the cognitive overload domain and that employees can only process a finite amount of information. An increase in relative information will aid a user in the decision process but when irrelevant information is present, it is extremely detrimental to the decision process. Two states of information overload were identified and these were driven by the concepts of pulling and pushing information. A number of studies have focused on the concept of pushing information and overload but relatively few discuss the domain of pulling information and this is where the recipient of the information would have the most control. The literature has shown that a multifaceted approach to reducing information overload caused by technology and improving the interactions of colleagues could present possible benefit to organisations and would represent new research within the field of information overload.

The knowledge sharing section, part of the multifaceted approach, showed the potential issues that exist in sharing knowledge and it identified a number of different methods of knowledge transfer and the difficulties involved. In particular, the barriers that exist that may prevent the sharing of knowledge. The barriers fell into one of three key categories:

- Potential technological issues
- Potential individual issues
- Potential organisational issues



One of the important points raised about these barriers is that they might affect different organisations in different ways, hence why they are deemed potential barriers. If the transfer of knowledge can be increased via information then it is feasible that there could be an increase in information overload, as more information will be shared in the attempt to transfer knowledge. However, if knowledge sharing barriers are not identified and addressed then they could hinder the commercial competitive advantage.

In order to help the end-user with their information needs, as part of the multifaceted approach, emphasis must be placed on search engines to deliver accurate information. The literature review highlighted that the majority of the research had focused on the improvements to the algorithms that power the search systems and the behaviour of how users search. Relatively little research had been focused on the presentation of search results. The research that was found focused on adding visualisation to search results, mainly in the field of web searching, rather than intranet based search systems. Many of these attempts have said to be successful in providing visual additions to search results, but the extent to the success was not provided by the results in the research papers. Out of the papers found on visualisation the most frequent researched area was clustering. The work relating to clustering has shown that alternative visualisation approaches might work at improving information retrieval and that it is possible to identify documents by key words relating to their content. The main advantage of clustering is that it does not require a significant computational overhead.

As part of the multifaceted approach, another area identified to aid information retrieval was that of tagging. The literature had shown that tagging can increase a user's ability to discover relevant documents that have been tagged previously. It also highlighted that tagging can be of benefit to discover relevant information that might not lend itself to full text search used by search engines. The literature also points to tagging as a lightweight alternative to heavier weight alternatives, such as ontology and taxonomies. It has been stated that the barriers to the entry of using a tagging system are far lower than those of taxonomies (Mathes 2004), however a number of barriers to tagging were identified. The effect of these

barriers however was not assessed in detail and the extent to which these barriers are present does not appear to have been determined in the literature.

In the final section of the chapter semantic technologies were reviewed as part of the multifaceted approach as well as one of the fundamental concepts of the semantic web, the ontology. It was identified that ontologies can be used to harness information and provide machines with a richer and deeper understanding of the content available to them. This understanding could potentially aid the end-user when trying to retrieve information as the system has a wider understanding of the information the user is looking for. However, there are a number of issues surrounding ontology development and this could be the reason why they are not widely used within many organisations, even though the benefits can be shown to outweigh the costs. The main problem is the time and cost it takes to construct one, as ontologies have been known to cost into the millions. To tackle this problem research has been conducted into automatically generating an ontology to overcome the associated costs. However, the quality of these automatically generated ontologies has been deemed insufficient and the issues associated with creating high quality ontologies still exist.

Several gaps have been identified by the literature review that provide potential areas for further research into improving information retrieval for the end user. Through information and knowledge sharing and the use of technology to overcome the barriers, it may be possible to improve the relevant information that is retrieved and help reduce the information overload problem. The methodology section that follows shows how this research will be conducted and outlines the methods used.

### 3 Research Methodology

This chapter will detail the research philosophies, methods and approaches taken during this research and the reasoning behind the decisions. Firstly, the chapter identifies some of the philosophical foundations behind the methodologies before presenting a number of research approaches. Following this, a plan of the research that was undertaken shall be given. Once this plan has been given, the methodology looks at the actual research methods that were selected and used within this thesis and explains where and why these methods were chosen.

#### 3.1 Research Philosophies

Within the area of scientific study, two major philosophical research methodologies have emerged and have been widely accepted and quoted for some time. They are namely positivism and interpretivism (Galliers 1992). Another philosophy has more recently emerged called the critical research philosophy. These three approaches are discussed in the following paragraphs.

Positivists believe that all reality can be described by measurable properties and that the knowledge we obtain is based on real life experience or observation in some form (Cornford, Smithson 1996). Positivism is strongly linked to empiricism meaning that "all knowledge must be sensed to be real; faith alone – knowing that it is true because you believe it to be so – is an insufficient basis for explaining a phenomenon or as a foundation for knowledge" (McNabb 2004). Positivism aims to provide objective facts that are measured and cannot be disputed (Cornford, Smithson 1996). The positivist approach is also called the scientific approach by some authors including Galliers (1992). Galliers states that the scientific approach is based upon the positivist philosophy and thus they can be considered as the same thing for the purposes of this work. The positivist approach also assumes that these measurable properties are independent of the observer of these properties.

Interpretivism has been called many different things: post-positivism, anti-positivism and realism are all examples. Interpretivism "asserts that reality is, as well as our knowledge thereof, are social products and hence incapable of being

understood independent of the social actors (including the researchers) that construct and make sense of that reality" (Orlikowski, Baroudi 1991). Put simply, this means that a *degree of interpretation is required to make sense of the findings or observations and that understanding is required*. It suggests that there is a social element to the world around us and attempts to explain why people act in the way, or that findings are, the way that has been observed. Interpretivism also relies on the *idea of shared meanings, something quite important within knowledge sharing* (Walsham 1995).

Along with positivist and interpretivist philosophies, Myers and Avison (2002) also mentions the critical approach. The critical approach has emerged more recently (Cornford, Smithson 1996) and is based upon the idea that there is a more social and political aspect to research and that this can impact upon that research. "Although people can consciously act to change their social and economic circumstances, critical researchers recognize that their ability to do so is constrained by various forms of social, cultural and political domination" (Myers, Avison 2002).

Along with the positivist and interpretivist philosophies exists the concept of both qualitative and quantitative research. Quantitative research collects data that is measured, exact and unquestionable, and that may be proven using statistical techniques. Qualitative data however is not necessarily based upon numbers and exact figures but can contain free explanations and opinion.

Positivism is often referred to as simply quantitative research, although it is argued by some authors that it is possible that qualitative research may be positivist also (Myers, Avison 2002). Equally interpretivism is often referred to as the qualitative approach. Although qualitative research for the most part falls into the interpretive field, qualitative research is not confined to interpretivism.

Depending on the underlying assumptions of the researcher and methods employed qualitative research may be interpretive but it does not have to be (Myers, Avison 2002). Given this, the choice of qualitative research method may be independent from the philosophical choice (Myers, Avison 2002).

Qualitative research is often highlighted as providing a much deeper understanding and a more meaningful set of data than that of quantitative data, however quantitative data is exact. Traditionally there was a feeling that qualitative results would not produce reliable results within research, although confidence has grown and qualitative data is more accepted today (Miles, Huberman 1994).

Remenyi and Money (2004) also states that positivist and interpretive research are not at polar opposites to each other, and suggests that a combination of both of these methods can be beneficial to research (Remenyi, Money 2004). This is a point re-affirmed by Johnson and Christensen (Johnson, Christensen 2007) who show that mixed research can be used and qualitative data can be used to supplement quantitative giving further insight.

The research in this thesis is based on a combination of the positivist and interpretive research philosophies, although it has mainly taken the interpretive approach. The research gathers both quantitative and qualitative data, however in many cases quantitative data was supplemented with qualitative data in a mixed approach to help to understand and interpret the results to provide a deeper understanding of the data. The research included questioning people within the target organisations and also students studying within the field of research. Building on these research approaches, there are a number of methodologies and approaches that can be taken. These will be highlighted in Section 3.2 "Research Methodologies and Approaches." Section 3.3 shall then outline the actual methods available and used by the research presented in this thesis.

### **3.2 Research Methodologies and Approaches**

Different authors name higher-level approaches or methodologies in different ways. Whilst some call them approaches (Galliers 1992), others may call them methodologies (Cornford, Smithson 1996), however they refer to the same thing. They are high level views of the way that research is conducted and the form that the research shall take.

Cornford and Smithson (1996) suggest three key areas of a research methodology and a number of sub-areas that exist within them.

- Constructive research methods:
  - Conceptual development
  - Technical development
- Nomothetic research methods:
  - Formal-mathematical analysis
  - Experiments, laboratory and field
  - Field studies and surveys
- Idiographic research methods:
  - Case studies
  - Action research

Constructive researchers create frameworks, refine concepts and make technical development. The constructive approach allows things to be modelled that do not necessarily have to be present or exist in reality (Cornford, Smithson 1996). In constructive research the models can describe a situation that might not physically manifest itself in reality but that can help to describe theories. Although in some situations it is possible to derive new models and frameworks from existing literature Cornford and Smithson (1996) advise that it may be easier and more appropriate to realise this form of research through empirical observation.

The two further forms of research address two key areas. One examines empirical data with a view to create a generalised law of theory that can apply to a range of cases. The second form examines particular cases or events to provide a more detailed picture of what transpires.

Nomothetic research deals with the field of statistical proof and empirical data or creating a hypothesis and then testing this (Cornford, Smithson 1996). These methods are far more related to the positivist approach previously mentioned. Many of the research approaches that fall under this category involve the collection of data and analysis with a view of providing generalised insight (Cornford, Smithson 1996). It is important to consider generality and aspects such as the sample chosen to correctly represent this when performing this type of research (Cornford, Smithson 1996).

Idiographic research is related to exploring cases and events and providing an understanding of what is happening. The aim of Idiographic research is to understand a particular phenomenon within its own context (Cornford, Smithson 1996). Idiographic research may involve case-studies or take the form of action research. In information systems research it is common that case-studies may form a large part of this research method (Cornford, Smithson 1996).

### 3.2.1 Practice Driven Research

Further to these typical research methodologies or approaches is a more recent concept called practice driven research (Zmud 1998). Although not so widely accepted as the methods suggested by Galliers, practice driven research is of great relevance to this research. Practice driven research differs from traditional research driven methods, as it involves far more input from sponsors regarding the work in hand.

Table 5 shows Zmud’s view of practice driven development as an alternative to researcher driven development.

**Table 5- Practice vs. Researcher driven research (Zmud 1998).**

| Practice Driven   | Researcher Driven  |
|---|--|
| Topic defined by sponsor’s end-point is a “moving target” framed by nature and phenomena designed jointly by the researcher and sponsor | Topic defined by researcher’s end-point is initially known framed by research model designed by the researcher |

Zmud (1998) also identifies four factors that characterise this form of development. The four issues are as follows:

- The sponsors, not the research team, determine the topic or phenomenon to be studied. The research team agrees to study the area put forward by the sponsor, although the exact details shall be worked out between both the sponsor and the research team.
- There is no initial specified research outcome. Sponsors feel the need to know more about the subject being studied before they are willing to commit to a specific outcome. This does not mean that there are no

objectives but that they are revisited throughout the lifecycle of the project and may change as learning occurs.

- Research is framed by the current understanding of the researcher and the sponsor rather than a well defined research model.
- The research team is expected to propose and direct the research design. However the sponsors react and suggest revisions to the design due to their perspectives and understanding of the issue at hand.

Practice driven research leads to two key forms of scholarly publications and academic output. The first are those that derive from project deliverables and outcomes of the research. The second comes from both extensive exploration and research within a research topic, and also insights that are gained from the project.

There are however, a number of issues associated with practice driven development. According to Zmud, (1998) these are centred on the following areas:

- Gaining access to research sites – both identifying sites and then convincing executives at those sites to participate in the research can form a challenge
- Gaining access to informants at a research site – both identifying informants and setting up interviews or research work with these informants may be difficult however; using a champion within the organisation can be of benefit here.
- Maximising information from informants – keeping the interviews short may be required in order to prevent research distracting employees from their core job. The recommendation here is to make interviews no longer than one hour.
- Ensuring that research findings are interesting and meaningful for sponsors. A common problem to researchers is to become so engrossed within the organisation that they forget the findings must be of relevance to both them and the sponsor.

Whilst these issues are important they can be overcome so the research is of value to both the target organisation and the researchers

The following section will now continue to look at the actual research approaches that can be taken by researchers.



### 3.2.2 Research Approaches

This section shall continue to investigate the actual approaches that can be taken by researchers to gather data for research. In 1984, Vogel and Wetherbe (Vogel, Wetherbe 1984) provided a classic taxonomy of research approaches. These approaches have been used many times as the basis of information science research approaches (Cornford, Smithson 1996). These approaches were further refined by Galliers(1992).

Galliers (1992) categorised the approaches laid out by Vogel and Wetherbe and expands upon these to form a table of approaches under the scientific (positivist) and interpretive headings. Table 6 shows Galliers' approaches.

**Table 6 - Positivist and Interpretivist approaches**

| Positivist             | Interpretivist             |
|------------------------|----------------------------|
| Laboratory experiments | Subjective / argumentative |
| Field experiments      | Reviews                    |
| Surveys                | Action Research            |
| Case studies           | Case Studies               |
| Theorem proof          | Descriptive / interpretive |
| Forecasting            | Futures research           |
| Simulation             | Role / game playing        |

Galliers (1992) provides a list of strengths and weaknesses for some of the approaches shown in Table 6 along with their key features. These approaches and their features can be seen in Table 7.

**Table 7 - Galliers approaches and their features** (Galliers 1992)

| Approach                          | Key Features   | Strengths   | Weaknesses  |
|-----------------------------------|--|---|---|
| Laboratory Experiments            | Identification of precise relationships between chosen variables via a designed laboratory situation, using quantitative analytical techniques, with a view to making general statements applicable to real-life situations  | The solutions and control of a small number of variables which may then be studied intensively  | The limited extent to which identified relationships exist in the real world due to oversimplification of the experimental situation and the isolation of such situations from most of the variables that are found in the real world.                        |
| Field Experiments                 | Extension of laboratory experiments into the real-life situations of organisations and/or society.   | Greater realism; less artificial/sanitised than laboratory situation  | Finding organisations prepared to be experimented on. Achieving sufficient control to enable replication, with only the study variables being altered   |
| Surveys                           | Obtaining snapshots of practices, situations or views at a particular point in time (via questionnaires or interviews) from which inferences are made (using quantitative analytical techniques) regarding the relationships that exist in the past, present and future. | Greater number of variables may be studied than in the case of experimental approaches. Descriptions of real world situations. More easy / appropriate generalisations. | Likely that little insight is obtained relating to the causes/processes behind the phenomena being studied. Possible bias in respondents.   |
| Case Studies                      | An attempt at describing the relationships which exist in reality, usually within a single organisation or organisational grouping.  | Capturing 'reality' in greater detail and analysing more variables than is possible using any of the above approaches   | Restriction to a single event / organisation. Difficulty in generalising, given problems of acquiring similar data from a statistically meaningful number of cases. Lack of control variables. Different interpretations of events by individual researchers. |
| Simulation                        | An attempt at copying the behaviour of a system which would otherwise be difficult/impossible to solve analytically, by the <i>generation/introduction of random variables.</i>  | Provision of an opportunity to study situations that might otherwise be impossible to analyse.  | Similar to experimental research in regard to the difficulties associated with devising a simulation that accurately reflects the real world situations.  |
| Subjective argumentative research | Creative research based more on opinion / speculation than observation.  | Useful in building theory that can be subsequently be tested.   | Unstructured, subjective nature of research process. A likelihood of biased interpretations.  |
| Forecasting/futures research      | Use of such techniques as regression analysis and time series analysis to deduce possible events   | Provision of insights into likely future occurrences in situations where existing relationships may not hold true in the future.  | Complexity and changing relationship of variables under study. Lack of real knowledge of future events.   |

Although the table presented by Galliers (1992) highlights a number of approaches, Cornford and Smithson (1996) presented six approaches most likely to be of use to researchers, especially those in the field of Information Systems.

These approaches were namely:

- Laboratory Experiments;
- Surveys;
- Reviews;
- Action Research;
- Two forms of case studies
  - Descriptive and
  - Interpretive.

Each of these options are explored in more detail below.

### **3.2.3 Laboratory Experiments**

The concept of a laboratory experiment implies that an experiment takes place within a controlled environment. Laboratory experiments can be used to control certain variables and modify others, whilst observing the results. Most of the data gathered by laboratory experiments will be quantitative, the experiment will also relate to a limited number of phenomena.

The use of experiments in a laboratory environment will not exactly represent an equivalent situation within an organisation however this does not have to be a major concern (Cornford, Smithson 1996). It may be possible to be confident that the sample used to collect the data will generalise sufficiently to represent a more generic audience. It is important to consider the sample chosen however and justify any generalisation that is intended.

### **3.2.4 Surveys**

Surveys present a cross-sectional representational view of state at a given point of time (Cornford, Smithson 1996). Surveys usually consist of either questionnaires or interviews. Although not mentioned by Cornford and Smithson (1996), this section shall also introduce focus groups as a method of gathering a collectively

formed qualitative consensus around a topic area and the Delphi method as an additional means of reaching group consensus.

#### **3.2.4.1 Questionnaires**

Questionnaires can take a number of forms. Questionnaires may either be self-administered or involve an interviewer. When self-administered, the participant must understand the question and interpret its meaning by themselves (Bryman, Bell 2003). This method provides a number of benefits. One such benefit is the ability for the participant to perform this questionnaire whenever they have time rather than at a pre-determined time. There is also little cost associated with this method and it reduces the chance of bias being introduced by an interviewer (Bryman, Bell 2003). In addition the questionnaire can be more anonymous. The key disadvantage of this method is an interviewer is not present, therefore the participant cannot ask questions for clarification and the interviewer cannot gain a deeper understanding of the participant's true intentions. "Because there is no interviewer present in the administration of the self-completion questionnaire, the research instrument must be especially easy to follow and its questions specifically easy to answer" (Bryman, Bell 2003).

*Questionnaires may be distributed using a variety of delivery methods.*

Interviewer administered questionnaires may be performed both in person, over the telephone or via a video conferencing system. Telephone questionnaires lack the presence that is gained when the questionnaire is administered in person. Likewise an interviewer administered questionnaire may either be given individually or on a group basis. Self-administered questionnaires may take place either physically with a hard copy of the questionnaire that can be filled in on paper or virtually through the use of an online questionnaire system. The benefits of self-administered questionnaires are present with an online system along with a higher degree of control over the system and the ability to invite the audience via email allowing an entire geographically dispersed department to be easily reached. Another important factor is all forms of questionnaire allow a large number of variables to be addressed within a short amount of time in comparison.

In terms of the type of questions asked, it is important to consider the implications of either multiple choice and fixed (also referred to as closed) questions or open ended ones. The use of multi-choice questions is recommended in the use of self-completion questionnaires (Bryman, Bell 2003). However if only fixed closed-ended choices are available then it is possible that bias may be introduced by the questionnaire designer (Krueger, Casey 2000).

Free text responses to questions can also be used in order to allow the participant to elaborate and to determine why participants give the answers they gave and to give a deeper understanding into respondents answers and allow a more qualitative understanding. Although open-ended questions are harder to analyse they have the benefit of allowing the respondents to express themselves in their own words instead of those chosen by the researcher (Weisberg, Krosnick & Bowen 1996).

Cornford and Smithson (1996) list a number of factors when developing a questionnaire that must be considered.

- Effort – The cost and difficulty in developing and deploying the questionnaire.
- Response – Response rates to questionnaires can often be poor. This can be worsened by poor questionnaire design however management buy-in can strongly support the response rate.
- Bias – Along with researcher introduced bias, bias is an important factor in questionnaires and should always be considered wherever possible. Even the sample selected may introduce bias into a questionnaire and where possible the sample should represent the intended audience
- Well understood topics – Questionnaires are only suitable for topics that are fairly straightforward. If questionnaires need detailed explanation than is possible in the question text then it may not be suitable to use questionnaires.
- Focus of questions – As previously mentioned the types of questions should be considered such as open or closed questions and in addition so should the number of questions and length of time questions will take to answer.

In addition Cornford and Smithson (1996) indicate it is important to form clear questions that are easy for respondents to understand. Questions should be constructed so that they do not cause the respondents to make unnecessary assumptions and so that they are comprehensive, allowing as much insight as possible from a single question. Finally, it is important that questions do not require respondents to seek additional information in order to be able to answer the question.

Bryman and Bell (2003) highlight the benefits of conducting “a pilot study before administering any self-completion questionnaire or structured interview to your sample”. Given that there is no interviewer present, pilot studies are particularly useful to assess the usability of self-completion questionnaires. This gives an opportunity to reduce any confusion that may occur in the interpretation of questions before the questionnaire is administered.

#### **3.2.4.2 Focus Group**

Focus groups have become an extremely popular and more recently, a widely accepted method of gathering research. Focus groups are said to have grown from the fact that some social scientists had reservations about the validity of data obtained from a questionnaire. A questionnaire with fixed closed-ended choices means the respondent was limited to the choices outlined by the questionnaire designer and therefore questionnaires alone may inadvertently bias answers that are given (Krueger, Casey 2000).

A focus group can be used to collect evidence from a highly specialised group of people (Remenyi, Money 2004) and can form a simpler method of collecting data from these experts.

A focus group is a method of interviewing more than one interviewee at a time, thus it may be described as a group interview. Krueger and Casey describe a focus group study as “a carefully planned series of discussions designed to obtain the perceptions on a defined area of interest in a permissive, non-threatening environment” (Krueger, Casey 2000). The discussions are intended to be relaxed and although group members can influence each other by responding and

commenting to their ideas, these can lead to in-depth discussions and insights into a given area.

The concept of having a given area of study is one of the key factors of focus groups. Bryman and Bell (2003) make the distinction between a focus group and a standard group interview. There are three key points mentioned that create a distinction.

The first, as already mentioned, is the fact that focus groups have a specific theme or topic rather than a wide area of discussion. The second point is that focus groups are not carried out in order to save time and money by interviewing more than one person at once. They do so to encourage the sharing and group discussion. This leads to the third and final point that the focus group is interested in the ways that issues are discussed between the group and the way that the group discusses and handles the issues rather than how each individual does.

Focus groups are about gaining the group's perception rather than combining those of the individuals. Given these points, Bryman and Bell still refer to an unclear boundary that often lies between focus groups and group interviews. In the majority of cases a qualitative approach is taken within focus groups. With researchers having a fairly open and unstructured approach to the focus group, there should be a moderator who facilitates the discussion, finding the fine balance between guiding the discussion and influencing or intruding upon it.

Focus groups are often recorded or transcribed and this allows for a number of things. Firstly, it allows the researcher to see who said what within the focus group and secondly, recording allows the researcher to see the tone in which it was said. This allows the researcher to gain more understanding of the emotions of the participants when they took part in the discussions (Bryman, Bell 2003).

Transcription also allows the users to later reflect on the roles that different people took within the discussions if this is deemed important. There are privacy implications of recording and transcribing however. Many participants may not be comfortable with being recorded and this may present an issue. Privacy is an important factor in many organisations and in research. In order to gain true insight it is important that privacy is preserved. The literature review has

highlighted already how competitive organisations can be. If participants of the focus group are worried that anything they say may have consequences for their jobs then they shall be considerably more reserved.

The size of a focus group should also be a key consideration. Remenyi and Money (2004) simply states that a minimum of four participants are normally required. Bryman and Bell (2003) clarify this however, stating that a number of different group sizes have been suggested and used in the past, however six to ten members is normal practice. Larger groups offer more variance or option, but they can become more difficult to manage. In smaller groups, more demand is put on the participants. The suggestion is to keep groups smaller if it is anticipated that many of the users will have a lot to say on the subject and are heavily involved or emotionally attached. The final point that should be addressed with focus groups is how the questions are asked. Some researchers believe it only necessary to ask a small number of questions and allow conversation and debate. Others prefer to have a structured set of questions to work from.

#### **3.2.4.3 Interviews**

Although focus groups can be of more benefit when interviewing multiple respondents, interviews may also be useful. One-to-one interviews are possibly the most widely employed method in qualitative research (Bryman, Bell 2003). The interview may be structured, semi-structured or unstructured and each has their benefits and disadvantages. It is important that the interview structure is flexible in order to allow it to adapt to the responses of the interviewee. Some authors also suggest that the interviewer should never discourage interviewee's from going off on a tangent, as this can give a greater insight (Bryman, Bell 2003). However, this may be expensive and time consuming for both the company and the interviewer. If it is important that consistency in the results gathered occurs, then structured or semi-structured interviews may be of more benefit than unstructured ones. However, unstructured interviews can often lead to a larger breadth of information as participants are free to discuss the issues relevant to them and are not confined by the structure proposed by the researcher. Semi-structured interviews can provide a good combination of both structured and unstructured



interviews allowing the interviewee to roam and describe freely whilst still retaining an element of control.

#### **3.2.4.4 The Delphi method**

In addition to the measures described in this section the Delphi method shall be discussed here in order to highlight the differences between the Delphi approach and a focus group.

The Delphi method is used for finding group judgements and follows a set of procedures based upon the concept that "two heads are better than one" (Dalkey 1969). The aim of the method is to discover the most reliable consensus of opinion from a group of experts.

The first Delphi experiment took place in the late 1940s and has seen several evolutions but sees relatively infrequent use today (Landeta 2006). The Delphi method structures the responses of respondents in order to resolve a complex problem. The characteristics of a Delphi study are as follows (Landeta 2006):

- It is a repetitive process. The experts must be consulted at least twice on the same question, so that they can reconsider their answer, aided by the information they receive from the rest of the experts.
- It maintains the anonymity of the participants or at least of their answers, as these go directly to the group coordinator. This means a group working process can be developed with experts who do not coincide in time or space and also aims to avoid the negative influence that could be exercised by factors in the individual answers in terms of the personality and status of the participating experts.
- Controlled feedback. The exchange of information between the experts is not free but is carried out by means of a study group coordinator, so that all irrelevant information is eliminated.
- Group statistical response. All the opinions form part of the final answer. The questions are formulated so that the answers can be processed quantitatively and statistically.

The Delphi method is not without weakness, for example the bias introduced by the researcher, the lack of motivation and reinforcement that might be provided by the support of other group members and the time to carry out the method are all highlighted by Landeta (2006) along with many other weaknesses.

### **3.2.5 Reviews**

Reviews can be considered to look backward into previous research. Most research will conduct a literature review, investigating previous literature relating to their topic of interest and research. Reviews can be used to develop frameworks and gain insight and understanding into research that already exists within an area of research. An introductory literature review can also be used to learn from the prior work and to formulate their plans avoiding repetition. A literature review may also help researchers to identify methodologies used by others conducting similar work thus allowing the researcher to perform their research in a more informed manner.

Although reviews can be used by researchers to gain a personal understanding, a good review can also form a contribution in its own right giving a more concise and refined understanding of the chosen area (Cornford, Smithson 1996).

### **3.2.6 Case Studies**

There are two key and related forms of case study identified by Cornford and Smithson (1996), namely descriptive case studies and interpretivist case studies.

#### **3.2.6.1 Descriptive Case Studies**

A case-study is an in depth exploration of one situation (Cornford, Smithson 1996). A case study might for example be used to chart the implementation of a new computing system within an organisation or the development of a strategy over time. The analysis of case study research takes place within a single situation providing cross sectional snap-shots (Cornford, Smithson 1996).

Due to the nature of case study research, the data collected may be obtained through a variety of means. Case study research is therefore often recommended for topics or areas of research that are novel or have little pre-existing theory (Cornford, Smithson 1996).

One significant disadvantage of case studies is that it can become difficult to generalise findings due to the study of a single situation.

### **3.2.6.2 Interpretivist Case Studies**

In an alternative view of that presented by Galliers (1992), case studies can be seen as a wholly interpretivist approach. Cornford and Smithson (1996) state that some authors argue that case studies can be used to provide a deeper understanding of a phenomenon rather than a particular situation.

Walsham (1993), one of the authors quoted by Cornford and Smithson (1996) is less concerned with the problem of generalising case studies stating that "the validity of an extrapolation from an individual case depends not on the representativeness of such cases in a statistical sense, but on the plausibility and cogency of the logical reasoning used in describing the results from the cases and in drawing conclusions from them"

In Walsham's view, case studies can be used to draw logical conclusions that are adequate for generalisation.

### **3.2.7 Action Research**

Action research may also be termed collaborative research. Action research occurs when the researcher, traditionally an observer, actually takes part in the events or scenarios along with the subjects of the research in the problem situation.

The key aspect of this form of research is that the researcher is actually involved and is an influencer of the research problem, conducting research whilst attempting to create change. As an example, a researcher may work within an organisation in a systems development role helping with design, programming, analysis and testing. Research output from action research takes two formats. The first is the change that occurs within the activity and results from the theoretical knowledge of the researcher. The second is the experience gained by the researcher that can be documented through reflection. Action research is most appropriate when the researcher has a particular and specific skill to offer and a research site may be found willing to allow the research to be put into practice. Due to the nature of this research, research takes place within a real world scenario, similarly to the case

study but in this situation the researcher has a much fuller involvement and therefore understanding of the issue and result (Cornford, Smithson 1996).

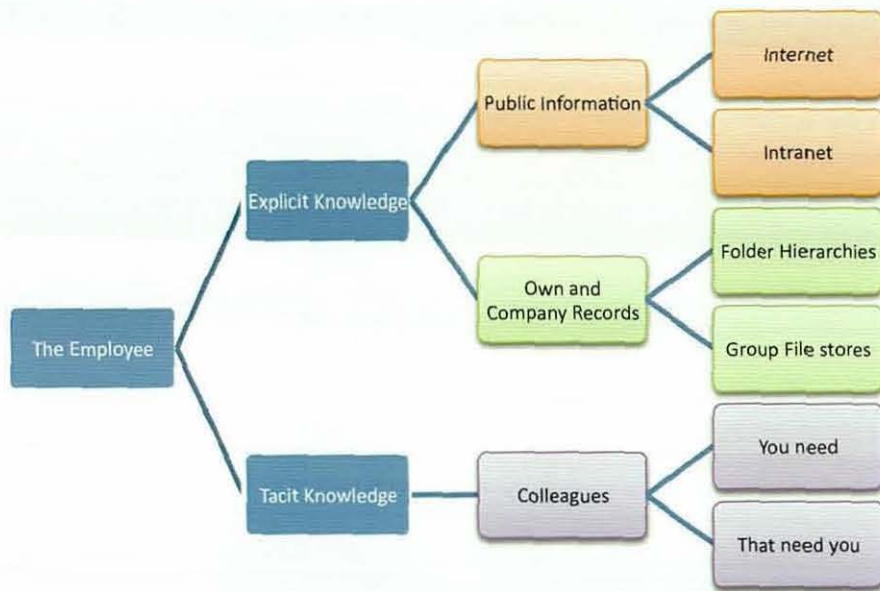
The key risk with action research surrounds the concept that the researcher may become too involved within the organisation and fails to see the wider picture and the research aspect of the task they are achieving.

Given the various different methodologies and approaches available, the next section discusses the overall aim of this research and the research philosophy, methodologies and approaches chosen to achieve it.

### **3.3 The Research Framework**

The research takes a multi-faceted approach towards reducing the information overload problem for organisations, as shown by Figure 9. This section details the areas of research that will be undertaken in-line with the aim and objectives outlined in Chapter One. The research philosophy and methods chosen to achieve the aim and objectives will also be detailed within the section.

The literature review found that information overload presented a significant problem to organisations and could impact upon an employee's ability to make decisions. Although a large amount of research existed within the area of information overload, the majority of that work focused upon information that was pushed and not pulled.



**Figure 9 - An overview of the sources of information investigated by the thesis**

The literature review identified several methods that could be used to improve a user's ability to find relevant information. In addition, the literature review highlighted that several barriers existed and might prevent a number of solutions from being applied. As set out in Chapter One, this thesis aims to reduce these barriers wherever possible and apply the research approaches to allow an organisation to improve its ability to find relevant information and reduce the information overload problem. In order to achieve this, the thesis concentrates upon three key sources of information. These sources are:

- Information from colleagues
- Information from Internets and Intranets
- Information from a users own and company records

The different colours on the diagram, Figure 9, show each of the information sources outlines above and their relation to the employee.

Once combined the approaches, developed as part of the research detailed within this thesis, will form a framework that can be used by organisations to improve their ability to find relevant information and reduce the information overload problem.

### 3.3.1 Information from colleagues

The literature review found a large body of work relating to knowledge sharing and found, although it is difficult to determine the effectiveness of knowledge sharing, the barriers to successfully sharing knowledge can be identified. The literature review also found that the barriers would impact different organisations to different extents. What was not found, however, was a method to measure the extent to which each of the barriers affects an organisation.

The planned research investigates the impact that the barriers have upon two organisations, PharmaCo and SoftwareCo. The aim of the research is to allow an organisation to determine the extent to which the knowledge sharing barriers affect their organisation. Once the barriers are determined then suggestions as to how to reduce these barriers can be made. With the barriers to knowledge sharing reduced then the information overload issue might also be reduced as more relevant and less irrelevant knowledge is available to the user. In turn, it is expected that with more relevant information available to an employee and better sharing practices in place, the employee can share more relevant information.

In order to determine the barriers that exist within an organisation, it would be necessary to gather a representative snapshot of the state of potential barriers towards knowledge sharing within the organisations. In order to achieve this representational cross-section of the state of an organisation, a survey was used. An alternative to the use of a survey could be action research, however, given that the literature review found it was difficult to measure the success of knowledge sharing, it would be difficult to assess any change that may occur during the action research.

Given that surveys would be used, there were a number of alternatives available for assessing the knowledge sharing environment and the barriers that existed within the two organisations. This could include the use of questionnaires, interviews or even focus groups.

It was decided that a questionnaire would be used to collect the data for a number of reasons. Firstly, using a questionnaire would allow a far greater number of employees to be questioned and more questions to be asked than other methods

given the limited amount of time employees of the organisation could spare. The questionnaire aimed to give as representative sample of each department as possible and was therefore sent to the entire department and championed by high level employees. The fact that questionnaires can help to preserve anonymity in answers was also a key consideration for their use. The questionnaire was self-administered which brought great benefits within the geographically dispersed organisations enabling data to be collected from locations that would otherwise have been out of reach for the questionnaire. In this case due to the geographically dispersed nature of the organisations and the difficulty and cost of scheduling specific times that an interviewer must be present, the self-administration method was of great benefit. This questionnaire was delivered as an online questionnaire in order to allow a greater flexibility for the participant and allow them to perform the questionnaire anonymously and at a time that suited them. This was an important consideration in order to increase employee participation.

Most questions within the questionnaires were multiple-choice questions. This would allow the participants to answer the questions quickly and allow quantitative analysis to take place to understand how participants felt. Free text responses to some questions were also requested in order to allow interpretation and gain a deeper understanding of a user's insight in a more qualitative way.

In addition, several pilot studies were performed before the questionnaire was given to its full intended audience and the questionnaire would also make use of a phased delivery to ensure that any issues with the questionnaire could be rectified before the entire population received it. This was also of benefit to establish understanding and ensure that participants who were not native English speakers could understand the questionnaire appropriately.

Chapter 4 titled "Assessing the Knowledge Sharing Environment" presents and discusses the questionnaire design to enable an organisation to determine the extent to which many of the barriers affect their employees.

### **3.3.2 Information from Internets and intranets**

The most common source of information overload described in literature was the Internet. The Internet represents a huge body of information that is increasing by

the day. In 2005, Gulli and Signorini estimated that there were over 11.5 billion indexable pages on the World Wide Web (Gulli, Signorini 2005). Using similar methods, a more recent estimate by De Kunder (2008) places this estimate to be closer to 60 billion pages.

When attempting to find information on the Internet, one of the first places that most people will visit will be a search engine. Search engines also provide the first point of contact for the discovery of information on many intranet sites. The literature found that although significant improvement to search engines had been made and research existed into the behaviour of search engine users, there was relatively little literature relating to the presentation of search results.

A small number of visualisation attempts did exist and the literature showed the potential for visualisation of search results, especially with the more visual and less systematic users of search engines. The literature identified a number of disadvantages that existed which might prevent the success of visually displaying the result of search queries. The disadvantages identified included requiring a significant processing time or even that the system is only of benefit when comparing documents rather than performing a navigational search. These disadvantages in many cases may actually have created barriers preventing the visualisations from being as effective as they could and preventing their adoption.

As detailed in the literature review, there are a number of approaches to aid effective retrieval of information from search systems. The approach chosen was to look at the affect of Concept Clouds upon the ability of an end user to find relevant information. The effects on the end user could have been determined by a number of different methods such as surveys using questionnaires, interviews or focus groups, laboratory experiments or even through implementation within the organisation and case studies.

Unfortunately it was not possible to implement the system in a real organisation in order to perform a case study. Even if this were possible, it would have required a huge effort to deploy across a representative sample and ensure that the system were sufficient to be used by employees all day. Several pieces of literature examined in the literature review actually made use of field or laboratory



experiments to test enhancements suggested by the authors (Paek, Dumais & Logan 2004) (Sebrechts et al. 1999). The use of experiments would allow a fully quantitative representation of the performance of the system and its ability to reduce the time required by users to find correct and relevant information. Two laboratory experiments were used to assess the ability of the Concept Cloud system to reduce the barriers experienced. Obtaining a representative sample was a key consideration for the experiments. Rather than using employees from a cross section of organisation the experiments used undergraduate students. The undergraduate students were on an Information Retrieval module that would soon be going into the industry within this field. It was felt that this sample would sufficiently generalise to a wider population of knowledge workers however this generalisation must be highlighted as a potential weakness. In order to test the system, users were matched based on their ability to answer preliminary questions and split into two groups. One group would use the system and one would search without the system.

Chapter 5 titled "Alternative Search Visualisation – Concept Clouds" presents and discusses the laboratory experiment design to enable an evaluation of the affect visualisation can have on users to improve their information retrieval abilities.

### **3.3.3 Information from own and company records**

In addition to visualisation, the literature review showed that tagging could be of benefit when users are trying to find relevant information. Tagging was especially of benefit when attempting to discover information that could not be full-text searches such as video and images. The literature review showed that although tagging had been used in other fields, its use on the Internet had grown significantly in recent years. In addition, although tagging was said to replace hierarchical structures, it had not been applied to one of the most common hierarchical structures found within an organisation, a computer's file system. After investigating tagging in general, Chapter 5 also presents and evaluates a system that allows users to tag files and then retrieve those files based upon the tags added to the file. The aim is to allow users to make use of tagging to help them to discover relevant documents more easily.

To determine the extent to which tagging could be of benefit to finding relevant information, a tag based file system was created and evaluated. The effects on the end user could have been determined by a number of different methods including survey methods, experiments or case studies. It might also be possible to use action research to see if the use of tagging could be fully implemented within an organisation.

It was decided that a combination of both questionnaires and focus groups would be used to collect data in this chapter. The key desire was to understand group consensus surrounding both tagging and the proposed systems ability to help users discover information and documents. It was also felt that due to the technical nature of the system, an expert group that would fully understand the system and its potential benefits would have to be established. Unfortunately the organisation involved was not happy with the focus group being transcribed or recorded, as is often the case. In addition, it was not possible to identify a large enough sample that would allow the sole use of a questionnaire that could be used to gather a representative quantitative view. Being outside the UK, interviews with each of the participants would not be possible because of the expense involved in performing this exercise and the time constraints of employees. Further to this, the use of the Delphi method would not be possible due to the time constraints of the organisation's employees and substantial amount of time it would take and the one-to-one time with individual users to perform the Delphi method.

In order to overcome the barriers presented, a combination of methods were used, questionnaires and focus groups. Using questionnaires would allow the opinions of participants to be recorded anonymously whilst the focus group would allow the group to reach consensus of opinion. Although the focus group could not be transcribed or recorded, the company did agree that 'sound bites' could be noted and used to highlight the arguments presented by the participants.

Chapter 5 titled "Alternative Search Visualisation – Concept Clouds" presents and discusses the questionnaire and focus group design to enable an evaluation of the affect tags and tag based filing has on users to improve their information storage and retrieval abilities.

### 3.3.4 Ontologies

The literature review showed that although a number of automated attempts to create ontologies existed, the quality of the ontologies produced were not sufficient to use. The objective of the next phase of the research was to develop and assess the potential of a semi-automated ontology development approach. The approach attempted to discover information from sources such as Wikipedia and Google in a semi-automated way to help reduce the cost involved in creating an effective ontology. The ability of a semi-automated approach to ontology development could be examined using a number of the methods already presented.

Several methods were not possible however. Firstly, it was not possible to perform field or laboratory experiments due to the significant amount of time it would take to create an ontology and the need for many systems to be available for users to create their ontology's with.

A case study approach or the use of action research, were not thought to be possible due to the time and involvement it would take to implement a full ontology during this research. It could therefore not be guaranteed that the ontology would be created by the time the research was complete.

Given this, the survey approach provided the best alternative, and two key surveys were conducted as part of Chapter 5 using questionnaires, focus groups and interviews.

The first element of data collection used a questionnaire. The questionnaire was used to ask a wide range of questions to undergraduate students that had experienced the ontology creation process via the semi-automated approach and were familiar with ontology creation. The students were used to gather quantitative data from a larger collection of respondents that would generalise to represent those who had worked with ontologies. Although the generalisation from students to those who had worked with ontologies was possible, it was felt that the students may not have encountered a sufficient number of ontologies or information retrieval problems.

A focus group was therefore created within SoftwareCo, one of the case-study organisations, in order to gather an expert consensus. Although not all of the members of the focus group were experts in ontology development, they were all working with or around ontologies and all worked within the search or information retrieval software field. It was felt that these participants would be able to provide a more focused expert opinion. Unfortunately, as in a previous focus group, transcription or recording was again not possible and so 'sound bites' and notes were taken and agreed by the participants. The participants were also given small questionnaires that they could use to describe their thoughts anonymously in a way that they were happy to be shared.

Following the surveys, an interview took place with an employee that had been using the approach developed to implement an ontology within the organisation but unable to attend the focus group. This employee was chosen for interview as at the time they were the only employee that had fully completed an ontology using the system.

Chapter 7 titled "Ontology Development" presents the semi-automated system and discusses the questionnaire design to enable an evaluation of the semi-automated ontology creator on the end users ability to become more effective and efficient at retrieving information.

### **3.3.5 The Research Methodology**

The research in this thesis comprised of a combination of both positivist and interpretive philosophies. In addition, it takes the form of both quantitative and qualitative research. Table 8 shows the chapters of this research and the philosophies, methodologies and approaches employed by this research.

**Table 8 - The approaches employed by this research**

| Chapter   | Evaluation Philosophy                      | Evaluation Methodology | Evaluation Approach   |
|---|--|------------------------|---|
| 2 - Literature Review   | Interpretive                               |                        | Review  |
| 4 - Assessing the Knowledge Sharing Environment               | Positivist with elements of interpretation | Nomothetic             | Surveys - Questionnaires within two case study organisations.   |
| 5 - Alternative Search Visualisation – Concept Clouds         | Positivist                                 | Nomothetic             | Two Laboratory Experiments  |
| 6 - Using Tagging to Discover Networked and Local Information | Positivist with elements of interpretation | Nomothetic             | Combination method using Focus Groups and Questionnaires within a case study organisation.  |
| 7 - Ontology Development                                      | Positivist with elements of interpretation | Nomothetic             | Questionnaires given to undergraduate students.<br><br>Surveys - Combination method using Focus Groups and Questionnaires given to a case study organisation<br><br>Interview with one member of the case study organisation. |

Finally Chapter 7 shall propose a framework showing how organisations can identify and combat the problem of information overload at the various sources discussed in the previous chapters.

Given the philosophies, methodologies and approaches taken Table 9 highlights the objectives of this research and which of the chapters aim to fulfil those objectives.

**Table 9 – The research objectives and chapters**

| Chapter   | Objective No. | Objective Details  |
|---|---------------|--|
| 2 - Literature Review   | 1             | Critically review literature on information overload and other information defects and the effect it has upon information workers.   |
| 4 - Assessing the Knowledge Sharing Environment               | 2             | To establish through the use of a questionnaire the extent that multi-faceted barriers hinder information and knowledge sharing.   |
| 5 - Alternative Search Visualisation – Concept Clouds         | 3             | To determine how information overload can be reduced through the investigation and development of summarisation techniques.  |
| 6 - Using Tagging to Discover Networked and Local Information | 4             | To develop and assess alternative approaches to storing information to improve information retrieval and reduce information overload.  |
| 7 - Ontology Development                                      | 5, 6          | To establish the role ontologies can play in the retrieval of relevant information and reduction of information overload, the complexities of ontology development and the barriers to their use.<br><br>Investigate alternative approaches to traditional ontology development tools that may be used by subject experts rather than ontology specialists to aid in the creation of ontologies that can help the discovery of relevant information. |
| 8 - Conclusions and Recommendations Framework                 | 7             | Establish an information overload framework to provide direction and solutions to the information overload problem experienced by information workers.   |

### 3.4 Summary

This chapter has identified two major research philosophies both positivist and interpretive. It has also stated that research in this thesis shall be based on a combination of these positivist and interpretive theories. The data gathered during this thesis was both quantitative and qualitative. The quantitative data was often used in conjunction with qualitative in order to help understand and interpret that data in an interpretive way.

Finally this research highlighted the actual approaches that would be taken within this work. These research methods were:

- Literature Review;
- Laboratory experiments;
- Questionnaires;
- Focus Groups and
- *Interviews*

Further information relating to how the questions were chosen for questionnaires can be found within the relevant chapters.

## 4 Assessing the Knowledge Sharing Environment

### 4.1 Chapter Preface

Knowledge sharing represents an area with the potential to help reduce the irrelevant information shared between employees. If employees are able to openly share relevant information and reduce the irrelevant information that they share with colleagues, then the information overload problem can be reduced. However, a number of potential barriers exist preventing the effective sharing of knowledge within an organisation and decreasing an employee's access to relevant information. As Cross et. al. (2001) state "By taking a look at the aspects of relationships underlying effective knowledge flow, we can offer more precise ways to improve a network's ability to create and share knowledge without overloading employees with yet more meetings or e-mail".

This chapter presents a method that enables an organisation to determine the extent to which knowledge sharing barriers impact upon an organisation. The literature review identified a number of potential barriers to knowledge sharing. However, the literature review did not find a method to determine the extent that knowledge sharing barriers affect a particular organisation or the knock on effect on information overload.

Using the barriers identified by the literature review as a basis, a questionnaire was developed that could be used to determine which of the potential barriers are present in an organisation. Following the questionnaire development a 'traffic light' system was also developed to allow organisations to quickly identify the barriers affecting their organisation and where it must focus its attention with regards to knowledge sharing. With the barriers to knowledge sharing identified the organisation can then aim to improve its knowledge sharing and in turn increase the relevance of information to employees. Following the development approval and piloting of the questionnaire it was then deployed within the two case-study organisations.



The questionnaire was deployed at both Pharmaco and SoftwareCo. The results from the first organisation, PharmaCo, followed by the results from the second organisation, SoftwareCo are discussed and the chapter concludes with a comparison of the differences between the two organisations and potential solutions for these organisations specifically drawn from the combination of results and literature.

This chapter satisfies objective 2 – “To establish through the use of a questionnaire the extent that multi-faceted barriers hinder information and knowledge sharing.”

#### **4.2 The Need to Assess Knowledge Sharing Barriers**

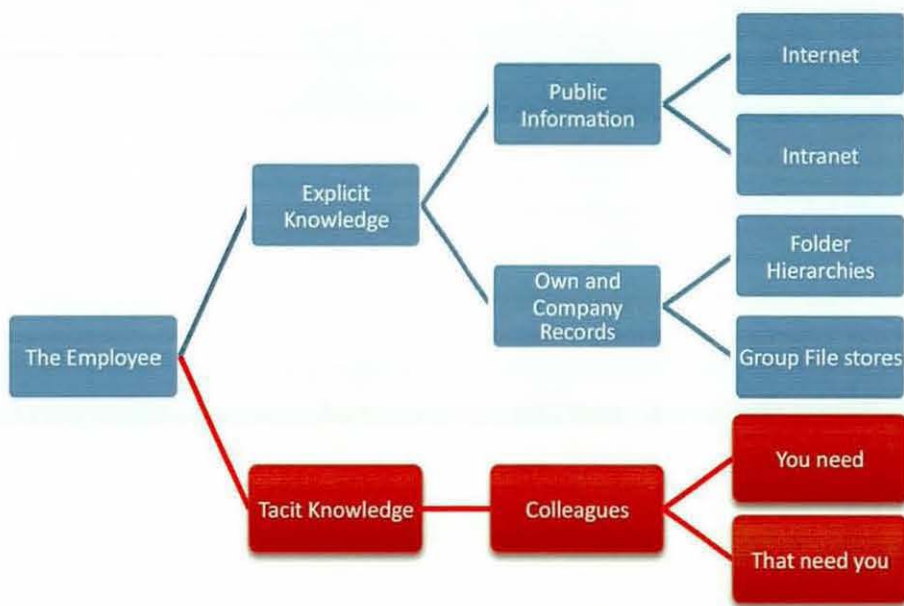
Literature has indicated that people searching for information would rather consult other people than use on or offline manuals (Tedmori et al. 2006b). Allen (1978) found that engineers and scientists were close to five times more likely to consult individuals rather than impersonal sources such as databases or file-cabinets for information. Even with advancements in information retrieval, computing and communications, the tendency to use people still exists. People remain the most valued and used source for knowledge (Tedmori et al. 2006a).

In knowledge intensive businesses, facilitating the sharing of knowledge between and organisation’s employees has been attributed to improving competitive advantage (Nahapiet, Ghoshal 2005). More purposeful and improved sharing of knowledge will bring with it increased learning and innovation at the individual, group and organisational levels. This increased learning and innovation will in turn bring about the development of better products and ideas that can be brought to market both more effectively and more efficiently (Riege 2005).

Increased sharing of knowledge, through the medium of information will make more relevant information available both within direct communications and through any potential system used by employees. The quality of information across the organisation will be higher and employees will be free to share information without the fear of how it might affect them.

The literature determined that, although it is difficult to measure the success of knowledge sharing (Riege 2005), it is possible to determine the barriers that exist

within an organisation. In addition, the literature review showed that the barriers towards knowledge sharing would have a unique impact upon each organisation (Argote, Ingram 2000). Although the literature review identified that each barrier may affect each organisation differently, and that in many cases some of the barriers may not be present at all, the literature did not identify a way to measure these barriers. In addition the literature also highlighted methods to combat the knowledge sharing barriers, but did not provide a method to determine which of the barriers are present. Therefore it is difficult for organisations to determine where effort should be focused to breakdown the existing barriers.



**Figure 10 - Information Sources - Colleagues**

As part of this research a tool was developed to assess the extent of the knowledge sharing barriers within organisations focusing upon the colleagues section of the diagram presented in Figure 10. The tool consisted of a questionnaire and a method to present the results to provide managers with a quick and easy way of determining the barriers that require immediate attention, that might need attention in the future, and finally, what areas are working well.

### 4.3 A Diagnostic Tool – Assessing the Knowledge Sharing Barriers

The barriers that would be investigated were based upon the barriers found in the literature review. The literature review identified a number of pieces of work

relating to the barriers to knowledge sharing. The literature review conducted by Riege (2005) combined many of the points contained within some of the most prominent pieces of literature to determine a comprehensive set of barriers in a structured format. These barriers were divided into a number of categories:

- Potential technological issues
- Potential individual issues
- Potential organisational issues

Each of the categories contained a number of potential barriers that may exist within an organisation. There were 36 barriers in total.

These potential barriers identified by Riege (2005) have been used to form part of the method to assess the barriers to sharing within an organisation. Riege's work provided the barriers, but did not provide a method to determine which barriers existed within an organisation and to the extent to which those barriers were present. The research detailed in this chapter shows how these barriers were taken and a framework added to them to enable a snapshot of the knowledge barriers to be determined at both PharmaCo and SoftwareCo.

#### **4.4 Capturing the Extent of the Barriers to Knowledge Sharing**

Given the large number of variables that exist and the other factors involved, it was decided that a questionnaire would be the most effective method to assess the knowledge-sharing environment within organisations. The questionnaire allows a snapshot of the feelings of participants at a particular time and allows a large number of participant's views to be collected that can be representative of the views across the department. The questionnaire method is especially useful as it allows a wide range of views to be collected in a relatively small time period compared to many other methods (Cornford, Smithson 1996). The questionnaire was created to establish how each of the determined barriers affects each organisation. The ability to distribute the questionnaires to a wide geographical area was vital as a number of employees were based at different locations throughout the United Kingdom in the first organisation and the second

organisation was based outside the UK. Using a questionnaire meant that it was possible to retrieve the data without having to physically be in the organisation to gather the responses.

#### 4.4.1 Categories of Questions

The three broad category titles identified by Reige (2005) were changed to aid the acceptance of the questionnaire.

1. Potential technological issues
2. Potential individual issues
3. Potential organisational issues

Negative connotations could be associated with “individual” and “issues” and were therefore changed to the following category titles. Questions that were originally constructed to fit under the original headings were then distributed throughout the new headings.

- Technology – that contained potential technological issues
- Organisational Factors – that predominantly contained potential individual but also contained potential organisational issues
- Organisational Sharing – that predominantly contained potential organisational but also contained potential individual issues

These categories formed the basis of the questionnaire and were further expanded to contain additional categories. Further details of the categories shall be given in this section.

PharmaCo decided that they would also like to use the survey to investigate how users interacted with technology during their daily processes. The organisation wished to know if any information retrieval systems that they were creating could be integrated into the current tools and processes of their employees. They felt that this would be important to prevent any disruption to the current practice of employees when a new system was introduced. For this reason the questionnaire included an additional section that would enable an organisation to determine the

tools and systems that were currently used by employees and their use of things such as the company portal. This section would be called 'daily processes'

Once these sections had been added the questionnaire contained the following categories.

#### **4.4.1.1 Participant Demographics**

This section asks questions about the age of the employee, level within the organisation and their length of employment along with other company specific questions.

#### **4.4.1.2 Technology**

Within the technology section preliminary questions were asked to give an overview of the level of competence the users had with technology in general. Following this, the technology section is predominantly derived from the potential technological barriers identified by Riege (2005).

#### **4.4.1.3 Organisational Factors**

The organisational factors and organisational sharing sections contained questions that were mostly derived from Riege's potential individual and potential organisational factors. These were split into two groups to prevent employees from feeling questions were aimed directly at barriers caused by them, as a section containing just individual barriers might cause concern and guarded answering.

The organisational factors section focused predominantly on how the organisation enables the individual to share knowledge. It also elicits the benefits they have seen and the issues with discovering employees with the necessary knowledge within the organisation.

#### **4.4.1.4 Daily Routine**

As mentioned earlier, the daily routine section was added at the request of PharmaCo in order to understand the daily tools and technology in use within the organisation. The section does not form part of the knowledge sharing barrier assessment.

#### **4.4.1.5 Organisational Sharing**

The organisational sharing section mainly focused on facilities provided and how knowledge was shared between teams and within the organisation as a whole. It was not restricted to only team based work.

#### **4.4.1.6 Rewards and Recognition**

Finally the rewards and recognition section was created to determine how users perceived being rewarded for knowledge sharing and to determine if they were aware of existing reward systems.

Although this is contained within Riege's barriers to some degree it was felt that it was important enough to warrant its own section. The first organisation requested this to be separate section in order to highlight that programmes existed even if participants were not aware of them.

#### **4.4.2 Developing the questions**

The questions within the questionnaire were predominantly based around the barriers discovered during the literature review. Each of the barriers was investigated and questions created to try to understand the feelings of employees with regards to these potential barriers. Some of the barriers would be difficult to assess and in many cases the way that things were phrased within the literature review prevented these being used directly as questions to the employees. Many of the barriers were addressed using a number of questions rather than just one in an attempt to discover more information about that subject or to take a less direct approach towards eliciting the information.

#### **4.4.3 Questionnaire Format**

The questions were mostly multiple-choice, based upon a four-point or two-point scale. Questions upon a two-point scale predominantly featured questions with a 'yes' or a 'no' answer. The even scales were chosen to prevent users from using an average 'middle' answer, often referred to as the central tendency error.

The management of PharmaCo felt that as an entire department was being asked to complete the survey the motivation of those participants may not be that high. For that reason they asked that the questions did not contain a central option.

Removing the middle option would not prevent acquiescence, a tendency to either consistently agree or disagree with a set of questions (Bryman, Bell 2003), however the organisation requested it was removed. In order to ensure consistency between the results of PharmaCo and SoftwareCo the same options were presented to both organisations.

A number of the multiple-choice questions were followed by a free text question asking the user to explain their answer or give reasons why they had answered the question in the way they did. The open questions provided an excellent insight into the views of those questioned and enabled them to expand on the answers that gave. This enabled a deeper view into their thoughts and led to less interpretation being required to alleviate some of the issues mentioned previously relating to surveys introducing bias.

The questionnaire contained 54 questions plus a number of questions at the beginning relating to participant demographics (Five for PharmaCo and Two for SoftwareCo) such as the length of time employees had been employed by the organisation and their level within the organisation.

#### **4.4.4 The Questions Designed and the Relation to Riege's Barriers**

Table 10, Table 11 and Table 12 show the barriers suggested by Riege (2005) and the questions developed as part of this research for the questionnaire used to analyse the barriers. Not all of the barriers identified by Riege were addressed to the same extent. In some cases the questions listed can only be used to help infer the answers rather than answer the question directly. The daily process section of the questionnaire does not appear in the tables, as none of the questions within that section relate to any of the potential barriers identified by Riege. Table 10, Table 11 and Table 12 each contain three columns. The first shows the barrier identified by Riege. The second column highlights which of Riege's three categories the potential barrier belongs to and the potential barrier number used to label that barrier in Riege's literature review (Riege 2005). The final column shows the questions that were constructed for the questionnaire to assess the knowledge sharing barriers. A full list of the questions contained within the questionnaire can

be found in chapter 10 within the Appendix. Abbreviations are used for the section titles in order to save space within the table.

#### 4.4.4.1 Potential Technological Barriers

Within Table 10 the word technology shall be abbreviated to Tech.

**Table 10 - Potential Technological Barriers**

| Barriers identified by Riege                           | Riege      | Question No.     |
|--|------------|------------------|
| Lack of system/process integration                     | Tech. 1, 4 | Tech. 7          |
| Lack of technical support                              | Tech. 2    | Tech. 8, 9       |
| Mismatch between individual's requirements and systems | Tech. 5    | Tech. 8, 10, 10a |
| Reluctance to use systems due to lack of familiarity   | Tech. 6    | Tech. 3, 5       |
| Lack of training                                       | Tech. 7    | Tech. 4, 5       |
| Lack of communication/demonstration of advantages      | Tech. 8    | Tech. 6          |

#### 4.4.4.2 Potential Individual Barriers

The following abbreviations shall be used within Table 11:

- The abbreviation "indiv." shall be used for Reige's potential individual barriers;
- Organisational Factors shall be abbreviated to "Org. Fac.";
- Organisational Sharing shall be abbreviated to "Org. Share." and
- "Rew. Rec." will be used in place of rewards and recognition.

**Table 11 - Potential Individual Barriers**

| Barriers identified by Riege  | Riege     | Question No.                   |
|---|-----------|--------------------------------|
| General lack of time to share knowledge, and time to identify colleagues in need of specific knowledge.                                 | Indiv. 1  | Org. Fac. 2-5<br>Org. Share. 9 |
| Apprehension of fear that sharing may reduce or jeopardise people's job security.   | Indiv. 2  | Org. Fac. 1a, 1b, 9            |
| Low awareness and realisation of the value and benefit of possessed knowledge to others.  | Indiv. 3  | Org. Fac. 7, 9                 |
| Use of strong hierarchy, position-based status, and formal power ("pull rank").   | Indiv. 5  | Rew. Rec. 3                    |
| Lack of contact time and interaction between knowledge sources and recipients.  | Indiv. 8  | Org. Fac. 2, 4, 8, 9           |
| Age differences.  | Indiv. 10 | Basic Details 1                |
| Lack of social network. (Used also to assess opportunities and places to interact)  | Indiv. 12 | Org. Share. 8, 9               |
| Taking ownership of intellectual property due to fear of not receiving just recognition and accreditation from managers and colleagues. | Indiv. 14 | Org. Fac. 1                    |



Riege's (2005) potential barrier number 15, "Lack of trust in people because they may misuse knowledge or take unjust credit for it." is assessed to a limited extent. It is difficult to obtain and understand whether there is a lack of trust in people. However, the concept of people receiving unjust credit or a fear or sharing because of this was assessed.

#### 4.4.4.3 Potential Organisational Barriers

The following abbreviations shall be used within Table 12:

- The abbreviation "Organ." shall be used for Reige's potential organisational barriers;
- Organisational Factors shall be abbreviated to "Org. Fac.";
- Organisational Sharing shall be abbreviated to "Org. Share." and
- Rewards and recognition shall be abbreviated to "Rew. Rec.".

**Table 12 - Potential Organisational Barriers**

| Barriers identified by Riege  | Riege     | Question No.                       |
|---|-----------|------------------------------------|
| Integration of km strategy and sharing initiatives into the company's goals and strategic approach is missing or unclear.             | Organ. 1  | Org. Share. 4, 5<br>Rewards also   |
| Lack of leadership and managerial direction in terms of clearly communicating the benefits and values of knowledge sharing practices. | Organ. 2  | Org. Share. 4, 5                   |
| Shortage of formal and informal spaces to share, reflect and generate (new) knowledge.  | Organ. 3  | Org. Fac. 6, 8<br>Org Share. 9     |
| Lack of a transparent rewards and recognition systems that would motivate people to share more of their knowledge.                    | Organ. 4  | Rew. Rec. 1, 4<br>Org. Fac. 3      |
| Existing corporate culture does not provide sufficient support for sharing practices.   | Organ. 5  | Org. Fac. 3<br>Org. Share. 7, 8, 9 |
| Deficiency of company resources that would provide adequate sharing opportunities.  | Organ. 8  | Org. Fac. 3, 6, 8                  |
| Communication and knowledge flows are restricted into certain directions (e.g.top-down).  | Organ. 10 | Org Share. 6<br>Rew. Rec. 3        |
| Internal competitiveness within business units, functional areas, and subsidiaries can be high.                                       | Organ. 12 | Rew. Rec. 2                        |
| Hierarchical organisation structure inhibits or slows down most sharing practices.  | Organ. 13 | Rew Rec. 3                         |

There were also two potential barriers, identified by Riege (2005), "Shortage of appropriate infrastructure supporting sharing practices" and "Physical work environment and layout of work areas restrict effective sharing practices", that

were assessed to a limited extent within the questionnaire due to the difficulty that would be found assessing these factors and the open-ended nature of the barriers.

#### **4.5 Questionnaire Deployment**

The survey was deployed as an online questionnaire. The first and foremost reason for this was that anonymity was required. The questionnaire asks a number of questions that may be considered quite sensitive and their answers may be perceived by the employees as a threat to their job security should they offend anyone. It was important that highly representative and unbiased answers were received. Secondly, recipients answered the questionnaire within their normal working environments to minimise disruption and to maximise the responses.

The questionnaire itself used the lime survey system. The system allows easy development of the questionnaire through the web-based interface. Since the lime survey project was open source the organisations were free to install the survey software on internal servers to increase the perceived security and privacy protection of the questionnaire.

The survey was deployed to two quite different companies. Although both were large multinationals the companies were both considerably different in terms of their organisation. One of the companies was a pharmaceutical and one a software company. Whilst it may have been interesting to compare two pharmaceuticals or two software companies, competition or rivalry would have caused issues. If another company in a similar industry were to be compared, the organisations would not have allowed their results to be shared with that company due to fear of one gaining competitive advantage from the other. The two companies chosen had no reason to feel in competition with each other allowing results to be shared and compared.

Before deployment of the questionnaire it was piloted by a number of Loughborough University staff and Postgraduate students. The participants acted as 'outside' participants, who completed the questionnaire in order to ensure that no problems were encountered. Before the questionnaire was deployed it had to be approved by members of the Human Resources department and Information

Services departments of PharmaCo. These members also acted as pilot respondents, filling in the questionnaires themselves. The questionnaire was also further piloted within the two organisations. The employees within the organisation that were helping to deploy the questionnaire first completed the questionnaire in order to assess how long the questionnaire would take and ensure potential problems were minimised. Bryman and Bell (2003) state that members of the pilot should not be part of the sample that would eventually be employed in the full study. It was therefore important that the testing took place with participants that would not be a part of the final study. However, in addition to the pilot study, both organisations deployed the questionnaire in two stages. This allowed any further alterations to be made before the questionnaire had been deployed to all of the sample population. Due to the extensive piloting of the questionnaire, no alterations were requested after the first phase of delivery.

#### **4.5.1 The organisations**

The questionnaire was deployed within two organisations, PharmaCo and SoftwareCo. Firstly it was deployed at PharmaCo and then within SoftwareCo several months later.

##### **4.5.1.1 Organisation One - PharmaCo**

Organisation one, PharmaCo, is one of the world's leading multinational pharmaceutical companies. They are active in over 100 countries with over 10,000 employees and have research and development sites around the world. The key focus of the company is in the research and development of new pharmaceuticals. The company takes the drug development process from start to finish from initial concepts to marketable medicines including clinical trials. The organisation prides itself on not just performing in its core research, but also creating a culture and environment in which people are valued and rewarded for their ideas and contributions. It is for this reason that this research was so well received by the organisation.

The participants of the questionnaire were from one department within the organisation. The department that participated in the questionnaire was involved in the testing of pharmaceuticals in humans, a highly research-intensive task. The

department was identified because of the managements desire to increase the knowledge sharing process presenting a case where employees could be asked to partake in the research by management. It was also felt by the management that this sample could generalise well to a larger section of the organisation making the research beneficial to the organisation as well as the researcher. Although all the participants belonged to the department, the department is geographically dispersed across two sites within the UK and has regular contact with teams around the world. The department contained close to one hundred employees and all were asked to participate in the survey. Participants were invited to participate in the survey via email from the key manager who runs the department.

Of the roughly one hundred employees asked to participate in the survey 60 participants responded. In PharmaCo a high response rate was expected and participants were asked to answer every question within the survey. The survey was placed onto a server within the organisation that would be available at any time of the day for a period of two weeks. It was placed within the organisation to alleviate any fears or concerns from high-level management as to the security of the information being recorded.

#### **4.5.1.2 Organisation Two - SoftwareCo**

Organisation two, SoftwareCo, which is again a fictitious name, is one of the largest software organisations in the world. SoftwareCo is within the top 10 of all of the major software rankings including the Forbes2000 and Research Foundation's top 100 with most indexes ranking the organisation in the top 5. The organisation employs over 50,000 people in over 50 countries. The company develops a range of software solutions in house and its products are used globally.

The organisation has a number of research and development departments along with rapid development and value prototyping. It is home to some of the best employees within their respective field and has a very unique structure. The department that was studied has attempted to create an environment specifically suited to rapid application development, testing and deployment. The department aims to have only a limited hierarchical structure with all members of the

department seen as equals and interacting with each other to take advantage of their respective skills

The employees were geographically dispersed around the world and therefore the online questionnaire was again well suited to collecting the information. The entire department was invited to take part in the questionnaire via email by someone known within the organisation as being a champion for knowledge sharing and information systems.

The department hosts close to 200 members and this number is growing rapidly. With this rapid growth the challenge of ensuring that employees continue to share knowledge also appears to be rising. Not all of the employees are full time and some members of the department work with the department infrequently so a high response rate was not expected. Of those that do frequently work with the department, a number are contractors rather than direct employees of the department.

It was decided by members of the organisation involved in the development of the survey that due to the nature of the department and how busy its employees are, participants should not be forced to answer every question as they were in the first organisation. The organisation felt that the survey would be better accepted and would receive a better response rate if users were free to opt out of answering a question at their discretion. The organisation also wanted to make the survey as "easy to swallow" as possible for employees as they wished to perform further surveys in the future. In order to help encourage employees, the department also entered all those employees who reached the end of the questionnaire into a prize draw to win a book token.

The survey software is capable of knowing which users got to the end of the questionnaire and which did not. Of those invited to participate there were 40 users who reached the end of the questionnaire and 76 responses in total. Since the users could miss any question they wished to it was decided to simply combine all of the answers leaving a maximum of 76 responses for each question, however some questions received a significantly lower response rate, for example the last question in the questionnaire only received 29 responses. Within this section any

percentages given will be the valid percentages, that is, the percentages taken only from those who answered the question. The organisation felt that although 29 responses for some questions was quite a low figure they were still interested in these results. The sections within the questionnaire also operated as standalone units and although the last section may not have had a great response rate, the earlier sections have a much higher response rate and thus a greater statistical value.

As already mentioned previously, not everybody within the department worked full time for the organisation and participants also included consultants who were working for the organisation. In this organisation consultants are seen as just as much a part of the permanent staff as any other member. Over 50% of the staff who responded to the questionnaire were consultants, many of which have been employed for a large period of time at the organisation. When asked what percentage of their time did the participants work at the organisation, 72% said that they worked for the organisation for 90% or more of their time. Although not all of the participants of the survey are technically employees of the company, they shall be referred to as employees.

The questions were kept as similar to those used by PharmaCo as possible. Almost all of the participants of the second survey were not native English speakers and this could have had an impact on the answers, but it was unlikely as the company's official language is English and most employees understand, speak and write English extremely well. However, some questions were re-worded slightly to reduce the possibility of misunderstanding for non-native English readers at the request and with the assistance of one of the employees of SoftwareCo.

#### ***4.5.1.3 Why these organisations were chosen***

The two organisations were chosen for a number of reasons. The size and type of the organisation were considered when selecting the organisation. Two different types of organisation were chosen so the results could be compared. If they worked in the same industry it is unlikely that the results could have been shared. The literature has shown that there is likely to be more barriers to sharing information and knowledge in international organisations with a large number of

personnel than that of a small to medium enterprise. The two organisations were both international and both had large number of employees. Both organisations are knowledge intensive which also fits the remit of this research. Finally, the organisations have existing relationships with Loughborough University in the area of information overload and knowledge sharing, and although they meet the selection criteria they were also opportunistic selections.

#### **4.6 Results**

The results are presented under the same headings as the questionnaire previously discussed in section 4.4.1. The sections are Technology, Organisational Factors, Daily Routine Organisational Sharing and Rewards and Recognition.

The results tables do not contain the responses given to free text questions for the sake of brevity, although many of these responses are included in a written results section found in the Appendix in Chapter 11. The daily routine section is also not included here, as section 4.4.1 stated these questions were only included for the benefit of the organisations and do not relate to the potential barriers identified.

**Table 13 - Technology Barriers experienced by PharmaCo and SoftwareCo**

| <b>Question</b>   | <b>Answers</b>            | <b>PharmaCo</b> | <b>SoftwareCo</b> |
|---|---------------------------|-----------------|-------------------|
| How would you rate yourself as a computer user  | Expert                    | 12%             |                   |
|   | Experienced               | 66%             |                   |
|   | Some Experience           | 20%             |                   |
|   | Novice                    | 2%              |                   |
| How would you rate yourself with regards to using technology in general (e.g. a video recorder) | Expert                    | 11%             |                   |
|   | Experienced               | 52%             |                   |
|   | Some Experience           | 35%             |                   |
|   | Novice                    | 2%              |                   |
| When you are given a new piece of technology do you   | Look forward to using it  | 49%             | 74%               |
|   | Use it only when required | 49%             | 22%               |
|   | Become apprehensive       | 2%              | 4%                |
| Adequacy of training  | 10%                       | 5%              | 5%                |
|   | 25%                       | 9%              | 14%               |
|   | 50%                       | 29%             | 41%               |
|   | 75%                       | 45%             | 30%               |
|   | 90+%                      | 12%             | 11%               |
| Is sufficient training given when a new system is introduced                                    | Always                    | 2%              | 9%                |
|   | Often                     | 48%             | 21%               |
|   | Sometimes                 | 42%             | 53%               |
|   | Rarely                    | 9%              | 17%               |
| Do you feel the benefits of a new system over the old are clearly explained                     | Always                    | 1%              | 8%                |
|   | Often                     | 26%             | 29%               |
|   | Sometimes                 | 6%              | 51%               |
|   | Rarely                    | 66%             | 12%               |
| Do you think that current IS tools and business processes are well integrated                   | Always                    | 0%              | 9%                |
|   | Often                     | 19%             | 43%               |
|   | Sometimes                 | 20%             | 24%               |
|   | Rarely                    | 62%             | 24%               |
| Are you given sufficient opportunity to give feedback on the suitability of IS provided         | Yes                       | 32%             | 71%               |
|   | No                        | 68%             | 29%               |
| Is there sufficient technical support available for the applications you use                    | Yes                       | 63%             | 91%               |
|   | No                        | 37%             | 9%                |
| Do newly implemented systems live up to your expectations                                       | Yes                       | 48%             | 76%               |
|   | No                        | 52%             | 24%               |
| Do you suffer from the lack of compatibility between IT systems                                 | Yes                       | 29%             | 32%               |
|   | No                        | 71%             | 68%               |



**Table 14 - Organisational Factors experienced by PharmaCo and SoftwareCo**

| Question   | Answers        | PharmaCo | SoftwareCo |
|--|----------------|----------|------------|
| Do you feel you receive sufficient credit when sharing knowledge                                     | Always         | 3%       | 15%        |
|  | Often          | 24%      | 28%        |
|  | Sometimes      | 47%      | 31%        |
|  | Rarely         | 26%      | 26%        |
| If Rarely or Sometimes does this make you reluctant to share knowledge in future                     | Yes            | 15%      | 33%        |
|  | No             | 85%      | 67%        |
| Are you given enough time to share knowledge   | Always (Yes)   | 3%       | 53%        |
|  | Often          | 24%      | 0%         |
|  | Sometimes (No) | 13%      | 47%        |
|  | Rarely         | 60%      | 0%         |
| Do you feel you can record 'Knowledge Sharing' in your timesheets                                    | Yes            | 77%      | 33%        |
|  | No             | 23%      | 67%        |
| Are you given enough time to meet and identify colleagues that have the knowledge YOU SEEK           | Yes            | 46%      | 68%        |
|  | No             | 54%      | 32%        |
| Are you given enough opportunity to meet and identify colleagues with a need for YOUR knowledge      | Yes            | 39%      | 59%        |
|  | No             | 62%      | 41%        |
| Have you benefited through sharing knowledge with others (including receiving knowledge from others) | Always         | 25%      | 32%        |
|  | Often          | 51%      | 42%        |
|  | Sometimes      | 22%      | 18%        |
|  | Rarely         | 2%       | 8%         |
| Are there currently knowledge capture tools available within your organisation                       | Yes            | 52%      | 47%        |
|  | No             | 48%      | 53%        |

**Table 15 - Organisational Sharing experienced by PharmaCo and SoftwareCo**

| Question   | Answers   | PharmaCo | SoftwareCo |
|--|-----------|----------|------------|
| Do you share knowledge outside your team   | Yes       | 82%      | 79%        |
|  | No        | 19%      | 21%        |
| Has your company made its Knowledge Sharing goals clear  | Yes       | 48%      | 32%        |
|  | No        | 52%      | 68%        |
| How regularly are you encouraged to share knowledge by your management   | Always    | 14%      | 3%         |
|  | Often     | 38%      | 19%        |
|  | Sometimes | 33%      | 36%        |
|  | Rarely    | 14%      | 43%        |
| Is sharing knowledge outside your team or group part of your work process  | Yes       | 57%      | 50%        |
|  | No        | 43%      | 50%        |
| Do you find it easy to actually share knowledge  | Yes       | 65%      | 50%        |
|  | No        | 35%      | 50%        |
| Are there enough formal (e.g. within meetings) and informal (e.g. coffee rooms) places to share, generate and reflect on new knowledge | Yes       | 59%      | 63%        |
|  | No        | 42%      | 37%        |
| Do you feel you are given sufficient opportunity to interact with colleagues outside your immediate job, for example at conferences    | Yes       | 34%      | 46%        |
|  | No        | 66%      | 54%        |

**Table 16 - Rewards and Recognition experienced by PharmaCo and SoftwareCo**

| <b>Question</b>  | <b>Answers</b> | <b>PharmaCo</b> | <b>SoftwareCo</b> |
|--|----------------|-----------------|-------------------|
| Do you know of any reward schemes present to encourage the sharing of knowledge within your organisation   | Yes            | 12%             | 9%                |
|  | No             | 88%             | 91%               |
| If yes do you feel these schemes offer sufficient reward to encourage Knowledge Sharing  | Yes            | 91%             | 50%               |
|  | No             | 9%              | 50%               |
| Do you feel you are in competition with other people within your department  | Yes            | 26%             | 34%               |
|  | No             | 74%             | 66%               |
| Does your organisational reporting structure hinder Knowledge Sharing, for example knowledge is only shared between yourself and your manager            | Yes            | 22%             | 14%               |
|  | No             | 79%             | 86%               |
| If Knowledge Management and Sharing were included within a yearly review process would you spend more time developing your skills in 'Knowledge Sharing' | Yes            | 48%             | 86%               |
|  | No             | 52%             | 14%               |

#### 4.6.1 Summarising the results and comparing the organisations

In order to manage the problem of information overload within an organisation it is important to identify the areas that must be focused upon. It was determined through discussions with the organisations that a method to assess how strongly each of the potential knowledge sharing barriers identified impacted the organisation would need to be coupled with a method to quickly highlight the barriers and summarise the information gained from the questionnaire.

In order to effectively summarise and compare the questionnaire results from the two organisations a method to summarise the results of the questionnaire was devised. The summary is presented as a table, showing the potential barriers that were identified by this questionnaire and the extent to which the barrier was present within the organisation as determined by the questionnaire results. The *summary method then uses traffic light colours, red, amber and green to provide a fast and simple summary for that question.* This method also allows more than one organisation to be placed into the table for comparisons to be made.

The questions from the questionnaire were interpreted to understand where knowledge sharing issues lie within the organisations, as shown by Table 17, Table 18 and Table 19. Many of the questions may be beneficial in isolation. However,

the answers to the questions may be used holistically, to gain an understanding the barriers proposed by Riege (2005). The answers to the questions were discussed with key knowledge champions within the organisation so that the areas that should be addressed could be identified. The tables make use of a traffic light system in order to highlight issues. The lights used are as follows:

- Green is seen as a small issue or no issue at all, something that does not need addressing at present.
- Amber is seen as an issue but not one that requires urgent attention. This issue shall be investigated by the organisation in more detail and recommendations formed at a convenient time for the organisation.
- Red is reserved for the most critical issues. Issues highlighted in red should be addressed immediately and pose a problem for the organisation.

In many cases multiple questions were used to assess each barrier and the results of the questions were combined. A degree of interpretation was also required as there was no one-to-one mapping between the barriers and the questions. Some questions may have more of an influence on the results than others. As a guiding rule, in most cases a 66% positive response would receive a green light. Anything with more than 33% positive responses would receive an amber light and anything with less than 33% positive responses would receive a red light.

The traffic light system was interpretive and thus these percentages were only used as a guideline and the answers to the questions were combined. In many cases, the answers to several questions were combined and the free-text results were consulted in an interpretive way to determine the state of that barrier. As an example, if it were found that people did not feel like they received credit but then stated that it did not affect how they shared in future, this would receive an amber warning rather than a red one. In many cases the free text answers were also used to understand the feelings of participants and address the strength of the issues.

#### **4.6.2 Comparison Tables**

Contained within Table 17, Table 18 and Table 19 are a list of potential barriers that were assessed by the questionnaire and the relevant traffic light symbol highlighting the impact of the barrier within the organisation.

**Table 17 – Traffic lights for Potential Technological Barriers**

| Barriers Assessed by the Questionnaire                 | Riege      | PharmaCo | SoftwareCo |
|--|------------|----------|------------|
| Lack of system/process integration                     | Tech. 1, 4 |          |            |
| Lack of technical support                              | Tech. 2    |          |            |
| Mismatch between individual's requirements and systems | Tech. 5    |          |            |
| Reluctance to use systems due to lack of familiarity   | Tech. 6    |          |            |
| Lack of training                                       | Tech. 7    |          |            |
| Lack of communication/demonstration of advantages      | Tech. 8    |          |            |

**Table 18 – Traffic lights for Potential Individual Barriers**

| Barriers Assessed by the Questionnaire  | Riege     | PharmaCo | SoftwareCo |
|---|-----------|----------|------------|
| General lack of time to share knowledge, and time to identify colleagues in need of specific knowledge.                                 | Indiv. 1  |          |            |
| Apprehension of fear that sharing may reduce or jeopardise people's job security.   | Indiv. 2  |          |            |
| Low awareness and realisation of the value and benefit of possessed knowledge to others.  | Indiv. 3  |          |            |
| Use of strong hierarchy, position-based status, and formal power ("pull rank").   | Indiv. 5  |          |            |
| Lack of contact time and interaction between knowledge sources and recipients.  | Indiv. 8  |          |            |
| Age differences.  | Indiv. 10 |          |            |
| Lack of social network. (*Used also to assess opportunities and places to interact)   | Indiv. 12 |          |            |
| Taking ownership of intellectual property due to fear of not receiving just recognition and accreditation from managers and colleagues. | Indiv. 14 |          |            |

**Table 19 – Traffic lights for Potential Organisational Barriers**

| Barriers Assessed by the Questionnaire  | Riege     | PharmaCo | SoftwareCo |
|---|-----------|----------|------------|
| Integration of km strategy and sharing initiatives into the company's goals and strategic approach is missing or unclear.             | Organ. 1  |          |            |
| Lack of leadership and managerial direction in terms of clearly communicating the benefits and values of knowledge sharing practices. | Organ. 2  |          |            |
| Shortage of formal and informal spaces to share, reflect and generate (new) knowledge.  | Organ. 3  |          |            |
| Lack of a transparent rewards and recognition systems that would motivate people to share more of their knowledge.                    | Organ. 4  |          |            |
| Existing corporate culture does not provide sufficient support for sharing practices.   | Organ. 5  |          |            |
| Deficiency of company resources that would provide adequate sharing opportunities.  | Organ. 8  |          |            |
| Communication and knowledge flows are restricted into certain directions (e.g. top-down).   | Organ. 10 |          |            |
| Internal competitiveness within business units, functional areas, and subsidiaries can be high.                                       | Organ. 12 |          |            |
| Hierarchical organisation structure inhibits or slows down most sharing practices.  | Organ. 13 |          |            |

#### 4.6.3 Comparison Discussion

Table 17, Table 18 and Table 19 highlight a number of differences between the two organisations.

Table 17 shows the potential technical barriers identified by the literature and the extent to which the questionnaire responses determined that those technical barriers impacted each of the organisations. The barriers in this section relate to the interaction between technology and humans in order to facilitate knowledge sharing. The results indicate that technology can act as a facilitator to encourage and support knowledge sharing however it is important to ensure that the fit between technology and humans is correct in order to ensure successful sharing (Riege 2007).

As could be expected SoftwareCo appears to suffer less from technological barriers. However, one key finding is that employees do not feel training is sufficient in both organisations. The issue does appear worse within SoftwareCo. It may be felt that because the respondents from SoftwareCo are seen as extremely

technical managers they do not feel the need for so much training. Increased training would appear to be beneficial however.

The employees of SoftwareCo were more computer literate than employees of PharmaCo and this may also help to explain the higher mismatch between the requirements of the individuals and the systems delivered within PharmaCo. As an example a number of the SoftwareCo employees felt that one of the document management systems used by the organisation did not meet their requirements. This is one example that might help to explain why differences between organisations occur. Often more technically astute employees will be more aware of the potential issues that may arise during the development of a system and be more sympathetic towards any difficulties encountered. In addition the employees of SoftwareCo may be able to give more targeted feedback being developers themselves helping the organisation to develop and implement tools that better suit their needs. The inability to leave feedback and receive technical support may also have influenced this, as employees of SoftwareCo felt that they could leave sufficient feedback. The benefits of any new system should also be explained and demonstrated.

Table 18 highlights the individual barriers found within both organisations. As Riege states "Just about every book written on KM comments on the distribution of the right knowledge from the right people to the right people at the right time being one of the biggest challenges in knowledge sharing" (Riege 2007). The barriers that originate from individuals form a crucial area that an organisation should address as it aims to reduce its barriers to knowledge sharing. Table 18 highlights a number of differences between the organisations, although there were some similarities with regards to individual barriers. Although both organisations suffered from a lack of time to share knowledge the problem was more severe within PharmaCo. In addition, although both organisations suffered from a lack of time to interact with colleagues that they could gain knowledge from and could impart knowledge to the problem was again more severe within PharmaCo. The lack of social network and opportunities for interaction available within both organisations also appeared to cause potential problems.

The final comparison, Table 19 highlights the organisational barriers identified by the questionnaire for the two organisations. Organisational barriers relate to the environment and culture provided by the organisation to allow effective knowledge sharing. Again the two organisations show some similarities and differences when it comes to the organisational barriers present. In particular the culture within SoftwareCo appears to be far less supportive of sharing practices in the eyes of its employees. This may be related to the feeling that the overall knowledge management strategy and goals are unclear to the employees and that there is limited managerial direction with regards to communicating the benefits of knowledge sharing.

Overall the tables have highlighted that the organisations have significant differences when it comes to these barriers, as the literature suggested. Once the barriers are identified through the use of these tables it is then possible for the organisation to determine where its effort should be placed with regards to increasing its knowledge sharing and reducing the apparent barriers.

The following section examines the barriers identified by the questionnaire and traffic light system and then examines possible solutions, taken from the literature for these two case-study organisations. This process should then be repeatable for other organisations wishing to reduce their knowledge sharing barriers.

#### **4.6.4 The final step – example recommendations for the reduction of the barriers**

Given the barriers to knowledge sharing that have been identified by the questionnaire method it is possible to make recommendations to allow organisations to combat these issues and aim to increase knowledge sharing and the flow of relevant information. The traffic light system has helped to identify the barriers that existed within both organisations. Each organisation appeared to have five key barriers. The results have shown that both organisations have different areas in which they need to improve. In terms of the non-technical department in PharmaCo, the key barriers identified are:

- Mismatch between individuals' need requirements and integrated IT.

- Lack of communication, and demonstration of all advantages of any new systems over existing ones.
- General lack of time to share knowledge, and time to identify colleagues in need of specific knowledge.
- Lack of contact time, and interaction between knowledge sources and recipients.
- Lack of a transparent rewards and recognition systems that would motivate people to share more of their knowledge.

Within SoftwareCo two of the key issues were the same, however there were three other issues that differed in this organisation. The five key areas of concept within SoftwareCo where:

- Lack of training regarding employee familiarisation of new IT systems and processes.
- Lack of communication, and demonstration of all advantages of any new systems over existing ones.
- Integration of KM strategy and sharing initiatives into the company's goals, and strategic approach is missing or unclear.
- Existing corporate culture does not provide sufficient support for sharing practices.
- Lack of a transparent rewards and recognition systems that would motivate people to share more of their knowledge.

The direction from higher-level management appeared to be an area where PharmaCo appeared better than SoftwareCo. Although SoftwareCo's employees felt free to share knowledge and information at all levels, not just to and from their direct managers, SoftwareCo did not appear to have communicated its knowledge sharing goals and strategic approach to the employees. Although this issue existed within PharmaCo it was not as great. It also appeared that more of the provisions for knowledge sharing were given by PharmaCo as there was more opportunities



to reflect and generate new knowledge and more support for sharing practices. Both organisations lacked sufficient rewards and recognition systems, but the possibilities to improve knowledge sharing practice may lie within rewards.

#### **4.6.4.1 *Overcoming the Barriers Identified in PharmaCo and SoftwareCo***

Once the questionnaire had been used to determine which barriers existed within each of the organisations and the summary table was used to highlight the areas that the organisation needed to focus upon, it was important to provide recommendations to help the organisations overcome the barriers they faced rather than a generic set of barriers.

Overcoming these barriers will help promote more successful knowledge sharing within PharmaCo and SoftwareCo. Using the questionnaire and summary tables it showed that PharmaCo and SoftwareCo each have five potential barriers that impact upon their organisation. Although there were five barriers in each company, this is purely a coincidence.

Following the deployment of this questionnaire, Riege, the author who outlined the potential barriers to knowledge sharing, released a limited number of recommendations to tackle some of the potential barriers (Riege 2007). There are not solutions for all of the potential barriers identified previously. The recommendations given here shall draw from those taken from the literature review by Riege (Reige 2007), and also recommendations of the author, gained from a wider literature review. The recommendations shall be outlined below, along with the barrier that they tackle.

Both PharmaCo and SoftwareCo had two issues in common. Although both organisations suffer from two of the issues the recommendations shall be handled individually in this section. The following two sections shall look at the recommendations for each organisation.

#### **4.6.4.2 *PharmaCo***

The first barrier was a “Lack of communication, and demonstration of all advantages of any new systems over existing ones”. Although Riege (Reige 2007) did not provide a recommendation to this issue, there are a number of steps that

can be taken. The simplest solution would be to give regular updates. These updates could show employees the benefits of the system during its development and the new features that the system contains. Whilst new features are introduced they can be highlighted along with the downsides of the old system. This would allow a build up of anticipation within employees who could be excited to experience the benefits once the new system is available. Informing employees of new features could be done in a variety of ways, from holding regular meetings to sending out email updates.

It appears that this barrier is, however, intertwined with a second. The second barrier is a "Mismatch between individual's requirements and systems". If users are involved in the process of designing or choosing a system from start to finish, then they will not only be clear of the advantages to them but also requirements of the users will be met. Although it is not always possible to meet the requirements of every user, if they have been involved as a group then they can see the benefit to others also and appreciate that user driven change is occurring. Riege outlined a number of possible solutions to such a barrier (Riege 2007):

- "Focus primarily on people, not technology, i.e. look at who needs which tools to support and facilitate the way things are done".
- "Define technological challenges and opportunities to match them as closely as possible with existing resources".
- "Encourage people to provide feedback on content and usability, and acknowledge those who do".
- "Inform people of resolution or changes that have occurred based on their feedback and thank them for their assistance".

Presenting employees with the ability to provide feedback and showing that feedback is being acted upon will be of benefit. Regular updates like this shall also help communicate the advantages of such improvements, not just to the individual, but also demonstrate that improvements affect the entire department.

PharmaCo also experienced two of the barriers experienced by SoftwareCo. Those barriers were a "General lack of time to share knowledge, and time to identify

colleagues in need of specific knowledge” and a “Lack of contact time and interaction between knowledge sources and recipients”.

Riege (2007) offered a number of solutions to the problem of “General lack of time to share knowledge, and time to identify colleagues in need of specific knowledge.”. Being able to identify colleagues that have the information that an employee requires, allows the employee to quickly gain access to the relevant information. Although this approach could increase the channels of information available to employees, the key is to identify the right employee that holds the necessary information, without further adding to the information overload burden.

The suggestions given by Riege are as follows:

- “Acknowledge user time pressures and allocate purposeful ‘slack’ time for knowledge transfer, e.g. set aside one hour per week to facilitate sharing initiatives”.
- “During training/launch of KM initiatives, provide examples that illustrate how specific actions can save people time, and perform or prioritise certain tasks in the future more efficiently”.
- “Provide formal sharing settings, e.g. fairs, expert networks, communities of practice”.
- “Offer informal areas, e.g. coffee rooms, bars, gymnasiums, game rooms, where people can meet and connect socially, enhancing their sense of belonging to the firm and sharing opportunities”.
- “Offer work-related and social occasions to interact with stakeholders and customers to enhance cross-functional thinking and gain external knowledge”.
- “Gather and share ‘success stories’ about how time can be saved or wasted”.
- “Stress the importance of transferring tacit knowledge over explicit knowledge for individual and organisational learning”.

To the issue of "Lack of contact time and interaction between knowledge sources and recipients," Riege (2007) suggested the following solutions:

- "Create frequent formal and informal meeting areas, and opportunities to provide regular contact for people working closely together, or having a reason or need to share knowledge".
- "Support occasional face-to-face meetings before establishing new project teams, especially if team will be primarily in virtual mode".
- "Encourage external network opportunities between stakeholders and customers".
- "Create superior physical and electronic environments that support sharing initiatives".

In general, it appears that although PharmaCo already provide a number of formal and informal meeting spaces, employees lack the time to utilise these. Setting aside time to facilitate sharing initiatives is necessary.

#### **4.6.4.3 SoftwareCo**

SoftwareCo also experienced a "Lack of communication and demonstration of all advantages of any new systems over existing ones". This is especially interesting within a software company. Within SoftwareCo, employees felt that they did have enough opportunity to feedback on systems, and get technical support for those systems. In order to highlight the advantages, a number of steps could be taken. The first step is to include employees in the development of any system. Although employees are currently free to give feedback, they perhaps do not know why the system is implemented in the way that it is. Employees may also be unaware of the full list of features, or benefits, of the system. Involving the employees in the development means that they can see new features as they emerge and they can gain a fuller understanding of what the system does. Holding regular workshops where users may come and use the system and be introduced to the latest features would also be of benefit. In addition to this, regular emails can be sent out highlighting changes that have been made. These emails can also highlight

requests and how they are being satisfied to continue to demonstrate that the users drive features of the system.

A "Lack of training regarding employee familiarisation of new IT systems and processes" was also a key issue for SoftwareCo. Training is again an issue that Riege (2007) does not discuss. Again, however, the recommendation is a simple one. It is quite possible that, because SoftwareCo has a number of highly trained experts in the field of IT, training is not necessary to the degree that it is in other companies. Increasing the training that employees are given when they join the organisation and when new systems are introduced should have clear benefits within the organisation.

SoftwareCo appeared to have a limited impact caused by personal barriers. The organisation itself, however, appeared to have a number of barriers towards knowledge sharing. The "Integration of km strategy and sharing initiatives into the company's goals and strategic approach is missing or unclear" within SoftwareCo. Riege (2007) offered a number of suggestions to combat this barrier:

- "Clearly link any initiatives to company goals and strategic direction".
- "Demonstrate how initiatives can support company goals and strategies in a clear and transparent manner to all people, to obtain their ongoing support".
- "Market any initiatives, not as something that enhances your. or the firm's, own glory, but because an otherwise valuable intangible resource may go unused".
- "Explain how sharing practices can support people in the performance of their work, e.g. during training, have people access the system to address issues that they will face on the job, to illustrate how it can how it can help them".
- "Show how people can save time and work more efficiently through collaboration, thereby, e.g. benefiting individual learning, enhancing productivity, reducing mistakes".

The first step as outlined by Riege is to link the organisation's goals and strategic direction to knowledge sharing initiatives. In an organisation where knowledge sharing is so important, this is crucial.

Another related barrier faced by SoftwareCo was that "Existing corporate culture does not provide sufficient support for sharing practices". Riege (2007) again offered a number of solutions to this barrier:

- "Assess dimensions such as vision and mission, norms and customs, means to achieve goals, management processes, focus on external environment, image and reputation etc, that impact on your corporate culture".
- "Integrate sharing activities into existing corporate values and style of the company, rather than change your entire culture to suit sharing objectives (do one small step at a time as nobody likes change – do you?)".
- "Make your sharing culture a part of organisational policy and people's individual KPIs [Key Performance Indicators]".
- "Communicate knowledge policies clearly to all people, especially new ones, as part of the firm's training and development, and induction program".
- "Ensure individual and collective understanding of the purpose, value and benefits of knowledge sharing".
- "Implement any cultural changes to support sharing practices slowly, and communicate them clearly".
- "People who resent necessary changes need to adapt or leave".

It appears that the issue within SoftwareCo does not lie with the employees but with the organisation itself. The organisation needs to highlight the benefits of knowledge sharing and integrate sharing activities into the corporate culture. Employees should be given the opportunity to share knowledge and its importance should be stressed by the organisation. Higher-level managers should encourage sharing as should the organisation's overall goals and aims. Finally, in order to promote knowledge sharing activities and practice, knowledge sharing should be rewarded.

Both organisations, PharmaCo and SoftwareCo suffered from a “Lack of transparent rewards and recognition systems that would motivate people to share, reflect and generate (new) knowledge”. The first step for both organisations would logically be to place knowledge sharing into the organisation’s key performance indicators. Both organisations have regular reviews of staff performance. If employees are rewarded, in these reviews, for knowledge sharing then they are far more likely to do this. Employees were asked whether having knowledge sharing within their review process would encourage knowledge sharing. In PharmaCo almost 50% stated that they would improve knowledge sharing if it were included in their reviews. In SoftwareCo this was even higher with 87%. Rewarding employees for successful knowledge sharing does not have to be purely monetary. Receiving credit for being an above average ‘knowledge sharer,’ or being praised in some other form, can have benefits also. Finally Riege (2007) offered a number of suggestions to ensure that a rewards system works. The recommendations were based upon the fact that a rewards and recognition system was in place, but was not working. These recommendations can also help when establishing a new system of reward. The recommendations are as follows:

- “Keep your system simple and transparent, and use the same parameters for everyone”.
- “Weigh up intrinsic versus extrinsic rewards”.
- “Introduce an incentive system that ensures that all people contribute to what and to whom it matters”.
- “Communicate reasonable and accountable practices that motivate people to maximise purposeful sharing”.
- “Use rewards and recognition to encourage people to spend time, invest in their expertise and assume responsibility for using the system”.
- “Offer incentives to unite efforts that individuals cannot achieve by themselves”.
- “Openly trumpet successes and recognise individuals or units as contributors to the knowledge domain and convey ‘what’s in it for me’”.

- “Make sharing practices part of internal staff development and performance reviews”.
- “Consider building a job certification program, which includes transfer practices, or incorporate transfer use into an existing certification program”.
- “Ensure that any reward and recognition system promotes individual and organisational knowledge sharing, rather than individual knowing (which is still too common)”.
- “Ask yourself if any reward and recognition system creates any long-term benefits and adds to the firm’s performance”.

With an effective rewards and recognition system, employees are not only encouraged to share knowledge but are also shown that the management and organisation itself are committed to increasing knowledge sharing.

It seems that both organisations, PharmaCo and SoftwareCo could benefit from the recommendations given and that both suffer from different knowledge sharing barriers. Although there were similarities in some places, the barriers experienced by one organisation and thus the recommendations given differed. It is important that organisations assess how the barriers affect them, rather than simply assuming that all barriers have the same impact. The questionnaire and traffic light system can help organisations to assess this impact. With the impact of the potential barriers identified the organisation can determine where to focus their effort to reduce the impact of these barriers. With the barriers reduced, employees gain access to more relevant information and less irrelevant information. This increase in relevant information can help them to make more informed decisions and ultimately reduces the problem of information overload (Cross et. al. 2001).

#### **4.7 Conclusions**

Both the literature review chapter and this chapter have highlighted the benefits of knowledge sharing for organisations and the potential of knowledge sharing to reduce the problem of information overload. Improving the knowledge that is



shared within an organisation increases the availability of relevant information reducing their subjection to the irrelevant (Reige 2005).

Successful knowledge sharing can bring great rewards, and with only slight improvements, it can lead to an increase in learning and innovation. The facilitation of knowledge sharing between an organisation's employees has also been attributed to improved competitive advantage (Nahapiet, Ghoshal 2005). Although it is difficult to measure the success of knowledge sharing and identify how well an organisation shares knowledge, this research has shown it is possible to identify the barriers to knowledge sharing and reasons why knowledge sharing may fail. The literature provided a number of potential barriers to knowledge sharing but did not provide a method to determine which of the barriers and to what extent the barriers were present within an organisation.

The potential barriers identified within the literature review were used as the basis of a questionnaire developed by the author. The questionnaire enabled both organisations to determine the extent that each barrier affected their organisation.

Building upon the categories of barriers developed by Riege (2005), the questionnaire asked questions in five key categories. The five categories that were developed for the questionnaire were technology, organisational factors, daily routine, organisational sharing and rewards and recognition.

Following the deployment of the questionnaire it was important to allow a summary of the results to be available for higher level management to review. The author developed a traffic light table that could be used to summarise the results of the questionnaire and allow the organisation to determine the key areas of concern that required focus. The summary tables also highlighted the differences between the responses given by both organisations.

The approach taken in this research has provided a method for identifying knowledge barriers within organisations. The multifaceted approach has avoided the traditional route of just surveying IT systems for knowledge sharing. Through the implementation of a traffic light system, an organisation can quickly determine where they are doing well and the areas that need further work. The results have shown that both of the case study organisations have different areas that they need

to improve, which demonstrates the need to perform a multifaceted assessment of barriers as detailed within this chapter.

## 5 Alternative Search Visualisation – Concept Clouds

### 5.1 Chapter Preface

The previous chapter investigated the sharing of information between employees, and this chapter continues that theme by investigating the discovery of information. Although many people make use of interaction with colleagues to share information, the Internet represents a growing source of information. The literature review estimated that there were over 60 billion indexable pages on the Internet. A number of authors cite the Web as now being the primary source of information for many people (Cole, Suman, Schramm, Lunn, & Aquino, 2003; Fox, 2002). In addition, the rise of information overload has been attributed to technology and more specifically the Internet (Nelson 1994).

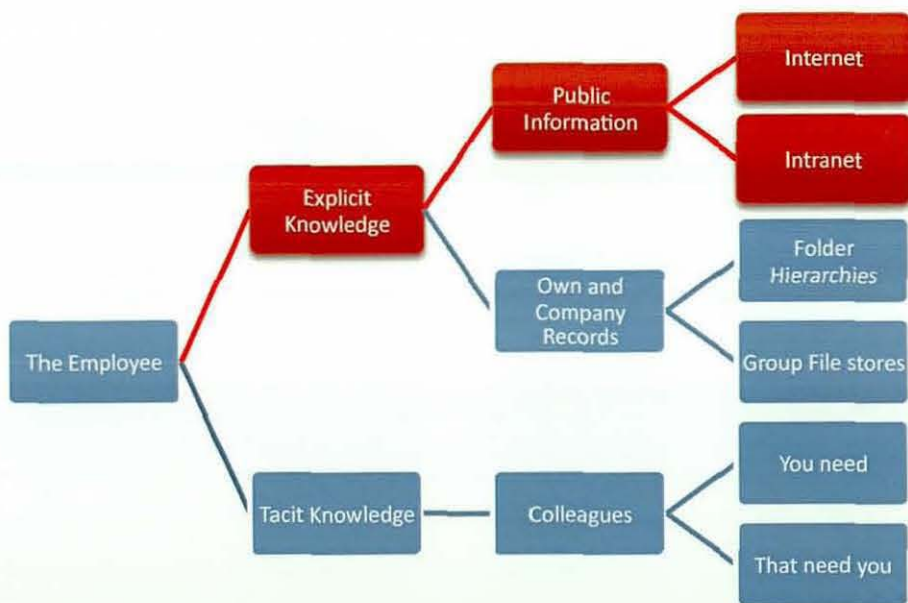


Figure 11 – Information Sources – Public Information

To address the issues surrounding information overload experienced by employees this chapter focuses upon search engines and how they can be used to increase the potential to find relevant information from both the Internet and corporate intranets, as shown by Figure 11.

## 5.2 The Need to Reduce Information Overload

The literature review identified the difficulty in discovering relevant information amongst the irrelevant. One of the key causes of information overload came from the inability to find relevant information within the large volume of information available today. With the increase of information available internally and externally to an organisation, and with the Internet cited as one of the key sources of information, it is the main cause of information overload (Nelson 1994).

The volume of information available to users on the Internet is growing dramatically. As previously stated, the literature review estimated that there were over 60 billion index-able pages on the Internet (De Kunder 2008), this number has grown from an estimated 11.5 billion in 2005 (Gulli, Signorini 2005), using similar techniques for estimation. Even if these estimations are slightly inaccurate the growth of available information is evident.

In addition to the information found upon the Internet most organisations also make use of internal corporate intranets, sharing company information. Due to their internal nature, these intranets often contain information more relevant to employees and form the basis for many knowledge sharing initiatives within organisations.

With the increase of available information, it is becoming increasingly difficult for employees to find the content that they require to perform their daily activities (Chen, Dumais 2000)(Kobayashi et al. 2006). With such an abundance of information available users must be presented with a method to determine where the relevant information lies amongst the information that is considered irrelevant to them.

For the user, attempting to discover information on the Internet or using intranet search engines often forms the first stage of their journey. Even in 1997, close to the dawn of the Google search engine, over 80% of Web searchers used Web search engines to locate online information or services (Nielsen Media, 1997). For many years there has been a critical need to understand how people use Web search engines (Amichai-Hamburger 2002).

The literature review showed that search engines have been the focus of a significant body of research in recent years. Comprehensive research existed from the algorithms that were used by search engines and even behaviour analysis to determine the way in that users searched upon the web. Although the core mainstream search engines such as Google, Yahoo and Microsoft Windows Live Search have added improvements to the core search ability, the representation of search results has barely changed since their conception years ago (Wiza, Walczak & Cellary 2004). The element of search engines that has remained untouched since the beginning and throughout all of the innovation of search engines is the presentation interface (Wiza, Walczak & Cellary 2004).

Given the search results the user will attempt to assess the relevance of the pages presented. The user will then choose the page they will navigate to. If the user chooses the wrong page then they will navigate to that site only to find the information is irrelevant and has contributed to the information overload problem.

Search results in the mainstream search engines are presented in a list structure. In the majority of search engines this consists of the title of the web page followed by a short summary of the web page's text. Traditionally this would be the first few lines of the document although more recently query based summaries have become popular, showing text that contains or is related to the terms found within the document (Paek, Dumais & Logan 2004). As the literature review stated, the query relevant text helps the user to determine how relevant the text is, but this functionality was not implemented in the majority of search engines until after this research had been conducted.

Enhanced visualisation approaches offer one alternative to traditional list based search results and although there was a limited amount of literature available in this area, it appeared to have received less attention than other areas of search engine research. It was determined that the presentation of search results presented a clear opportunity for research as it could potentially help reduce the problem of information overload.

### 5.3 Conceptualisation of Visualisation Techniques for Improving the Presentation of Search Results

Many of the visualisation methods shown in the literature review were successful but had significant disadvantages. Some needed browser plug-ins, some resulted in severe processing delays and some were simply not very effective or specialised in only one type of search. The literature highlighted the potential for visualisation to aid the search process, however none of the methods provided appeared to offer a suitable alternative to the traditional list-based results.

After analysing the visualisation techniques within the literature review and their success factors or weaknesses, a list of recommendations were created. There appeared to be a number of factors that lead to the success of a system and its successful adoption. There are also a number of factors that appeared to deter users from using the system. The issues encountered and success factors from across the studies within the literature review have been combined to create the following recommendations:

- Supplementary – A visualisation should be able to supplement existing plain text results as there is often a large adaptation to new methods of presenting results.
- Efficient – Although some delay has proved acceptable, a visualisation should not take so long to load that a user is hindered or even switches back to plain textual results. Any pre-processing that can be performed will obviously provide benefit.
- Plug-in free – Is possible browser plug-in should be avoided and pure html and/or java script should be used. This will avoid compatibility issues and help reduce loads times of a system.
- Familiar – If possible well known and proven concepts such as clustering and tag clouds could be used to reduce the barrier to entry of any visualisation.
  - As a sub point to this, it appears that three-dimensional systems provide too great a leap for many users.

- *Intuitive* – Although training should not really be necessary at all for any system, it is accepted that some degree of adaptation from the user may be required. However, a system should be easy to use from the outset.

Based on the recommendations any new visualisation technique must remain familiar to a user. Although a certain degree of adjustment is acceptable as long as the benefits are perceived, systems that present too much change simply do not appear to work effectively. Chapter 4.4 “Capturing the Extent of the Barriers to Knowledge Sharing” also highlighted that users within the two target organisations had a number of issues with new systems. If users could see the benefits of a system, then employees would be much more comfortable having to adapt to such a system. In some cases it was felt that new systems did not live up to expectations. When this is combined with many people feeling that the training given when a new system was introduced was inadequate, it is important to deliver a system that is familiar to users rather than something that is alien to them. If a system is unfamiliar to them then the need for training is going to be dramatically increased.

The system should be efficient, with minimal loading times and no requirement for additional downloads. Therefore, if possible, a pure html visualisation technique would be highly desirable. It was important that the visualisation did not replace existing search results. If the existing search results were replaced then that would alienate users who were happy with the existing search result presentation method. The literature review also mentioned the differences between ‘left-brain dominant’ and ‘right-brain dominant’ users. It would be important not to make things harder for users who were already happy with text-based search results and whom are brain-dominant toward this form of presentation.

These recommendations were developed through a review of the literature discovered relating to visualisation and as such were not requirements but recommendations. It is possible that a visualisation technique developed may not satisfy all of the factors identified, but the recommendations should aid during the development of a visualisation approach.

One highly accepted advancement in visualisation has been Tag Clouds. Tag clouds are an html driven visualisation of content showing a summary of the tags that have been manually assigned to all of the content of a site. Tag clouds actually have a number of different names on the Internet. McFedries (2006) gives reference to the names tag clouds, 'folksonomic zeitgeist' and 'tagroll'. More importantly McFedries also describes tag clouds as a 'list of the tags used on the site, although with some kind of visual indication of each tag's relative popularity'. Figure 12 shows an example tag cloud taken from the Guardian Newspaper shortly after a world cup football match between France and Italy.



**Figure 12 - The Guardian Newspaper's Folksonomic Zeitgeist**

Tag clouds have become an integral part of web-based systems within the concept of Web 2.0. They provide a lightweight and quite informative overview of the content of a site. In addition to the tag clouds visualisation the existing techniques displayed in the literature, and the lessons they can teach, should be considered for the development of a new visualisation. The next section provides a number of recommendations for visualisation development based on the literature reviewed in Chapter 2. Based on the recommendations and given the success and rapid adoption of tag clouds, they will be used as a starting point for this research to improve the visualisation of search results to help reduce information overload. The next section discusses the development of a new visualisation system called Concept Clouds.



## 5.4 Concept Cloud Development

As already mentioned, tag clouds provide a summary of all articles within a site rather than just one article. However, this visualisation technique could be used to present a visualisation description of a single article. This could aid the end user in searching for information. Instead of showing the tags that have been provided throughout the site, the same method of visualisation could be applied to the frequency that words appear within a document. Based on Tag Clouds a new visualisation system was created by the author and was named Concept Clouds.

The idea behind Concept Clouds is to supplement the existing search results that people are used to, whilst providing the user with additional information to allow them to determine the relevance of the information to them. The new system integrates into existing search result systems. As search results are displayed, a small visualisation is also presented to the user along with the existing results. This provides a quick summary, enabling users to gather an overview of the content within the search result, but it does not detract from the existing method of presenting results. This is important because it does not present something entirely different to the user but supplements the results that already exist.

The definition of Concept Clouds is, 'a list of concepts with a visual indication of their relative importance'. This definition could be further refined to give a 'weighted list of concepts'. In this case the importance or weight is the frequency of the occurrence. Therefore, the contents of the document are summarised by the list of weighted words and the key concepts, which enables the relevance of the document to be quickly established. With Concept Clouds being based upon tag clouds, it also provides a visualisation familiar to users.

When search engine results are created each result will contain a Concept Cloud that will act as a summary of the document. One of the great potential benefits of this concept is that it does not require pre-created categories. The images simply rely on the document content and, therefore, evolve as the content of the documents evolve. The Concept Clouds could also be pre-created and therefore add minimal additions to the search time. Using pure HTML and with minimal processing involved the Concept Clouds create a very efficient visualisation that

does not cause a processing delay for the user. The clouds are also very acceptable to html browsers, being made entirely from html mark-up, without the need for plug-ins or even the use of server generated images. Figure 13, shows an example of a document that you might find on a web site. In this case it is a news article from the BBC web site. Figure 14 shows a Concept Cloud that has been used to summarise the BBC news article.



**Can car parking enforcement be improved?**

**Will proposals to improve car parking enforcement make a difference?**

Plans to curb clamping and provide better training for traffic wardens in England have been unveiled by ministers.

Wheel clamping will only be used on the most persistent offenders and their details added to a national database. Also, the appeals process will be simplified.

The measures are part of a package designed to make car parking enforcement in England more friendly to motorists.

**What do you think of the government's proposals? Can car parking management be improved? Send your comments and experiences.**

**Figure 13 – An extract from the original BBC news article**

Figure 14 is an initial prototype and does not include advanced options. It is quite a good representation of the original article shown in (Figure 13). The Concept Cloud shows the importance of topics within the article based upon the weight and size of the font used for each word. A more frequently occurring word within the article will be shown in a bolder and larger font. Words that occur infrequently and words of no relevance do not appear at all leaving only the most frequently occurring words in the visualisation.

added also appeals **can car**  
clamping database designed  
**enforcement** england friendly  
government's have improve **improved**  
make management measures  
**parking** proposals

**Figure 14 – Concept Cloud created from a BBC news article the enforcement of clamping within English car parks.**

As Concept Clouds are based upon the features of a tag cloud system, they should present a familiar idea and visualisation to many users. Literature has stated that 28% of online Americans have already used the Internet to tag content themselves (Rainie 2007). Sites like Flickr and Del.icio.us have also brought tagging into the mainstream and given the emergence of tagging it was decided that Concept Clouds were based upon a similar visualisation to tag clouds in order to provide an element of familiarity and decrease the time to adoption.

To create the Concept Cloud from a web page, a number of steps are involved. The plain-text only content of the webpage must be extracted. This is performed through the use of a regular expression. The regular expression parses the html page and extracts the content that occurs within the tags of the page body, but removes all of the tags themselves. The regular expression also removes specific tags that do not contain useful content.

The Concept Cloud system also had limited support for known phrases. This allows the insertion of a number of key phrases that should be treated as a whole rather than as individual words. One example of this would be substituting 'Phase III' with 'Phase\_III'. Inserting the underscore character simply enables the system to treat all instances of 'Phase\_III' as the one word. In addition to this it may be desirable to replace 'Phase 3' with 'Phase\_III', which is again possible. Other text substitutions are also performed in order to remove undesirable elements from the text. Things like punctuation and quotes for example are removed. The pre-

processing system actually allows for any number regular expression based substitutions to be loaded and performed upon the text.

Once the pre-processing phase is complete the content is then tokenised, or split into individual words, and words are added one at a time to a list. Words are added one at a time, but if the word is within the list of 'stop words' then it is not added to the list. Stop words are words that have previously been determined to add little value. For example 'a', 'to', 'and' and 'the' are all within the list of words to be removed. The list of words used for this tool were taken from the open source Lucene indexing project.

The list is a dictionary which, allows an efficient storage method. As the words are added to the list they are converted to camel case, with an upper case first letter and lowercase for the remainder, to prevent any comparisons from being affected by case. Words that contain more than one capital letter are assumed to be abbreviations and the system will intentionally preserve the case.

As words are added to the list, a number of items are recorded. Firstly, the number of times the word has been added is stored. Secondly, the position that this word first occurred within the text is recorded. Once all of the words have been added to the list they are ordered by the number of times that the word occurs within the text. The system then takes the top occurring words and discards words that do not occur as frequently. The number of words that are kept may be chosen by the user but the default is 25. The word list also maintains the maximum number of occurrences of any word within the list for rendering later.

The words that are stored, the most frequent ones, are then ordered by one of three methods. The first method is to order the words based on the order that they first occurred within the text. The second is to order the words based upon frequency, although this disrupts the aim of the system, as the visualisation appears linear. The third and default option is to order the words alphabetically. Once the word list is completely generated it is passed to the rendering system, which renders the Concept Cloud from the word list.

The rendering system has a number of parameters. The first parameter states the minimum and maximum font size for the words within the system. The minimum

and maximum colours are also available as parameters but default to a very light grey for less frequent words, and a very dark grey for the most frequent words. The system then takes each word in turn and calculates the size and colour it should be. Once the size and colour has been calculated it then outputs the word within a span html tag. The span tag's style attributes are used to set the size of the span and the colour of the text within the span, as previously determined.

Figure 15 shows the output of a Concept Cloud. The Concept Cloud was generated using the start of a paper published relating to Concept Clouds (Smith, Jackson & Adelman 2007) and shows the concept of documents, 'search' and 'search engines' along with 'information' and 'information discovery', all of which are key topics of the paper.



Figure 15 - Example Concept Cloud

Currently there is no system in place to cater for the differentiation between plural or singular forms of the content, although this could be implemented in future work.

The Concept Clouds do not require significant overhead to display the cloud, but there is a possibility of a slight overhead involved in creating the Concept Cloud. In order to prevent any performance issues it is possible to pre-generate the Concept Cloud content and cache this content for retrieval later. The next stage in the research was to assess the system's potential in providing an improved search facility to the end user.

In summary the Concept Clouds system is supplementary and designed to be used alongside existing search results to provide an additional overview. Concept clouds

do not require any rendering of images or intensive graphical processing. The Concept Clouds can also be pre-generate if required allowing the content to be cached and stored to ensure an even faster response time. The output of the Concept Cloud is also in pure HTML. This brings three key benefits. The first benefit is that the bandwidth required is minimal. The second is that the visualisation is efficiently rendered by all major web browsers and finally no plug-ins are required to perform the rendering. Being plug-in free helps to ensure that the visualisation has maximum compatibility. Being based upon the concept of tag clouds also makes the visualisation familiar to a wide range of Internet users. This should also help the system to be intuitive although this shall be explored further, as the system is assessed in section 5.5.

Since the development of this research many search engines have now introduced query relevant search text. The traditional list results presented to the user would not simply contain the site's description, as determined by the author, but would include extracts of the page content that the search engine deemed relevant. This query relevant text was not included in the assessments performed and may affect how beneficial the findings are today. However, the Concept Cloud system still has the advantage of showing the overall theme of the document and how frequently certain concepts appear within it rather than an abstract. The Concept Clouds also present the findings in a more visual way allowing the user to quickly see the focus of the entire document. Finally the Concept Clouds can still be used in addition to this query relevant text and their use does not have to be exclusive. Future work would be required to determine the benefit of the concept cloud system along side the query relevant search text.

## **5.5 Assessing the Potential and Performance of Concept Clouds**

In order to assess the performance of the Concept Clouds system, two assessments were performed. The first study was used to gather an overview of the potential of the Concept Clouds system. The secondary study was performed in order to gain at a deeper understanding of the system's performance.

To assess the performance of the Concept Clouds system, a tool was developed to measure a user's search speed whilst answering multiple-choice questions. The

web-based assessment tool presented users with a question and then loaded a page within a frame below the question. As the frame was loaded, a timer was started. The user could then choose the answer from a list of possibilities. If the answer the user chose was correct then the next question would be presented. If the answer was incorrect then this would be recorded but the timer would continue until the correct answer was chosen.

Whilst the tests were undertaken, an XML file was created containing the user's id, the amount of time they had taken to answer each question correctly and the number of incorrect answers the user had given. These results were then combined and evaluated using Microsoft Excel.

### **5.5.1 Assessing a users search performance**

To assess the potential and performance of the Concept Clouds system, users were split into two groups. Those that would use the Concept Clouds and those that would not. This approach could introduce bias, so a matching pair study was performed in order to split the group equally. The matching pair study asked a number of questions and presented the standard Google search engine home page. It allowed users to enter their own queries in order to find the results or presented the users with a specific page to begin their search.

Users answered questions relating to the three types of search discussed in section 2.6.1 'Types of search'. Namely searching for a theme, searching for a specific document and searching for content within a document.

Users were then ranked, based on their number of incorrect answers and the time it took them to answer the questions. This was then used to split the users into two groups, those who would use the Concept Clouds and those who would not. The matching pair study also acted as an introduction to the assessment tool and provided a more balanced assessment. An example of the initial assessment tool is shown in Figure 16.

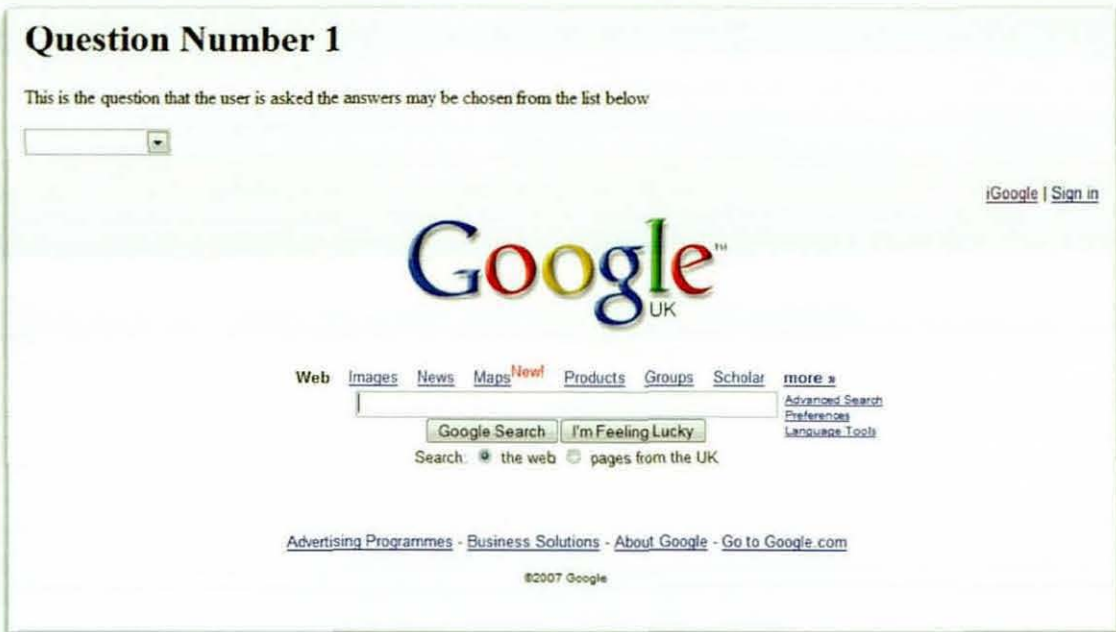


Figure 16 - The initial assessment system

### 5.5.2 Study one

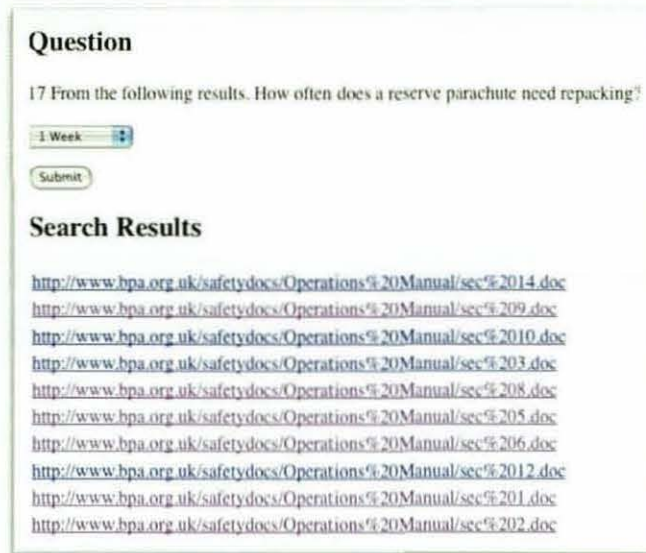
The initial assessment comprised of two phases, the participants search ability and the Concept Cloud assessment.

Unfortunately it was not possible to assess the Concept Cloud system within either of the case-study organisations or within another organisation. It was not possible to find an organisation that would allow a representative sample of its employees to take part in the laboratory experiment. Instead in order to get a sufficient sample undergraduate students were used to perform the experiment. The undergraduates were chosen as it was felt they could be sufficiently generalised to a population of knowledge workers. The students were on an Information Retrieval module and would hopefully soon be going into industry within this field. They would represent a selection of people that would hopefully generalise to those working in organisations and may even suffer from information overload themselves.

There were 90 participants that took part in the matching pair assessment and of those, just 18 participants took part in the comparison of the two systems. The reduced number was due to not all participants turning up for the second phase of the research even though they had been invited. This meant that there were 18



participants, 9 of which used the Concept Clouds and 9 that did not. The groups were divided so that their abilities were as equally matched as possible from the ranking results. Following this, the users were asked questions and were presented with a number of Uniform Resource Indicators (URL's) that might contain the required result results, as shown by Figure 17.



**Question**

17 From the following results. How often does a reserve parachute need repacking?

1 Week

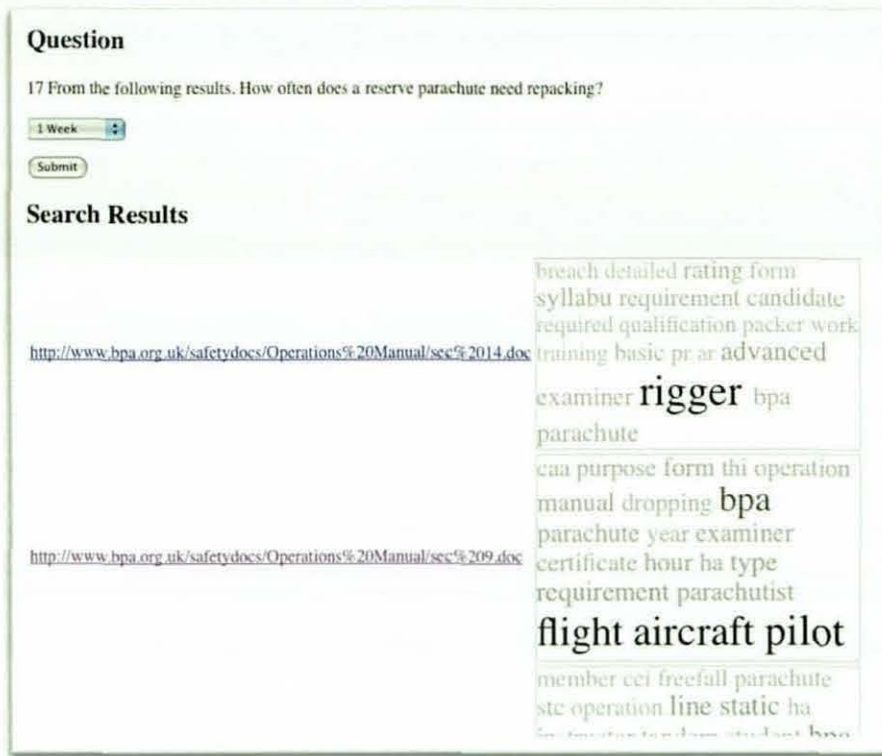
Submit

**Search Results**

<http://www.bpa.org.uk/safetydocs/Operations%20Manual/sec%2014.doc>  
<http://www.bpa.org.uk/safetydocs/Operations%20Manual/sec%209.doc>  
<http://www.bpa.org.uk/safetydocs/Operations%20Manual/sec%2010.doc>  
<http://www.bpa.org.uk/safetydocs/Operations%20Manual/sec%203.doc>  
<http://www.bpa.org.uk/safetydocs/Operations%20Manual/sec%208.doc>  
<http://www.bpa.org.uk/safetydocs/Operations%20Manual/sec%205.doc>  
<http://www.bpa.org.uk/safetydocs/Operations%20Manual/sec%206.doc>  
<http://www.bpa.org.uk/safetydocs/Operations%20Manual/sec%2012.doc>  
<http://www.bpa.org.uk/safetydocs/Operations%20Manual/sec%201.doc>  
<http://www.bpa.org.uk/safetydocs/Operations%20Manual/sec%202.doc>

**Figure 17 - The search assessment system – without Concept Clouds**

Those with the Concept Cloud system were also presented with a Concept Cloud created from the page that the URL linked to, as shown by Figure 18.



**Figure 18- The search assessment system – with Concept Clouds**

The speed at which the users answered the full set of questions and the number of incorrect answers given by each user was recorded.

### 5.5.2.1 Results

Table 20 shows the search response times and the number of incorrect answers for each of the users. The results have been ordered by the total time taken to answer the questions by the user.

**Table 20 – Concept Cloud Study One Results**

#### Without Concept Clouds

| User   | No. Incorrect | Total time (sec) |
|--------|---------------|------------------|
| User 1 | 0             | 910              |
| User 2 | 2             | 877              |
| User 3 | 0             | 784              |
| User 4 | 0             | 764              |
| User 5 | 0             | 708              |
| User 6 | 1             | 701              |
| User 7 | 3             | 685              |
| User 8 | 0             | 656              |
| User 9 | 0             | 506              |
| Total  | 6             | 6591             |

#### With Concept Clouds

| user    | No. Incorrect | Total time (sec) |
|---------|---------------|------------------|
| User 10 | 0             | 857              |
| User 11 | 0             | 687              |
| User 12 | 0             | 397              |
| User 13 | 0             | 326              |
| User 14 | 1             | 207              |
| User 15 | 0             | 107              |
| User 16 | 0             | 90               |
| User 17 | 0             | 90               |
| User 18 | 0             | 88               |
| Total   | 1             | 2849             |

The results show that users of the Concept Cloud system answered over two times faster than users without the system and they also made fewer mistakes.

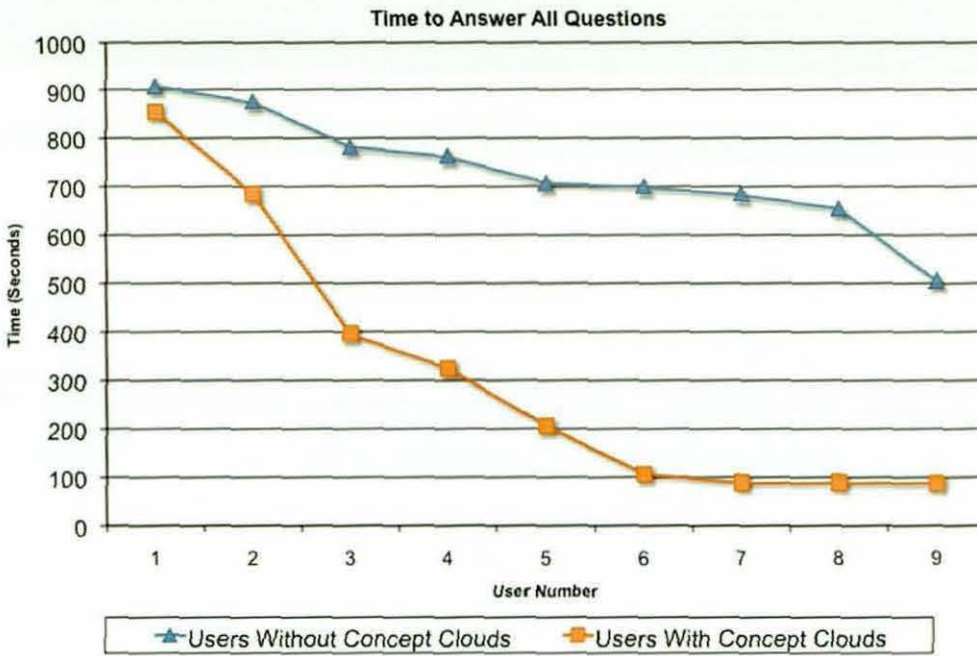


Figure 19 - Total times of each user

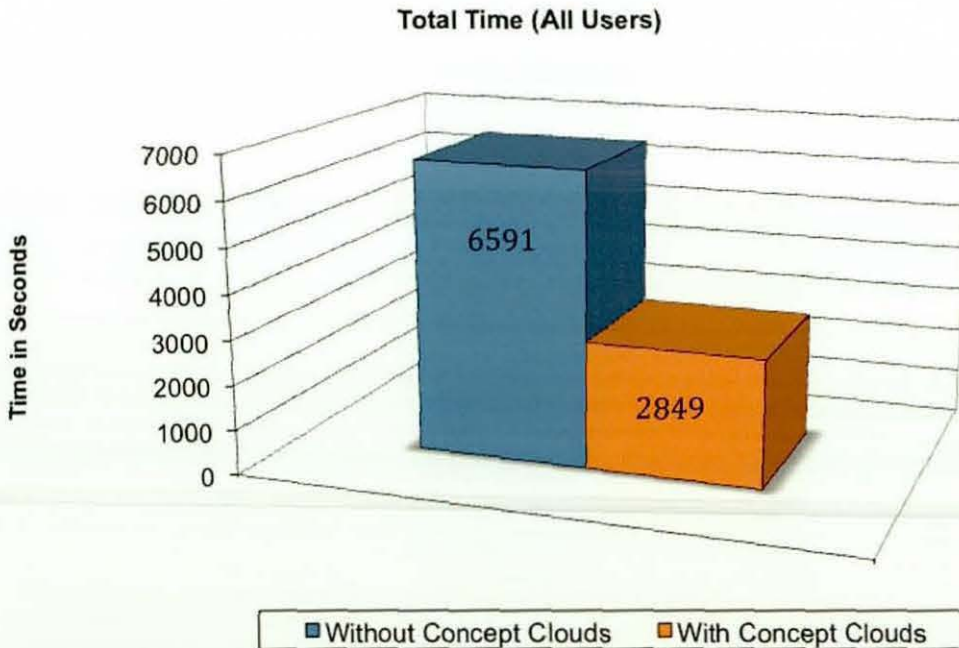


Figure 20 - Total times of all users

Figure 19 shows all of the users plotted in order of the most time to the least time, and Figure 20 shows the difference between the total times taken to complete the exercise. The Wilcoxon signed rank test was used to assess the significance of

these results. It was chosen as there was no guarantee that the data was normally distributed and two matched samples were being compared. The Wilcoxon signed rank test gave  $W = 45$ , since the sample size was 9 it could be inferred that the data was significant far beyond that required .05 confidence level. Further to this the test proved significant to the 0.1 confidence level.

There were two users, numbers ten and eleven in Table 20 that used the Concept Cloud system that have comparable times to those who didn't use Concept Clouds. They were the first two users to use the Concept Cloud system. The users had not been introduced to the Concept and had not received any training. These two users noted that they did not understand what the Concept Clouds were for or how to use them. The following seven users however, received a quick briefing on how to use the system, following the early feedback. Subsequently the other participants were far more comfortable using the system and understood how it could be used to improve their search following the short briefing. This suggests that although there is a benefit to the system, that training is necessary before any benefit will be seen. Although the original intention was that training should not really be necessary, many of the participants of this study had never seen a tag cloud before and thus would not understand the idea of a Concept Cloud. Only a very small amount of training should also be required. As tag clouds continue to become more commonplace on the Internet, the acceptance of Concept Clouds should grow.

This test does show a bias against a mainstream search engine because only links were presented to the users who did not see the Concept Clouds. This is representative of the document retrieval system based within the organisation that the Concept Cloud system is to be implemented within both PharmaCo and SoftwareCo. It is also believed that the Concept Clouds will be of benefit when used to supplement mainstream search engines as currently only a small sequential description is shown as a summary, not a collection of key terms. Although this study had a reduced number of participants than hoped for, the results are encouraging and show that it can aid an end user in their searching by cutting down their search time by over half and reduced the number of incorrect answers chosen by participants by 83%.

### 5.5.3 Study Two

The second study of Concept Clouds was conducted with the aim of obtaining a greater understanding of the tool and to obtain a higher participation rate than the first study. Again it was not possible to obtain a sample from actual organisations and undergraduate students taking the same module as those in the first study were used. The participants were a different group of students to those that took part in the first study. It was felt that the students could again generalise but this time more emphasis was placed on obtaining a larger sample size.

The second study had 79 participants. The participants were split into two groups by surname and attended two one-hour sessions. In the first session, the users undertook an assessment of their general search abilities, like study one. In the second session, the comparison between the Concept Cloud based system and the text only system was performed.

As with the first study, the users were split into groups based upon the number of incorrect answers they gave and the amount of time taken to answer the initial assessment questions. Within the two groups the fastest person took the text only system and the second fastest the Concept Clouds system. This continued with all odd numbered participants using the text only system and all even numbered participants using the Concept Clouds system.

There were a number of lessons learnt from the first study. Firstly, users would be encouraged not to simply try each answer in order to finish the test first. The need for a fair trial was explained to users and the number of possible answers was increased from four to ten so that any users attempting to 'cheat' the system could easily be detected. Secondly, many users would ask others who had already finished to give them the answers. In order to try and prevent this, the benefits of a fair trial were explained again and also all users were started at exactly the same time to try to improve the situation. The final observation from the first study was some users using the Concept Cloud system simply did not understand what they were for. For this reason all users were given a brief introduction to the Concept Clouds system and how the system should be used.

Greater effort was taken to ensure that as many people as possible who performed the matching pair assessment would return to take part in the comparing the Concept Clouds against the text only system. Of the 82 people who participated in the matching pair assessment, 79 users completed the second study. There were 40 people who acted as the control and used the text only system and 39 who used the Concept Clouds.

A total of 8 questions were asked and they related to type one or type two searches. Type one was based on specifically searching for a range of articles around a particular subject. Type two was based on searching for one article in particular. Participants were free to use search facilities within the browser to find content within a page if they desired, for example, using the Ctrl+F find function.

The questions the users were asked were as follows:

1. How often does a reserve parachute need repacking?
2. In the following manual Section 2.1.5 of the Training document covers which aspect of training?
3. What is the capital of Kazakhstan?
4. How many pages does the book written by Obie Fernandez contain?
5. Which of the following links shows an extract from Scott Guthrie's post about improved performance in .Net applications?
6. Which of the following talks about securing something and mentions finger print scanning?
7. Which of the following blog entries mentions something called EPUB?
8. Which of the following discusses an enhanced version of the little Gem Battery powered amplifier?

In most of the cases, users would be searching for a particular article. However, the system could also be used to filter a number of articles to discard the ones that are not of use. Once irrelevant information has been removed the user is then free to

use the old system of manually searching through the remaining documents to find the correct one.

#### **5.5.3.1 Results**

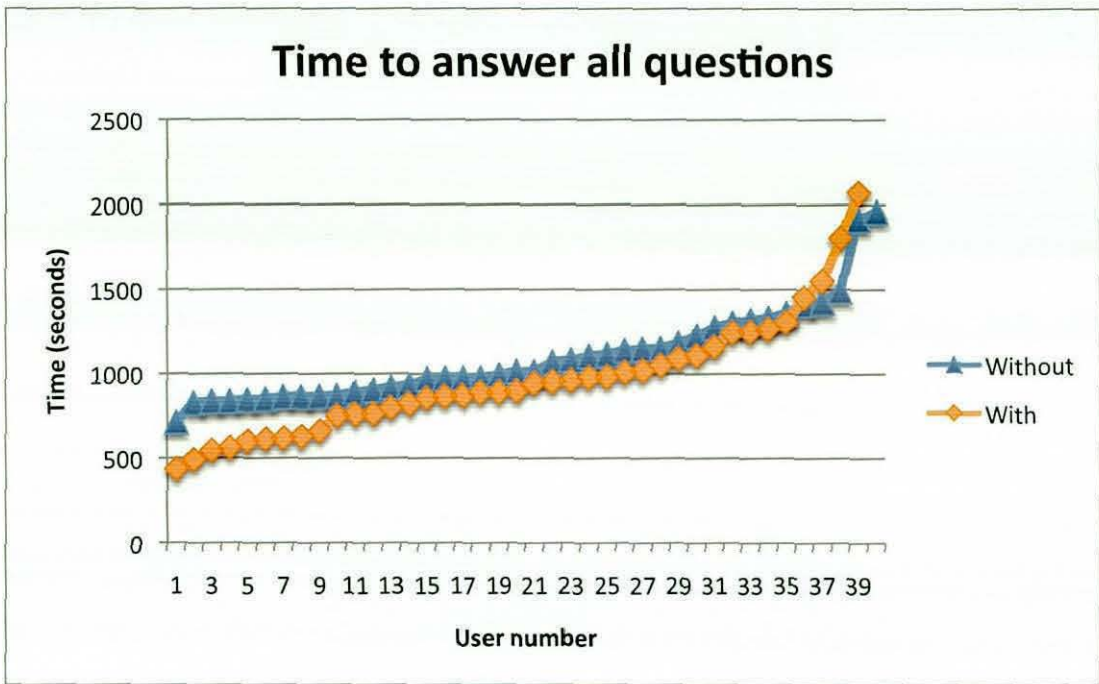
Results were gathered using the assessment system and then exported as a Comma Separate Value (CSV) file. The CSV file was imported into Microsoft Excel and this was used for analysis. Table 21 shows the total time to answer all questions and the number of incorrect answers given by each user. The number of incorrect answers given by the user was used as the metric to order the results.

**Table 21 – Concept Clouds Study Two Results**

| Without |                  |                   |
|---------|------------------|-------------------|
| User    | Total Time (Sec) | Incorrect Answers |
| 1       | 712              | 0                 |
| 2       | 829              | 0                 |
| 3       | 837              | 0                 |
| 4       | 845              | 0                 |
| 5       | 863              | 0                 |
| 6       | 866              | 0                 |
| 7       | 869              | 0                 |
| 8       | 890              | 0                 |
| 9       | 904              | 0                 |
| 10      | 932              | 0                 |
| 11      | 972              | 0                 |
| 12      | 974              | 0                 |
| 13      | 993              | 0                 |
| 14      | 1014             | 0                 |
| 15      | 1015             | 0                 |
| 16      | 1115             | 0                 |
| 17      | 1138             | 0                 |
| 18      | 1148             | 0                 |
| 19      | 1152             | 0                 |
| 20      | 1275             | 0                 |
| 21      | 1421             | 0                 |
| 22      | 1914             | 0                 |
| 23      | 1958             | 0                 |
| 24      | 836              | 1                 |
| 25      | 864              | 1                 |
| 26      | 975              | 1                 |
| 27      | 1069             | 1                 |
| 28      | 1082             | 1                 |
| 29      | 1106             | 1                 |
| 30      | 1180             | 1                 |
| 31      | 1300             | 1                 |
| 32      | 1330             | 1                 |
| 33      | 1411             | 1                 |
| 34      | 1491             | 1                 |
| 35      | 919              | 3                 |
| 36      | 850              | 4                 |
| 37      | 1225             | 6                 |
| 38      | 1358             | 9                 |
| 39      | 1318             | 10                |
| 40      | 976              | 30                |
| Total   | 43926            | 73                |
| Average | 1098.2           | 1.8               |

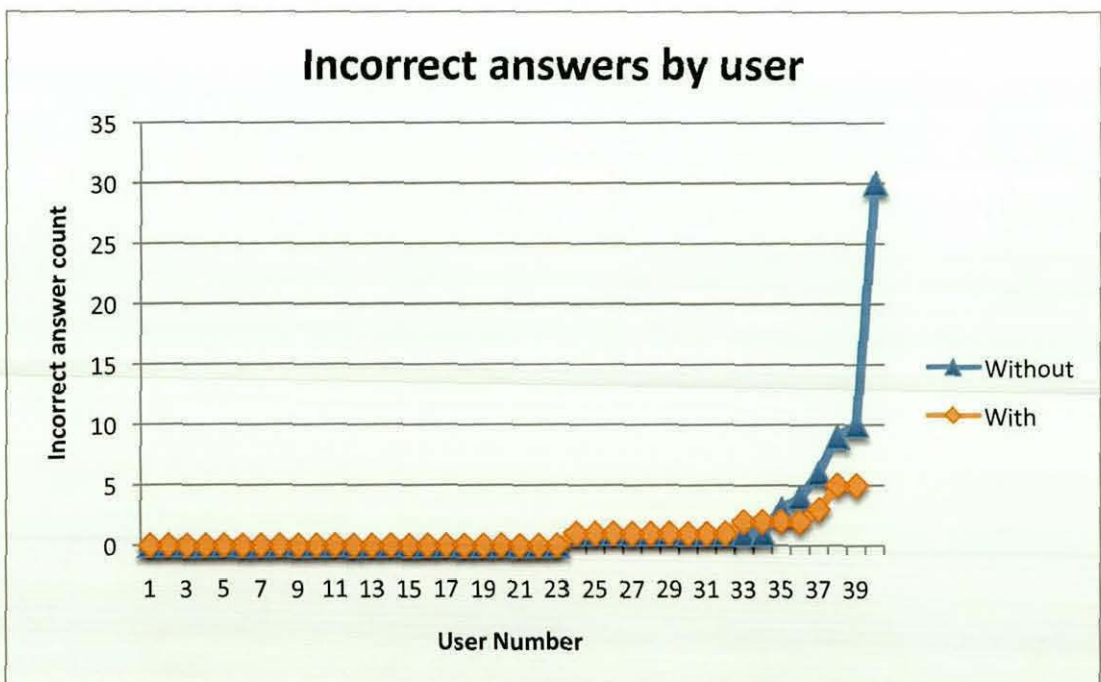
| With    |                  |                   |
|---------|------------------|-------------------|
| User    | Total Time (Sec) | Incorrect Answers |
| 1       | 437              | 0                 |
| 2       | 603              | 0                 |
| 3       | 620              | 0                 |
| 4       | 762              | 0                 |
| 5       | 764              | 0                 |
| 6       | 823              | 0                 |
| 7       | 857              | 0                 |
| 8       | 867              | 0                 |
| 9       | 890              | 0                 |
| 10      | 891              | 0                 |
| 11      | 897              | 0                 |
| 12      | 953              | 0                 |
| 13      | 960              | 0                 |
| 14      | 980              | 0                 |
| 15      | 983              | 0                 |
| 16      | 1008             | 0                 |
| 17      | 1020             | 0                 |
| 18      | 1051             | 0                 |
| 19      | 1091             | 0                 |
| 20      | 1110             | 0                 |
| 21      | 1246             | 0                 |
| 22      | 1454             | 0                 |
| 23      | 1806             | 0                 |
| 24      | 493              | 1                 |
| 25      | 615              | 1                 |
| 26      | 631              | 1                 |
| 27      | 663              | 1                 |
| 28      | 871              | 1                 |
| 29      | 1162             | 1                 |
| 30      | 1251             | 1                 |
| 31      | 1267             | 1                 |
| 32      | 1308             | 1                 |
| 33      | 547              | 2                 |
| 34      | 751              | 2                 |
| 35      | 799              | 2                 |
| 36      | 946              | 2                 |
| 37      | 563              | 3                 |
| 38      | 1552             | 5                 |
| 39      | 2070             | 5                 |
| Total   | 37562            | 30                |
| Average | 963.1            | 0.8               |





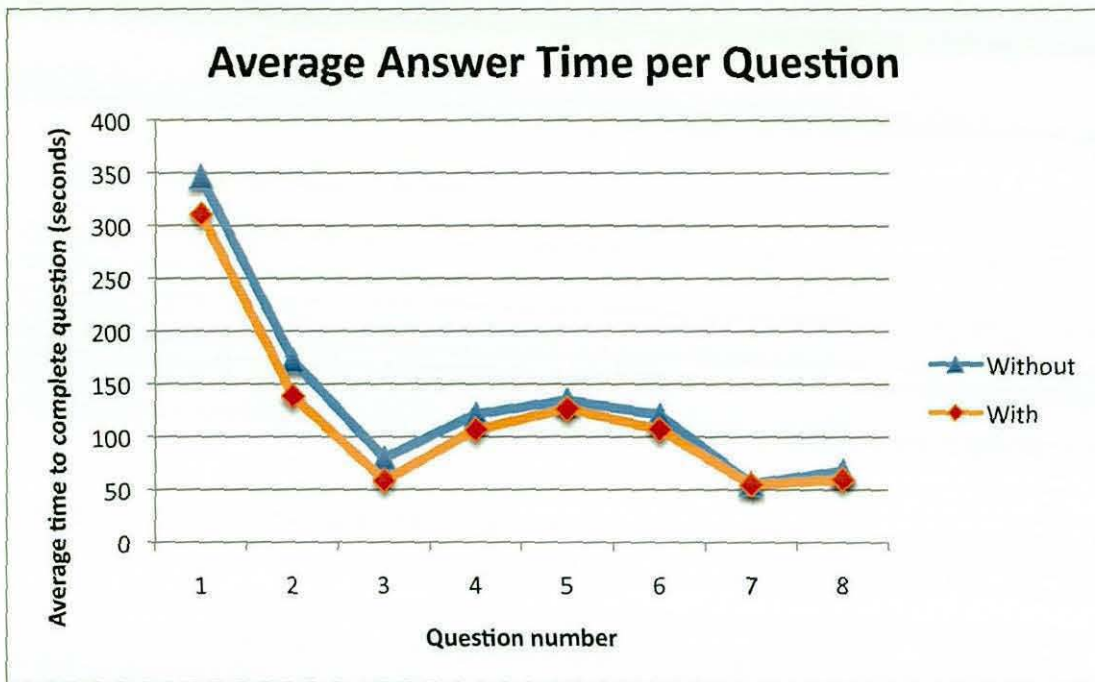
**Figure 21 - Like for like user comparison of the two systems by user**

Figure 21 shows the total time taken by each user for both systems. The total time taken to answer the questions for each user was calculated and then the users were split into groups depending on whether they used the Concept Clouds system or not. The results were then ordered so that the fastest user from the Concept Clouds system could be compared with the fastest user of the text only system.



**Figure 22 – Comparison of incorrect response count with and without Concept Clouds**

Figure 22 shows the number of incorrect responses that each user gave and is again ordered so a like for like comparison can be made.



**Figure 23 - Average time per question with and without Concept Clouds**

Finally, Figure 23 shows the average time taken by all users to answer each individual question.

The results show an improvement when using the Concept Cloud based system over using the text-based system. Although the improvement is not dramatic it is noticeable and consistent. The data within Figure 21 shows that the decrease in total time required is higher for the fastest users. Those users that were quite fast at finding results had the largest advantage using the system. The difference begins to reduce as the results move towards the less experienced users, although the majority still had an advantage. Towards the slow end of the scale, there were a number of users that took longer using the Concept Clouds system than using the text-only system. These users might have been confused by the system and took longer to answer all of the questions because of this.

Looking at the number of incorrect answers given by users in Figure 22, the incorrect answers are relatively similar. There were more incorrect answers given by users using the text-based system and this may have been due to the users of

the Concept Clouds system being able to narrow down the possible answers. There also is one anomaly, one user of the text based system answered incorrectly 30 times, which is not an accurate reflection of all users that used the text based system.

When looking on a question-by-question basis as shown in Figure 23, the improvement given by the Concept Clouds system appears fairly consistent although, the questions four to eight appear to have provided slightly less benefit than questions one to four. Interestingly, the questions four to eight began with the statement "which of the following links...". The question then went on to ask if the article talked about something in particular. These later questions, four to eight, asked users to find the answer to a specific question from a group of answers and involved the type two searches. These questions saw more benefit than those which simply asked the user to find which article in particular talked about a certain subject.

Perhaps it was just as easy to scan each of the links to see if anything relating to that content appeared rather than looking for a specific fact within the articles. It is of course harder to see the difference in the results when the question took less time but there are improvements throughout.

Questions one to four showed between a 10 and 25 percent decrease in the speed at which users answered the questions. Questions four to eight, the ones that stated "which of the following links", showed between a two and 12 percent decrease. Overall, there was a 13 percent decrease in the time taken to respond using the Concept Clouds system. If the one user who answered incorrectly 30 times is removed, there is also still a 28% decrease in the number of incorrect answers given by those users who used the Concept Clouds system.

If the number of incorrect answers is reduced then it is possible that information used within the organisation will be more accurate. The real saving however comes in the reduction of time taken to find the 'correct' article that participants were looking for.

The literature found that one of the most common sources attributed to causing information overload was the Internet (Hemp 2009). With over 60 billion

indexable pages the Internet represents a huge corpus of potentially relevant information. Unfortunately due to its nature it also represents a mass of information that is potentially irrelevant and hence it can be difficult to find the relevant information amongst the irrelevant.

Within the questionnaire in Chapter 1 employees of both PharmaCo and SoftwareCo were asked how often they made use of the World Wide Web for work. Of the PharmaCo respondents 30% said all the time, 17% hourly and 39% once or twice per day. In SoftwareCo web usage was more prolific. Of the employees of SoftwareCo 52% said all the time and 23% said hourly. Given that the literature review stated that even in 1997 over 80% of users used search engines to locate online information or services (Nielsen Media, 1997), these improvements in both time and accuracy would represent a significant advantage in determining the location of relevant information to an organisation.

Concept clouds have presented an adaptation of a known technique to a new area, supplementing search results with a visual ability to determine the key themes of the document. They have presented a way to manage, by interrogation, the large amounts of information available to users and help people identify relevant links. In turn helping users navigate the *copious amounts of information* to reduce overload burden. With more relevant information available in less time, the information overload problem can be reduced with adopting the Concept Clouds approach.

## **5.6 Reviewing the Recommendations for Building and Implementing Visualisation Systems**

After assessing the Concept Clouds system it appears that the recommendations developed in section 5.3 could be used to create an *effective tool* to aid the discovery of relevant information. The Concept Clouds system allowed users to find information more quickly and accurately than with, traditional, list based search results and received positive feedback.

The Concept Clouds system was supplementary to existing search results. The Concept Clouds system is used in conjunction with the current list based approach

in order to give users an overview of the page content and concepts found on that page. Being based purely upon HTML mark-up, the Concept Clouds system was also very efficient visualisation. The browser could render the visualisation extremely quickly and minimal bandwidth is required to transfer the visualisation to the user's browser. It is also possible to pre-process pages and cache visualisations in order to improve efficiency of the system further. Being based upon HTML also had the added benefit of the system being plug-in free. The visualisation required no browser plug-in, such as flash in order to display the results.

Whilst the system met a number of the recommendations, it appears that two of the recommendations were difficult to apply. The system was intended to be familiar, however, many of the participants of the system were not familiar with tag clouds and thus the Concept Clouds appeared to represent a new idea to them. This issue also affected the intuitiveness of the system. The original intention was that no training would be required in order to use the system. During the initial study, a number of users appeared to misunderstand the Concept Clouds and became confused by them. In order to combat this, a small amount of training was given before further users were introduced to the system. After initial training users were able to understand the system and use it to its full potential. It appears that it is very difficult to create a system that is intuitive and familiar to all users, but with a small amount of training users were able to use the system very quickly.

## 5.7 Conclusions

The research detailed in this chapter has shown an increasing need to provide information seekers with a more effective approach of analysing search results. The literature review highlighted a number of attempts to provide alternate visualisations each with their own advantages and disadvantages. As a result of the literature review and the knowledge sharing questionnaire in Chapter 1, a list of recommendations have been constructed to provide a better understanding of how to build and implement a visualisation system.

The recommendations were used to adapt an existing system called Tag Clouds to provide a new system called Concept Clouds. In line with the recommendations,

the Concept Clouds required no additional browser plug-ins and was extremely fast to render using standard html. The visualisation could have also been pre-computed to prevent any processing delays. The idea behind the new system was to integrate the visualisation into existing search results. As search results are displayed, a small visualisation is presented to the user along with the existing results. This provided a quick summary, enabling users to gather an overview of the content within the search results, potentially saving them time and improving their accuracy.

Two studies were conducted to assess the potential of the Concept Clouds. The results of both studies showed an overall improvement in the user's performance when compared to a traditional method of presenting the search results.

Although the benefits were not as high as potentially expected they do represent a significant difference. On average there was a 13 percent decrease in the overall time that users took to answer all of the questions. Whilst this is a moderate saving and may not at first seem substantial but a reduction of this size would equate to a saving of almost 8 minutes in every hour spent searching the web

The benefits of the system also appeared to be higher for those users who were shown to have the most experience when using search engines. This highlights the benefit of the system to those users who are already experienced searchers on the Internet and not just those who are not.

The Concept Cloud system has shown that the recommendations can be used to form the foundation for building new visualisation approaches to aid the discovery of relevant information. In addition the visualisation has helped demonstrate areas where the recommendations might be refined. Providing no initial introduction to the system did negatively affect the results. The original intention was that training should not really be necessary, but many of the participants of this study had never seen a tag cloud before and thus would not understand the idea of a Concept Cloud. Therefore after gaining the early initial results, minimal training was given to educate users of the benefits of this system. This enabled the users to use the tool more effectively.

In summary this chapter has developed a set of recommendations that can be used as the basis for developing new visualisation techniques to present search results. The chapter has shown it is possible to increase the performance of search users through the use of visualisation which in turn can aid users to discover relevant information from a source often attributed to Information overload, the Internet, and from intranets.

## 6 Using Tagging to Discover Networked and Local Information

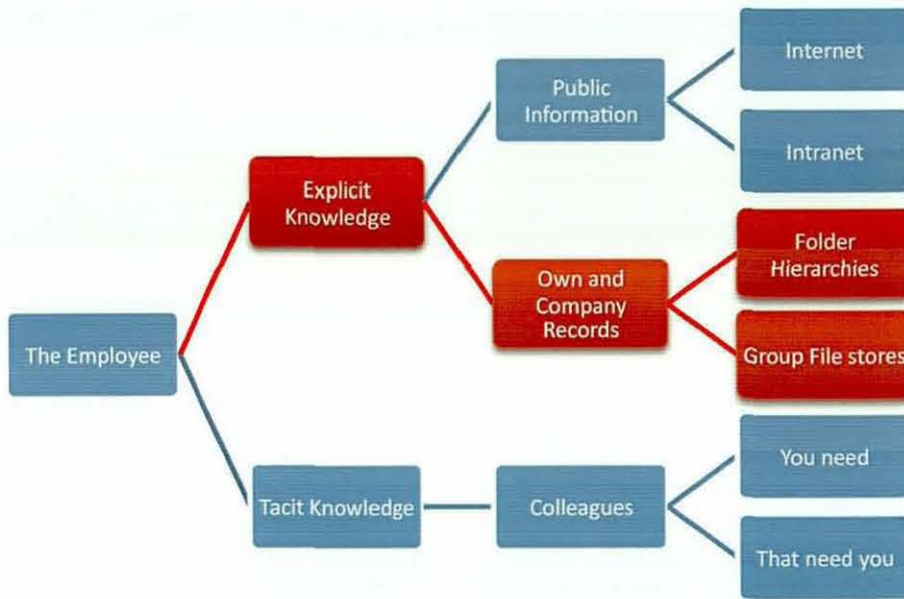
### 6.1 Chapter Preface

The previous chapter discussed the discovery of information on both the Internet and company intranets and proposed a possible solution to help aid the information overload problem when searching for relevant information online. *This chapter continues to investigate the documents available to users and further explores the information created and stored within the organisation.*

Although the Internet was cited as one of the major sources of information, it was also cited as one of the key sources of information overload because too much irrelevant information was available (Farhoomand, Drury 2002). The documents within an organisation may contain specific information that may also be unavailable outside the organisation. For this reason this chapter investigates a possible solution to reducing the information overload problem for documents within a company. Including those that may not be as easy to find as full text html pages.

The previous chapter also introduced the concept of tag clouds and tagging. This chapter, as shown by Figure 24, shall build upon the traditional use of tagging and investigate its potential benefit in the retrieval of relevant information within an organisation. The chapter shall investigate two key areas. The first is the ability of users to make use of tagging and the barriers that may exist when tagging. The second area proposes an entirely new and novel file system based upon the use of tagging instead of the traditional hierarchical directory structure.





**Figure 24 - Information Sources - Own and Company Records**

## 6.2 Searching for Local and Networked Information

### 6.2.1 Discovering Relevant Information from Local and Networked Sources

The previous chapter and literature review outline the problem of discovering relevant information. Searching for files within an organisation has been a source of problems for employees for a number of years (Chen, Dumais 2000)(Kobayashi et al. 2006). For this reason many companies invest substantial amounts of money in searchable portals and document storage facilities. Ineffective searches and wasting time looking for information has been said to cost up to 10% an employees time (Dubie 2006).

Many organisations employ highly sophisticated search engines to search the full text of all of the files stored within the system. The literature review in section 2.6 has already shown how not only are these systems often not sufficient, but their presentation of the search results is often inadequate, causing a user to manually sift through the results returned to find the relevant document.

Although the previous chapter has shown that visualisation can help users to discover more relevant information, the process of visiting a web page is not always an efficient one. Many search systems also cannot help retrieve documents that have not already been indexed by that search system such as those within a

user's corporate network profile. In addition web based search systems, such as Google or even a corporate intranet search engine, do not help users when finding files stored on their local computers. The user can resort to using their operating systems search facility but this suffers from the same problems as the web based search systems. Using such search engines also requires users to visit the search page and wait for results to be returned. Further to this, if the user wishes to browse for a file whilst continuously narrowing down their search as they go, they are restricted to a traditional hierarchical file structure.

Ultimately there are a number of barriers for users attempting to find information using traditional search approaches. This increases the difficulty in discovering relevant information and can lead to an inability to find what the user is looking for. To illustrate the severity of this problem Nelson references Naisbitt (Nelson 1994) arguing that, "Inundated with technical data, some scientists claim it takes less time to do an experiment than to find out whether or not it has been done before."

### **6.2.2 Can Techniques from the Internet Provide a Solution?**

The task of discovering relevant information has seen specific attention within the field of the Internet where, as already mentioned in section 2.6, the corpus of information available is huge (De Kunder 2008). Over time the Internet has moved from its traditional text and hyperlink driven state to a frenzy of media rich websites containing masses of information for, often, millions of users. Within this transition some interesting sites and approaches to discovering information have emerged. Flickr and Del.icio.us provide two interesting case studies.

Flickr is an online photo sharing website. It allows users to upload their photos for other people to see. Del.icio.us allows users to bookmark pages and then find their own and other users' bookmarks when needed. To quantify the volume of information that is stored in these systems, the Flickr website receives thousands of new photos per minute (Flickr 2006). All of these photos are stored and retrievable by its members. Del.icio.us is also a high traffic site; it received 150,000 posts per day in June 2008 (Keller 2008). Sites such as Flickr and Del.icio.us faced the challenge of allowing users to find content within huge numbers of photos and

bookmarks respectively and have become widely accepted names as part of the concept of Web 2.0. As has their solution to the discovery of relevant information which both of these websites have in common. Instead of using traditional methods and categories to allow users to find photos or bookmarks, they make use of tagging, a technology that has become synonymous with the concept of Web 2.0. It is also claimed that 28% of online Americans have tagged content on the Internet (Rainie 2007). Although tagging is not a new concept, its use within web based applications is certainly increasing. Tagging offers a strong alternative to traditional hierarchical structures and has allowed many web based systems to dispense with manually created categories for users to place their content into.

Tagging has proven extremely popular, however before it can be used on a large scale, there are a number of potential barriers that must be addressed. The literature review identified a number of these potential issues and suggested that training within the field of tagging might be beneficial. The literature review also provided a number of recommendations to help ensure success when tagging is used. Based upon the literature review and a number of sources (Golder, Huberman 2006)(Mathes 2004)(Sood et al. 2007), the following issues have been highlighted and some recommendations have been developed during the research into tagging.

- Single use tags – or tags which have not been used before should be avoided unless necessary.
- Pluralised or singularised versions of words – A decision should be taken on whether to allow both or just one, if only one is used then the singular form is recommended.
- Spelling mistakes – can cause the creation of a new tag and make this tag of no benefit.
- Personal tags - should be avoided however they could be prefixed with a period to allow distinction.
- Spacing and capitalisation should also be considered.
- Synonyms – including as many synonyms as possible can help to reduce the issues associated with only entering one.

- More tags - Entering a larger number of tags may help to improve search efficiency and actually allow more focused searches if necessary, preventing the problems caused by differing granularities of search.

The use of stemming in tagging systems was also rejected as it may actually hinder the search operation rather than provide benefit. The above issues are problems that should at least be considered by tagging systems in order to ensure that they are used efficiently and effectively and whilst many systems offer solutions to some of the problems, they must all be a consideration for the development of a new system. They should also be the focus of any organisation aiming to implement tagging as a potential solution to aid in the discovery of information.

Systems such as Del.icio.us already offer a solution to a number of these problems. The first thing Del.icio.us offers is a number of recommended tags, this helps to re-use tags that have already been used before. The system also shows the user how many times they have used a tag before, as they are adding a tag to an item. This is an extremely useful tool. Along with this, Del.icio.us also makes the recommendation that users do not use spaces at all within tags.

Given these findings a questionnaire was developed to help determine which of the features of tagging users or an organisation understand, and which areas training should be given in, this is discussed further in section 6.5. Full tagging does appear to offer an interesting solution to the problem of finding information, and this is especially true for documents that would be hard to discover with a full text search, such as photos and videos.

### **6.3 Could Tagging be applied to a Traditional File System?**

Tagging offers a very lightweight alternative to traditional structured storage systems. This is often the key reason attributed to the success of tagging in recent years. One of the most obvious hierarchical structures is the traditional file system used to store files on personal computers. The current directory structure found on many computers, along with those used by shared document systems, require a user to place a file into a folder, the file is then retrieved by navigating to the same

folder. The literature review has already shown the pitfalls of this using Golder and Huberman's (2006) example as follows:

"Consider a hypothetical researcher who downloads an article about cat species native to Africa. If the researcher wanted to organize all her downloaded articles in a hierarchy of folders, there are several hypothetical options, of which we consider four:

1. articles\cats - all articles on cats
2. articles\afrika - all articles on Africa
3. articles\afrika\cats - all articles on African cats
4. articles\cats\afrika - all articles on cats from Africa"

Placing files into a system whereby the user was guaranteed to retrieve the file that they wished to find would be extremely difficult. The user may even have to add the file a number of different times in different directories to ensure that it was found. Whilst one possible solution is to make use of one of the existing corporate search systems already discussed another possible solution to this problem would be to replace the file system with one that makes use of tagging. If tagging were to be used the user is free from the hierarchical structure and is enabled to make use of a tagging system instead. Using a tagging based system would allow the user to simply tag their files and then retrieve the file using these tags. It would remove the need to place files into multiple folders or even worry about where the files were stored because the tags would be used to retrieve the file and not just the location that they are stored. The system would also function for files that did not lend themselves to full text search such as images and videos.

#### **6.4 Proof of Concept: Building a Tag Based File System**

Given the potential benefits that could be found in tagging based systems and the issues currently associated with finding documents, the development of a file system based upon tagging was undertaken. The application of tagging to a standard operating file system had never been undertaken before but would allow users to make use of a tag oriented file store from within their common workflows, and without the need to resort to using a web-based search system.

Within Chapter 4, "Assessing the Knowledge Sharing Environment" the daily processes section of the questionnaire specifically asked questions relating to the current working processes of the employees of PharmaCo and SoftwareCo. Within both organisations it was found that employees made extensive use of the Microsoft Office suite and also that the departmental portals within the organisations were rarely utilised. Given these two findings the recommendations stated that the portal might not represent a prominent enough place to put a possible system. However, integration with Microsoft Office would be essential as the office suite was in such constant use within both PharmaCo and SoftwareCo. Having access to the files and benefits of the tagging system from within the employees' applications would be a significant advantage, making the tool well integrated into the employees work process.

A server was developed that would allow users to mount the folders of the server as if they were folders within the computer, as traditional Network Attached Storage (NAS) devices would but used a tag oriented navigation method rather than the traditional hierarchical approach. The benefit of integrating with the operating system is that the files will appear as if they are stored on the local computer. This removes the need for the user to open a web browser and perform a search. It would also allow the user to access these files from inside any application that was running on the computer instead of having to download the file first. The ultimate benefit is that the files would appear to be stored on the computer like any other file in an attempt to allow users to improve their ability to find relevant information whilst remaining in a familiar environment.

The server was developed in Ruby on Rails, due to the familiarity of the author with this programming framework. The server was used to create a virtual file system. The server would not retrieve files directly from a given directory but would work in conjunction with a database. Storing the files in conjunction with a database allows additional information to be stored with each file such as the tags that were used to tag that file. As the tags were also all stored in a database, it would be a very quick process to create a list of all of the tags used by files stored within the system making the system extremely responsive and prevent the delays that are associated with alternative systems such as searching the full text of a

document. The system would also have a much lower storage requirement than traditional search based approaches as indexes would not be required in this method.

The server would not store multiple versions of the files but would simply store one version of the file and make use of the database to allow the server to generate a virtual structure that is sent to the client. The database would maintain a list of files and the tags associated with that file. Each of the tags would then appear as a 'virtual directory' within the root folder. When the user chose that directory, the tag would be used to find all files that contained that tag and display those files. Along with those files, all of the other tags that were used by the files shown would also be present as directories to allow the user to further refine their search and narrow the list of files. This enables the user to choose as many tags as they like until the files that have all of these tags are listed within the directory. The files would be refined based upon the tags that are associated with those files rather than where they had been stored. Although to the user it would appear they were navigating a list of files they were actually just communicating with the server that was dynamically creating the folders they would see. An additional benefit of this method is that users may browse for the files that they wish to find, refining their search by clicking on directories, or tags, as they went. This is a strong contrast to search based approaches where search terms must be entered before any results are displayed. The system also allows the user to quickly narrow down their search using the directories to represent tags when they are searching for something specific. Instead of using a pre-defined directory structure the directory structure is automatically created based upon the tags of the file. Although the user is essentially browsing a directory structure, the directory structure is continually narrowing down the list of files that contain those tags to allow the user to search for the file.

The author named the system TagDav meaning Tag-based Distributed Authoring and Versioning. TagDav was used as a proof of concept to show how a tag based file system may work. As an example, three files were uploaded into and where named as shown in Table 22. The table also shows that tags that were applied to these files. The files in this table were created and placed into the TagDav system

and used as an example of how the files would appear with their respective tags to the user.

**Table 22 - Example files in TagDav**

| Name                     | Tags              |
|--------------------------|-------------------|
| File_Tagged_One_and_Both | tag_one, tag_both |
| File_Tagged_Two_and_Both | tag_two, tag_both |
| File_Tagged_One_Only     | tag_one           |

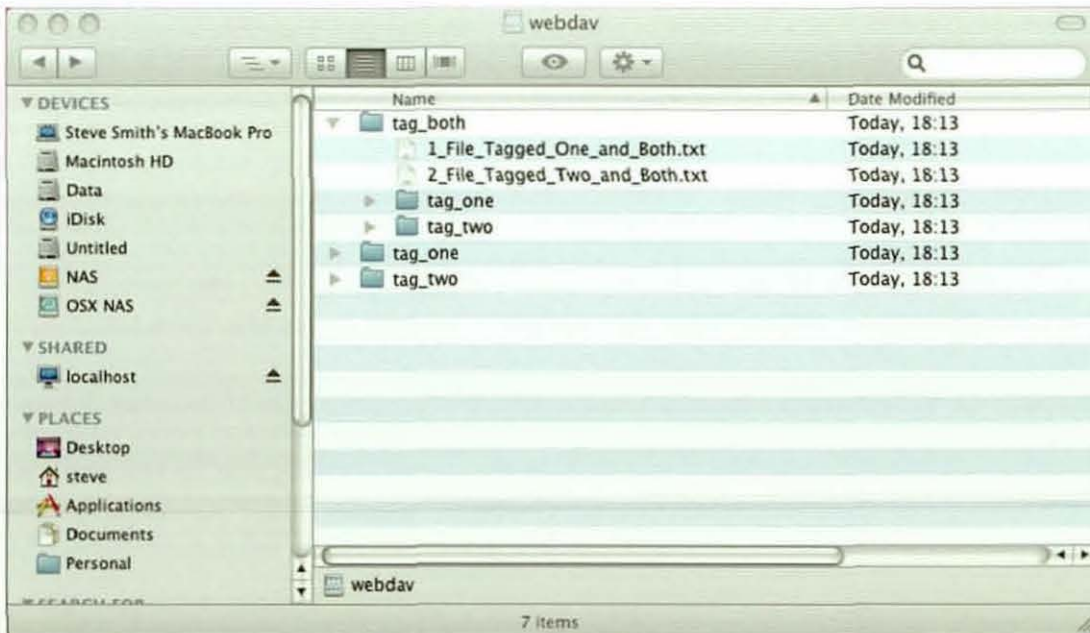
Figure 25 shows the TagDav system in the root folder. All of the tags used are listed so that the user may begin their search. Although it may seem like there would be an issue if there were thousands of tags entered into the system this would not be the case. The tags would be ordered alphabetically and could quickly be found. In many operating systems simply beginning to type the directory name shall highlight that directory and in this case the tag. Once one tag had been chosen, the list of remaining tags would be significantly reduced. The process could continue until the desired file was found. This root folder also has the option to show all of the files that exist in the system but for simplicity it is not shown in Figure 25. The tags from Table 22 can be seen as directories within the file system that contain the files.





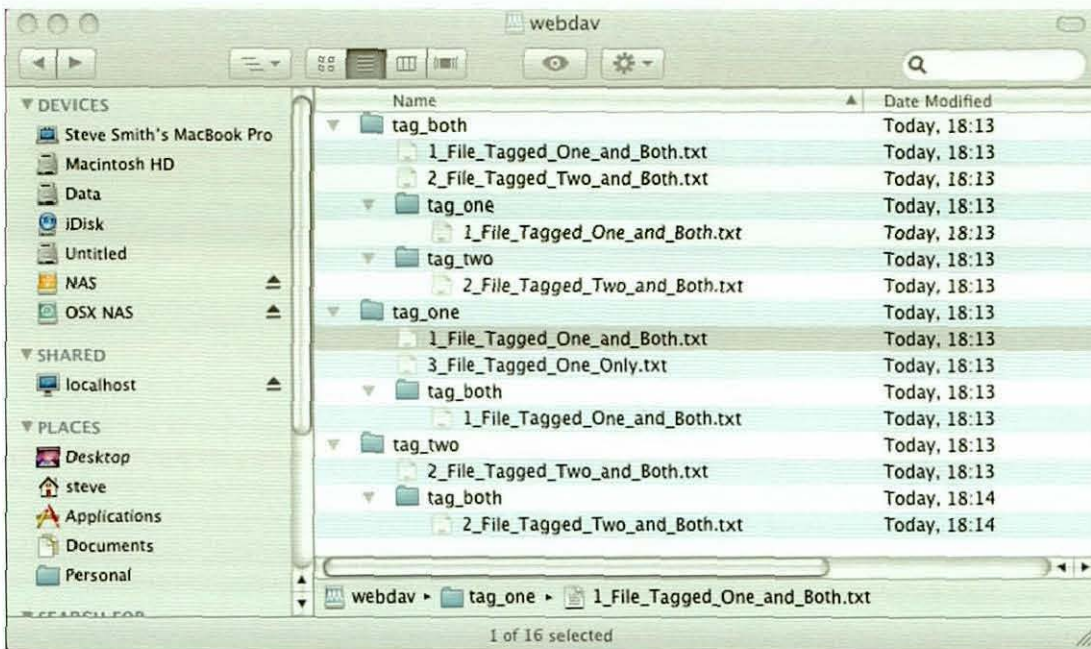
**Figure 25 - TagDav with closed folders**

It is then possible to expand and navigate within the folders to see the files that are tagged by that content. Figure 26 shows the tag 'tag\_both' expanded as a folder showing the two files within this folder and also the two tags that those files are also tagged with.



**Figure 26 - TagDav with one folder or tag expanded**

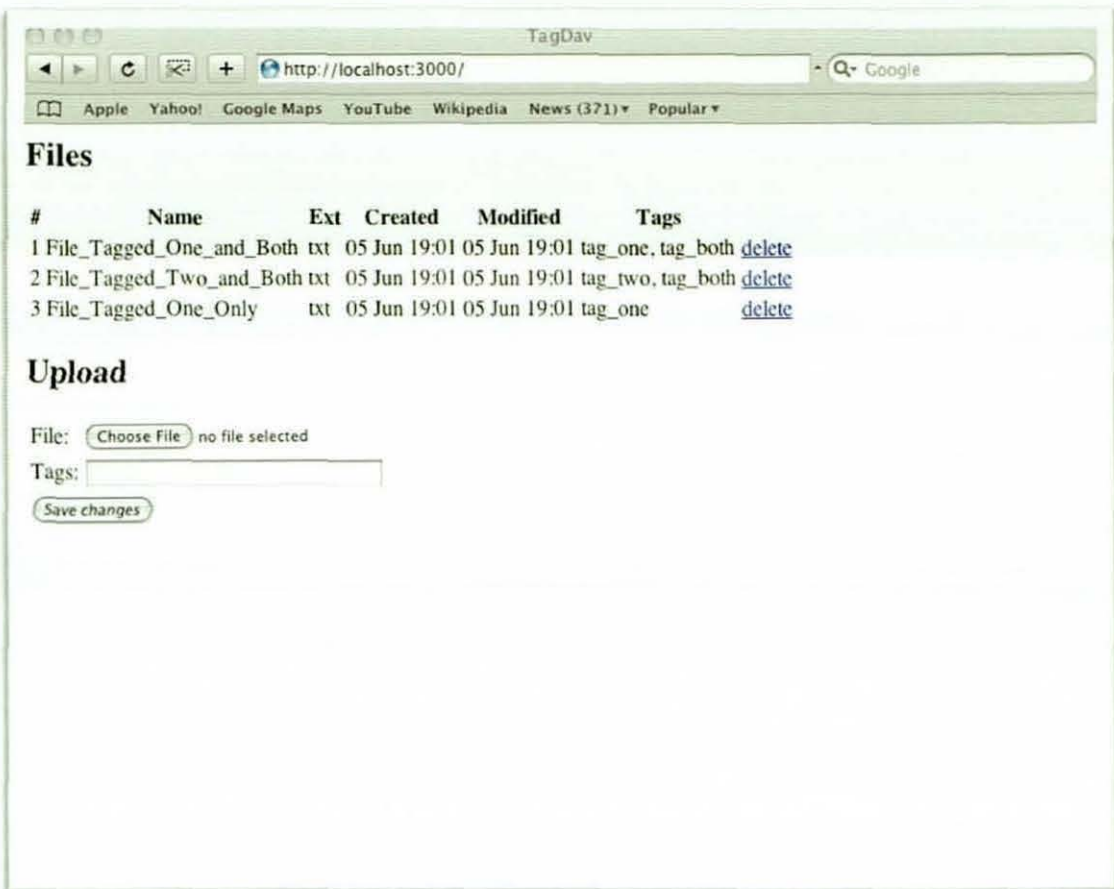
The tags may be combined to further refine the search within the folders. Figure 27 shows the full expansion of the example files and their folders.



**Figure 27 – Fully expanded view**

Although in Figure 27 there are 9 files listed there only actually three physical files stored within the system. Each file is only stored within the system once and each of the files shown in Figure 27 is actually a virtual reference to a file that will be retrieved at run time should the user wish to actually download or open this file. In order to work with this file, the user simply has to click the file as they would if the file were in any normal file system.

The current system does not allow the user to create files from within the file view shown in Figure 27. In order to add a file to the system the user has to visit a web page and upload the file. This was for two reasons with the first being purely for simplicity in this proof of concept. The second reason is that using the web front end allows greater flexibility in the interface and allows restrictions to be imposed in the tags that can be created. Figure 28 shows the basic upload interface that allows users to add files to the system.



**Figure 28 – The upload interface of TagDav**

The upload interface does not take note of the recommendations created earlier as this system was only design to be a proof of concept for the retrieval system. The system was not designed to showcase the upload procedure. However, any production system could make use of the recommendations. Taking these recommendations into account would allow more effective tags to be added to the files. Integrating systems to enforce the recommendations made would be a simple process when a production system is created. For example, in order to prevent plural words the system could check the tags and highlight any words that were not singular forms. Similarly it would be quite simply to inform the user how many times a tag had been used so that they could try to use tags that had already been used frequently previously. Using tags that have been used previously makes files easier for other users to find. Re-using tags means that more files shall have those tags and thus more files shall appear for each tag. However, if enough tags are used to tag each item then refining the search in order to find the file that the user is

looking for will be far easier and the user is more likely to find the file they are looking for.

A further example using Golder and Huberman's (2006) example is shown in Figure 29. The figure shows one file called "Big Cats.txt" that has been tagged with the tags "article", "cat" and "Africa". The file appears to have been replicated several times but in fact each link is merely a reference to that file categorised by the tags. All of the tags or directories in this case have been expanded to show every location where the file is located. It is important to reiterate that although the file appears in this list multiple times only one copy is present on the server and the directories are being dynamically generated.



Figure 29 – TagDav with Golder and Huberman's example of file structure

## 6.5 The Application of Tagging in a Business Environment

The use of tagging has apparent benefits for organisations and the TagDav system offers a potential solution to help improve the information overload problem within organisations.

This section investigates how tagging might be useful within the business context for the discovery of relevant information stored in a user's own and group file stores. The section shows how the concept of a tag based file store could be used by an organisation and the assessment, through a focus group and questionnaires,

within one of the case-study organisations, SoftwareCo. However before tagging can be introduced, an organisation must assess its potential barriers to the use of tagging in general.

Before the focus group and questionnaires took place an initial questionnaire was given to the respondents to assess any barriers that they might have to tagging. The concept with the initial questionnaire was to very quickly determine where the respondents might need training in the use of tags and tagging in order that they might make effective use of a tagging system. The assessment would give an idea of whether such a questionnaire might aid an organisation to establish where training may be needed before the employees of the organisation could comfortably make use of tagging. This initial questionnaire can be found in the appendix in section 12.

Following the initial questionnaire the respondents could see the TagDav tag based file system and an assessment of its potential within the organisation could be made.

In order to test the tagging system two methods were used, the first was a focus group. A focus group was used because it allows the group to focus on a specific area, namely tagging and the TagDav system. A focus group also allows the open discussions and participants to discuss a large number of aspects that are relevant to them as a group. It also enables the group to talk about subjects that might not be approached should they be within a questionnaire because the author of the questionnaire may not think of the issues raised by participants. The second method that was used was a questionnaire. The questionnaire can be found in the appendix in section 12. This questionnaire was administered in addition to the initial questionnaire relating to the potential pitfalls of tagging. Deploying a questionnaire allows a large number of questions to be answered in a short space of time and allows some anonymous responses to be included. The questionnaire was self administered, although the author was present to allow participants to ask questions should they need to. In particular, the aim was to determine if using a tag based system for document retrieval could help employees to discover relevant information.

The assessment took a hybrid approach of a focus group and questionnaires that was similar to but didn't follow a Delphi style approach. The approach was chosen because of the difficulties and tradeoffs that had to be made between the number of champions and people available to take part in the investigation, the time that they could allow and the fact that transcription or recording was not an option. The users were free to use the questionnaires and a SWOT analysis to record their thoughts anonymously whilst the focus groups allowed elaboration and discussion to help understand the true feelings and thoughts of the group collaboratively. Although transcription or recording was not allowed by the organisation a number of notes and 'sound bites' were taken that were approved by the participants.

The format of the assessment including both the initial questionnaire that would be used to assess the need for training in the field of tagging within an organisation and the focus group and questionnaire that followed. The format of the assessment can be seen in Figure 30



**Figure 30 – The assessment format**

The focus group consisted of 9 members from the SoftwareCo. As stated in previous sections, SoftwareCo is one of the largest software organisations in the world, within the top 10 of all major software rankings. It had over 50,000 employees in over 50 countries and designs software solutions used throughout the world. SoftwareCo is seen as one of the leaders in its field and the members of the focus group that also answered the questionnaire were all members of the value prototyping and rapid development department of that organisation. This department houses some of the company's most technically competent employees and has some of the most respected employees within their fields.

The focus group members were selected by one of the champions within the organisation, who selected on the criteria of someone who would give valuable input and would be interested in the application of the system. The participants were also chosen to include the most varied views including those who would be

expected to be advocates of such a system and those expected to be against it. Each member within the focus group was involved in the development of software, and has some connection or interest in semantic technologies. The members were also selected so that some of them were familiar with tagging and highly involved with the development of semantic applications whilst others were unfamiliar with tagging. Being unfamiliar with tagging would mean that the concept would be new to them but their ideas would still be valuable as they brought a different perspective. Some members also came from the field of search and document retrieval.

The size of focus group is an important factor (Bryman, Bell 2003). It was decided that nine employees would be invited. Although nine participants is at the upper limit recommended for a focus group, there were a number of reasons for choosing this number. Nine participants were chosen in order to give an even spread of ability and thus a greater variety of opinion with regards to the system. It was not possible to record or transcribe the focus groups due to privacy concerns within SoftwareCo. As a transcription of the focus group discussion was not a possibility at the end of the focus group users were given a matrix on which to record any notes and sound bites that they were comfortable sharing, in addition several sound bites were recorded by the researcher in order to capture the key points made within the focus group. In addition because transcription was not possible the respondents were given a matrix on which they could record their thoughts within four key areas namely the strengths, weaknesses, opportunities and potential risks that might be introduced by the use of the tag based approach to document recovery.

Two questionnaires were used, during stages one and three in Figure 30, to gather responses to questions relating to two key areas. The first (stage one Figure 30) was used to try and assess any potential barriers to the use of tagging and investigated the field of tagging in general. It would be used to determine whether any training in the use of tagging might be needed before tagging could be implemented within the organisation. This first questionnaire (See Appendix in section 12) asked questions relating to tagging and how the user currently tagged content, if they tagged content at all. The aim of these questions was to identify

how the issues with tagging previously identified affected the users of the organisation. If the answers showed that the issues associated with tagging were present then it would show that training would be required before users would properly use a tagging system and where any training would be best focused.

The questions in this questionnaire were predominantly derived from the issues with tagging identified earlier in this chapter in section 6.2.2. The questions began asking users if they had used sites that make heavy use of tagging such as Del.icio.us and Flickr. The questionnaire (See Appendix in Section 12), which contained eight questions, then asked questions such as the number of tags used and how frequently they re-used tags.

The second questionnaire, stage three in Figure 30, (See Appendix in Section 12) related to the TagDav system itself and was used to help record the opinions of the participants of the focus group as transcription was not an option. This second questionnaire contained 16 questions that questions focused on the way that users currently store and retrieve files and then asked some generic questions about the TagDav system they had seen regarding how easy it is to use and whether they felt that training would be necessary in order to use the system effectively. The aim of these questions was to establish if the participants had a common system or method of working or if they all used different filing schemes. These questions would also show how many of the users currently resorted to using search engines and duplicating files in order to retrieve them more effectively. Then the questions asked how the TagDav system and tagging approach might improve searching in the participant's working environment. They were asked questions such as, could this system replace their current method of filing and ultimately how much time did participants feel that the system might save them, how it could be of benefit to them along with its disadvantages.

While the focus group (shown in section four of Figure 30) discussions took place, as mentioned previously the participants were given the opportunity to record any notes and sound bites that they were comfortable sharing and in addition, focus group participants recorded the perceived strengths and weaknesses then opportunities and threats of the system (shown in section five of Figure 30). These



were really given to prompt the users to give more details relating to what had been discussed within the focus group that they were comfortable recording. These four features were described to the participants as follows:

- Strengths - The advantages of the tool (i.e. what does the tool do well?)
- Weaknesses - The disadvantages of the tool (i.e. what does the tool do badly?)
- Opportunities - The advantages of the tool to the organisation (i.e. how will the listed strengths benefit you and/or the organisation)
- Threats or Risks - The disadvantages of the tool to the organisation (i.e. how will the listed strengths or weaknesses threaten you and/or the organisation)

The features were used to record the thoughts and opinions of the participants and to record the topics that had been discussed during the focus group in place of transcription, due to the privacy fears of the participants.

### **6.5.1 Results**

#### **6.5.1.1 Determining the Issues associated with Tagging**

As the literature review had suggested might be the case the results showed a large variance when it came to tagging. Four out of the nine employees had never tagged content before. The selected mixed group would add value to the results as it would also test the usefulness of a tag based file system to those who were technically minded but had not yet experienced tagging. Of the five members of the group that had tagged content before, there was a wide variance in the answers given to the initial questionnaire regarding how they use tags. This section shall be based upon the five users that had previously tagged content.

When asked how many tags they use on average to tag content, the employees were quite close to the average discussed in the literature review, which showed an average of 3.3 tags per user even for users that had tagged a semantic web conference website. The literature found that one participant (20%) used just 1-2 tags and only 3-4 tags were used by the other four participants (80%). This highlights that, as found in the literature review, many users do not use as many tags as they perhaps should. Two (40%) re-used them always and two (40%)

sometimes, with one participant (20%) rarely re-using existing tags. This was worse when the tags were those created by another person. One participant (20%) always re-used tags used by others, one (20%) sometimes re-used them, and the other three (60%) rarely re-used the tags (20%) used a mixture and one (20%) used plural. Again there was no consensus over spaces, with two users (40%) stating that they used them, three (60%) stating that they did not and the other using them and some not. Only one participant (20%) used synonyms of a word when creating tags in order to help them to retrieve the content in future.

During the focus group, participants discussed that, in the most part, they were not aware of the issues associated with tagging. Many of the participants had only thought of tagging as being of use for them personally and not considered that other people would be using these tags to retrieve content. The participants also commented stating that many different systems ask users to tag differently or do not give any guidelines as to how tags should be entered into the system. All participants agreed that with the correct training and if all participants were to make use of the same tagging scheme, then the benefits of tagging would be far greater.

The results highlight the differences in the way that users perform tagging and also that with a small amount of training the barriers towards tagging can be overcome. The questionnaire helped to identify the way that users were currently tagging and through discussions during the focus group it was possible to expand upon the potential issues with tagging. One of the group members stated "although there are barriers towards tagging if they can be overcome then this could actually lead to increased consistency, ease of finding docs/data" with another user stating that "after training tagging might present a powerful option but that they were not aware of all of the potential problems"

These issues really highlight the need for training but also for a system to remind users of the benefits of considering these options when tagging systems are in place. It is only a matter of training users in the correct use of tags in order to improve their use. Tagging existed on the Internet as a very simple method and as such very little training or guidance has been provided. During the discussions

most members of the focus group were extremely interested in the issues associated with tagging and agreed that training was necessary for these systems to be successful.

Through the use of the questionnaire presented in this section organisations can determine where training might be required with regards to the use of tagging. This training can lead to a more consistent approach to tagging, allowing tagging to become a far better solution within an organisation attempting to overcome the difficulties in finding the correct information.

With the barriers towards tagging understood the assessment of the Tag based file system took place and is detailed in the following section.

#### **6.5.1.2 Assessing Search and TagDav**

In order to assess the potential of the TagDav system to help users obtain relevant information the second a questionnaire was given to the group of nine SoftwareCo employees. Once the members had answered the questions in the questionnaire a group discussion was held centred around the benefits and weaknesses of the system and any potential it may have. Participants were also reminded that this system was merely a proof on concept not a fully functioning production ready system.

The questionnaire contained some preliminary questions. When asked if they have difficulty finding files they have created because they do not know which directory they have been placed in, 33% of participants stated often, 44% sometimes had a problem but 22% of participants rarely experienced a problem. Of the participants 22% stated that they always resorted to using the search facility of a file system to find files, 22% often did and 44% sometimes resorted to using the search facility, with 11% never using the search functionality. Of the 8 participants that did use the search facility only one (13%) felt that it was rarely sufficient to find the files that they needed to find.

One (11%) of the participants always placed the same file in more than one location or created symbolic links in order to make it easier to find. Three of the participants (33%) sometimes did and one (11%) rarely did, with four (44%)

participants never doing this. This does highlight that 55% of the participants at some point have created a duplicate file in order to ensure it can be found in future. Copying a file to another location does carry a high level of unnecessary risk, if for example, the file is updated. Creating symbolic links is better practice but should not be necessary in the TagDav system. When participants were asked how they store their files every single response was different. Some stored files by project, some by topic, some by year and some used completely different schemes entirely. With all of these different storage methods, a collaborative file store could easily become quite difficult to use. There would be duplication of files and it could be difficult to find a file if it is not where it is expected to be stored. The responses highlight that there are a number of issues associated with traditional file systems and storage of documents.

The following questions and responses relate to the TagDav system and its potential to improve upon these issues. After using the system, participants were asked whether they felt that tagging files would help them to find the files and prevent the need for them to have to resort to using a search engines. Of the participants 67% said that they definitely felt that the system would help them to find the files they were looking for without having to resort to the use of a search engine and 11% stated that it often would. Two of the participants (22%) felt that the system could help sometimes. When asked whether tagging files would save time when it came to the retrieval of files, 78% felt that it definitely could and 22% of the participants felt that it would often be possible.

The participants also stated that they would use this system to replace their current way of filing and retrieval. Two of the participants (22%) said that they would definitely use the system, 44% stated that they would make the switch and 22% stated that they would use the system sometimes.

All of the participants felt that the system had benefits over traditional directory based systems and all but one (11%) of the participants felt that browsing a file structure was harder than uploading a file and tagging it. Therefore they would have no objection to the initial issue of uploading and tagging a file.

Participants were also asked whether they felt they could save time using this system and all participants agreed they could. When they were asked how long they felt that they could save each day when using this system, participants were allowed to enter any value of their choice, as no pre-defined categories were offered. Although answers ranged from 10 minutes to 120 minutes, on average participants felt that they could save 40 minutes per day through using the system. This is an extremely impressive number, especially when it is considered that some of the participants had not used tagging before they had seen the system.

One important factor to note is that all of the participants felt that training would be required before it was used. Although most of the participants (56%) felt it would only be necessary sometimes, three (33%) of the participants felt training would always be required and one (11%) felt that it would often be required before the system could be used.

Although participants used the questionnaire because the focus group could not be recorded or transcribed the participants were also asked to make notes where possible and fill in the chart containing what they felt the strengths, weaknesses, opportunities and risks or threats posed by the system were. A number of strengths, weaknesses, opportunities and threats were given and are as follows. Most of the participants felt impressed by the way the system allowed them to find a document. The system was said to be very easy to use and was praised for the ability to allow participants to find files from any number of approaches or viewpoints "files can be found from a number of different approaches and multiple file representations will appear making it easier to drill down and find documents". One participant also stated that it would allow users to "reach consensus about file system structure" the users also appreciated the ability to create a "reduction in storage space required" as files would not need to be saved in multiple locations. A number of interesting weaknesses and worries were also raised. One of the worries was that a file with no tags would be "lost in the system" and that if enough tags were not used then the file may be difficult to find. The key thought from participants was that the system would have to be used with a lot of care and "taken seriously" from the start and if it were not then there would be a risk of losing files or having an "unstructured mess". Although the participants felt

that training would certainly help there were still worries about the consistency that would occur if everyone using the system did not fully understand the potential of the system and how to correctly structure files. The other worry was that some form of consistency would be required in order to ensure that the tags used by some users would be retrievable by others. Some participants also felt that once there were too many tags it might be difficult to drill down and find the files that were required. However, as a group it was felt that this was "more of a benefit than a disadvantage, normally you would not have the option of drilling down further". One of the participants also raised the issue of the time it would take to store the documents however as a group they were not concerned about the amount of time it would take to store the documents as in many cases "it takes ages to find the directory to save something in now".

*With regards to the actual system, it must be highlighted that a minority of the participants were worried about the implications of using a server based system for storing files, but the implications of this would have to be addressed for any production system and group based storage systems were already in place in the organisation and shared the same concerns.*

## **6.6 Conclusions**

The research has developed and identified a proof of concept for a new tag based file system that has not been tried before. Information overload is a significant problem for business in today's working environment. The system took the concept of tagging, that has been used traditionally and more recently on the Internet and made it accessible to a users everyday work process.

*Documents stored on a computer can often be extremely difficult to find and participants are forced to seriously consider where they place these files in order to be able to retrieve them in future. When files are stored for groups rather than individuals the problem can potentially worsen. Users use a wide variety of different filing methods to store their files so that they can be retrieved in future. These methods include storing them by content type, project, date customer and many others.*

This research has shown that tagging can potentially benefit an organisation. Tagging has seen rapid growth on the Internet in recent years and offers an alternative to traditional hierarchical structures and full text search systems. This chapter proposed the use of a tag based file system, named TagDav, to replace traditional hierarchical file structures that could be used by both individuals and groups.

The results show that before tagging can be of benefit to an organisation the potential issues identified by tagging must be identified. The research has shown a simple questionnaire can be used to identify the extent to which employees require tagging and where training may be required in order to promote more effective collaborative tagging.

The system potentially provides a way of reducing duplication of information and time spent finding expert information. If a user cannot find a file, then they may have to re-create the file that they were looking for. This would come at a considerable cost to the organisation. This cost comes in addition to the time spent searching for a file that the user knows exists but cannot find.

As identified by the literature review finding more relevant information represents a significant challenge and a system such as this has shown potential to aid users in the discovery of relevant information and lower the information overload problem.

## 7 Ontology Development

### 7.1 Preface

This chapter introduces OntoFarm, a tool created to aid the development of ontologies. This chapter satisfies objectives 5 “To develop and assess alternative approaches to storing information to improve information retrieval and reduce information overload.”, 6 “To establish the role ontologies can play in the retrieval of relevant information and reduction of information overload, the complexities of ontology development and the barriers to their use.” and 7 “Investigate alternative approaches to traditional ontology development tools that may be used by subject experts rather than ontology specialists to aid in the creation of ontologies that can help the discovery of relevant information.”. This chapter begins by looking at issues involved with ontology development. The chapter then outlines a number of requirements for creating an ontology and shows how the concept of harvesting information can be useful to ontology development. Finally, the development of the OntoFarm tool is shown and its ability to aid the development of ontologies is assessed. Ontologies represent a more structured format for storing information that if harnessed correctly could greatly improve the storage and discovery of relevant information. Ontologies offer great potential for the reduction of the information overload problem.

### 7.2 Introduction

Chapters 5 and 6, in addition to the literature review in Chapter 2, have already described the issues faced by employees when trying to find relevant information. Information overload is making it increasingly difficult to find the information that is relevant to employees to perform their jobs. Improving access to relevant information and helping employees to filter out the irrelevant information can help to reduce the problem of information overload as there is a reduction of irrelevant information, the key cause of information overload.

Ontologies are often used on the World Wide Web to help in the retrieval of information. Ontologies can range from simple taxonomies used for categorisation



to more complex ontologies containing numerous types of relationships and links. Ontologies offer a wide range of possibilities and offer a machine readable format for storing information. They can also be used in conjunction with traditional searching in order to aid the search process. (Noy, McGuinness 2001).

“The use of ontologies to overcome the limitations of keyword-based search has been put forward as one of the motivations of the Semantic Web since its emergence in the late 1990s” (Vallet, Fernández & Castells). Ontologies have many different applications in the domain of search and semantic technology and their use can help to provide structure and understanding to query based searching.

During the literature review, the concept of ontologies was introduced and some of the difficulties in creating them were highlighted. These difficulties included the fact that ontologies must be designed rather than created (Gruber 1995) and the significant costs involved with ontology creation. Due to the difficult nature of ontology development, it comes with great risk and expense. This is true even when development is done correctly. The literature review highlighted how one project, the Gene Ontology, had cost at least an estimated \$16million by 2006 (Good et al. 2006). The literature review also showed a number of limited approaches that attempted to make use of technologies such as wiki's to create ontologies. This method was given the title of creation by 'proxy' as the creation of the ontology was not the primary aim of the tool. Even with tools such as this, ontology development is still difficult. This method may even make things more difficult as the people editing the wiki may need more training to create a meaningful ontology from a wiki article. Ontology development tools such as Protégé and TopBraid composer help ontology development but it is still a complicated process.

Although these tools are very good at allowing the user to actually model an ontology, they make no effort to help the user with the task of determining the content that needs to be added to the ontology. There are many disadvantages of the tools that already exist and in many cases it was felt that the tools required the users to have an extremely deep knowledge of their ontology already. The tools were extremely capable of modelling the ontology but very little effort was made

by the system to suggest concepts that already existed or to help them to find something that was already within the tool. The section entitled "Appendix Four – OntoFarm Questionnaires" in the appendix shows details relating to some of these tools.

One of the most powerful examples of ontology searching comes from Powerset. Powerset provides semantic meaning to content from Wikipedia and freebase and allows users to search in a more effective way. Figure 44 shows a Powerset search for the people who married Henry VIII.



Figure 31 - Powerset "who married Henry VIII"

In addition to providing this intelligent search, ontologies allow representations of concepts to be linked and inferences to be made from these links. For example, if searching for an employee within an organisation that is good at Java programming, a user may search for the term Java. The search for java might include any members of the organisation who included the word java in their profiles. However, what about employees with J2EE, or JDK or even object oriented programming languages within their profile? Through the use of an ontology more understanding and meaning can be given to a search and more effective results obtained.

One of the key problems identified with the tag based file system in Chapter 5 was the fact that different users may use different tags to describe related content. One suggestion to combat this was the use of ontologies. The combination of both tagging and ontologies can provide a hybrid approach taking benefits of both approaches to create "a user-friendly system that encourages collaboration and makes information easier to find." (Barbosa 2008)

In order to take advantage of the benefits afforded by ontologies, SoftwareCo decided that it would create a business specific ontology along with a number of other related ontologies. However, the creation of ontologies is not a simple task and requires a significant amount of thought from its creators. Given the extremely high cost and difficulties often associated with ontology development, it was decided that a tool should be created to help ontology masters design an ontology. If a tool could be developed that offered help when creating and maintaining ontologies, significant cost savings could be found by the organisation. This chapter shows the requirements and process that led to the development of an ontology development tool for use within SoftwareCo. The chapter shall then show how the tool was developed and its features and then assess its ability to aid in ontology developments.

### 7.3 Understanding the Requirements

To gain a deeper understanding of ontology development a focus group was formed. A focus group was used because the focus group method enables an understanding of how all of the members regard ontology development and the requirements of a system.

Ten members were selected for this focus group comprising of employees and contractors of SoftwareCo. The members of the focus group were all members of the rapid prototyping and development department of SoftwareCo. The focus group contained ten members who all had an understanding and work-related interest in semantic technology. The focus group participants had a mixed understanding of ontologies although all were familiar with them and all had knowledge of OWL. Although some of the members of the focus group were contractors rather than employees they shall be regarded as employees for the purposes of this work.

Some participants rated their knowledge of ontologies highly, having worked with them quite intensively. Other members had a very limited awareness of using and developing ontologies. Three users (30%) stated they had a strong awareness of ontologies and ontology development, three (30%) had a fair awareness and four

(40%) had a low awareness. None of the users stated that they didn't have an awareness of ontologies.

The participants of the focus group were also asked a selection of questions (See Appendix Four – OntoFarm Questionnaires) regarding the difficulties associated with ontologies. The answers were on a scale from defiantly, maybe, rarely and never. When asked if they felt ontologies could help to retrieve information more effectively, 70% said definitely and 30% stated maybe. There was clear feeling that ontologies could be of use within the organisation, especially within the field of information retrieval. The main barriers identified by the employees to creating ontologies centred on a number of factors. The first and foremost was a lack of time or money, and in many cases these were treated as the same thing. The next key answer was a lack of knowledge or expertise and finally a lack of tools was mentioned. Another extremely interesting observation was ontologies are "often purpose specific, but that purpose can change". This comment highlights that ontologies need to be adaptable and easily maintained.

Of the ten users, eight (80%) felt that having one user to create and maintain an ontology was a problem and that a system to allow different users to concurrently add to an ontology was more desirable. Amongst other reasons the problem of one person maintaining the ontology was related to the time that it takes for a user to create an ontology and that more than one area of the organisation may need to update the ontology. It was inferred from these discussions that collaboration would be necessary. Only one user felt that having one user create or maintain the ontology was rarely a problem and one user said it was never a problem.

There was a general feeling (80%) that having a lack of time to create an ontology was also a problem, but it was worth the effort. The time that it takes to discover the concepts that should be added and modelled in the ontology was also seen as a problem with 66% of the users agreeing with this statement. There was also a feeling that users needed training before they could create an ontology.

The participants were then asked to explain any issues they may have with ontology development. Along with suggestions that time and money cause problems, there is also a lack of adequate tools to create an ontology. One

participant raised the issue that "Ontologies impose a view with little flexibility" and that view often changes. It is likely that employees will view the ontology differently and coming up with one definitive view is difficult. Other users noted that it is extremely difficult to make ontological commitments from the beginning of the development process, especially as the structure will inevitably change.

Given the above findings a number of conclusions can be made. Firstly, employees feel that ontologies can be of use to the organisation. Secondly, the lack of time or money devoted to ontology development causes a problem and thus any tool developed should aim to save time and make ontology development easier for employees. The tool should also help users without them having to have a great understanding of a specific tool. The ontology development tool should also allow an ontology to adapt and change as required. One of the key findings was that the tool should be collaborative and allow more than one user to use the tool at the same time.

#### **7.4 Development of the OntoFarm System**

Given the difficulties in ontology development SoftwareCo employed a philosopher who was charged with the creation of the ontology. It was decided by the organisation that some additional restrictions on the system would be necessary in order to abstract the difficulties that might exist when creating ontologies.

A number of things were done to perform this abstraction. Firstly, there is no scope within the system to create instances. This system was designed in order to allow the quick creation of a framework that classes could be placed into.

Classes and instances in ontologies are extremely similar to those within object-oriented programming. A class represents a type of object that can be used to describe many different actual occurrences of that class that are called instances. Since ontologies model the world around them, instances within an ontology are often representations of physical objects that the ontology is modelling. As an example, a class called 'car' could be created, the physical car with registration plate "ab01 cde" would be an instance of the class car. The 'car' class could be

further refined to be a class called `Aston_Martin`, describing all of the cars created by the company Aston Martin.

The classes form the basis of the ontology, and other systems would be capable of both using these classes and also creating instances if necessary. However, the manual task of bootstrapping and creating the ontology would only initially involve the creation of classes and the relationships between these classes.

The relationships that could be created were also restricted. An overall ontology master could define properties but the individuals working with the ontology on a daily basis were restricted to the properties that were already defined. There was much debate over the properties that would be included and in this instance it was decided that only two relationships would exist. The first would be an 'is a' relationship. This 'is a' relationship would be expressed using the RDFS 'subClassOf' attribute. The other relationship that would exist would be a 'related to' property that would simply allow the ontology creator to state that two things were related. There were also discussions as to whether the 'part\_of' relationship should be included during this initial phase but at the time of writing it has still not been included in the version of the system used by the SoftwareCo. Different organisations could of course add any number of properties using the proposed system.

The system also had to be modular. In some cases only certain parts of the ontology would be necessary and at other times the entire ontology would be needed. Fortunately the concept of namespaces is a frequent one in ontology development with RDF and OWL both allowing namespaces to be incorporated and different files to be imported by an ontology. This allows the different namespaces to each be added to a different file and imported as necessary.

Another important observation was that a class or concept might have many different lexical representations and that each of these lexical representations could describe the same concept. In order to simplify this, it was decided that lexical representations would be handled separately by the system and added to the concept. This would allow the person creating the ontology to simply add the lexical representation and not have to be concerned with the way that the concept

was represented within OWL. The key requirement of the system was that the system should help the user to create an ontology easier than before.

Namespaces are another feature of OWL. Namespaces simply provide an area in which concepts may be placed and allow the user to separate the ontology. When different namespaces are created they are often placed into different files although this is not technically a requirement. Namespaces can then be referenced within the OWL document to show that a concept that is being referred to actually belongs to a different namespace than the current one and that namespace can be imported if required.

#### **7.4.1 Requirements**

Based upon the review of previous tools (seen in section 13), the literature review, the focus group and the restrictions already mentioned in this chapter, a number of requirements were constructed. The requirements included:

- Users being able to quickly bootstrap and create an initial ontology containing a number of concepts whilst being given as much help as possible.
- The system must allow concurrent access. As there may be many occasions when different users would be adding to the ontology at the same time and would need to see the updates that the other person had created. Further to this, the same user may also forget that they had already dealt with a concept.
- The system should aid users in determining if a concept already existed in the ontology and thus help reduce the number of duplicated concepts entered into the system.
- The system should be designed to minimise the complexities of creating an ontology wherever possible. In many cases, it is desirable to ensure that an employee that understands ontologies in great detail creates the ontology. In this scenario SoftwareCo determined that the ontology master did not need an in depth knowledge of ontologies and the technologies surrounding

them. Rather the ontology master should be of the mindset to create the ontology from a philosophical point of view.

- The system should be able to create something that enables a very fast creation of a basic ontology. Detail could be added later but the organisation needed a starting point. Many pieces of literature detail how ontologies can take a significant amount of time and investment before they can even be used (Farquhar, Fikes & Rice 1997), (Good et al. 2006), (Shadbolt, Berners-Lee & Hall 2006). The aim with this tool is to start simple and then add greater detail later.
- The system should allow different lexical representations of a concept to be entered into the system. This would allow the same concept to be reached from different synonyms within text.
- Restrictions should be allowed to be imposed on what may be entered into the system:
  - Control over properties available to those who work with the ontology should be added so that users are not free to enter any properties they choose and are confined to those already entered into the system.
  - Instances should not be allowed from the tool
- Modularity should feature in the system by allowing users to create a number of namespaces and place concepts into those namespaces.
- The system should allow users to search for a concept rather than have to find the concept in the existing hierarchy. This again could help to reduce the likelihood of duplications.

#### **7.4.2 Harvesting Information**

Research into harvesting information was undertaken to determine if information could be extracted from sites such as Google and Wikipedia to help users develop an ontology. The following section outlines the research behind the harvesting



process and how it was developed. Finally this section shows how information can be extracted in order to aid the development of an ontology.

The Concept Cloud method, that was introduced in Chapter 5, has shown that it can aid users in understanding the content of a web page and the concepts discussed within that page. The research question that was posed during the development of the OntoFarm system was 'would the Concept Cloud system be able to show the concepts surrounding a known page to give information on a certain subject?'

In order to test this theory and as a prototype experiment a number of concepts were searched for and a Concept Cloud created for the results. The Concept Cloud created and the concepts that surround the given subject of the page were examined in order to quickly determine whether Concept Clouds help suggest other topics that frequently relate to a given subject.

The first idea was to search Google for a specific term that was known to the author and see how the Concept Cloud related to that term. As Microsoft .Net Framework has a large presence on the web it was decided that one of the languages within that framework should be entered as a search term. As there could possibly be issues relating to 'C#' due to the sharp character, Visual Basic was entered as a search term. The phrase knowledge management was also used.



**Figure 32 - Google search for 'Visual Basic'**



**Figure 33 - Google search for 'Knowledge Management'**

Figure 32 and Figure 33 shows a Concept Cloud of Google searches for visual basic and knowledge management respectively. Interestingly along with the expected search terms, their abbreviations VB and KM both appear. A number of other terms related such as Microsoft Net MSDN appear. This highlights that it is possible to see a number of related terms. Previous literature has included systems that can automatically detect relationships from content on the Internet (Ponzetto, Strube 2007). In order to see whether relationships might be extracted by simply entering the search terms into the Google search engine, a search term followed by 'is a' was entered into Google.

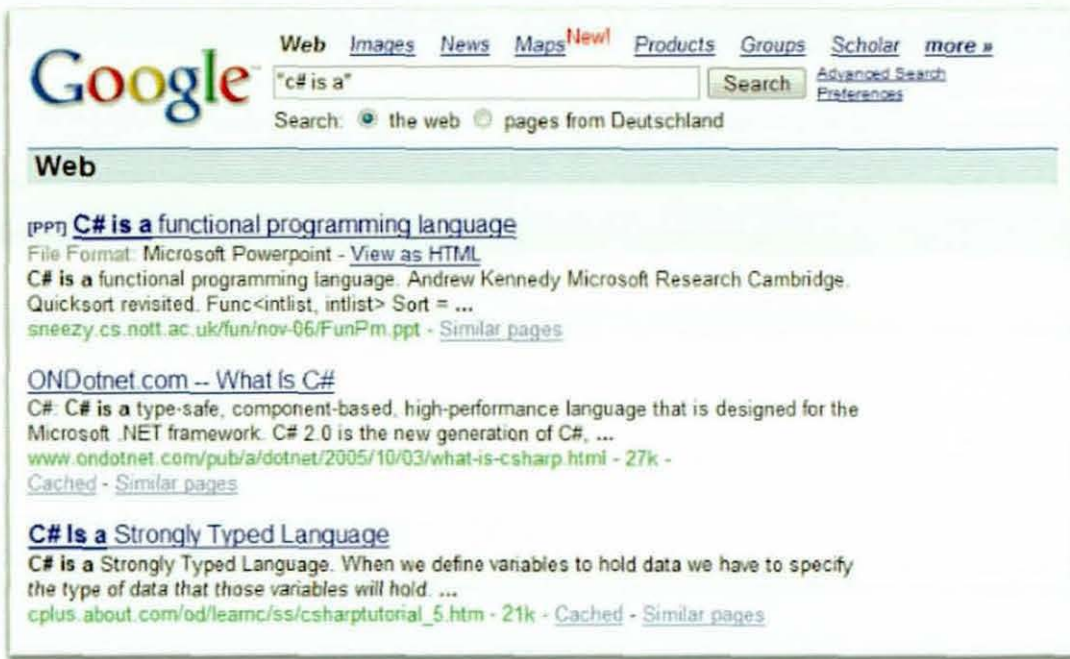


Figure 34 – A screen shot of the Google search results for “C# is a”



Figure 35 – A Concept Cloud showing the Google search for "VB is a"

Figure 34 shows the results of a Google search that searches specifically for the string 'C# is a' and then lists the results. The aim of this search was to use the search relevant query text that is returned with the links in order to find the answer to that specific question. Figure 34 shows that C sharp is a functional, type-safe, component-based, high-performance and strongly typed language. Most importantly it can be inferred that C# is a language. With a good enough ontology as a background it is also possible to infer things like 'a strongly typed language is a programming language'. Again this highlights the use of ontologies. Figure 35 shows the Concept Cloud for the search 'VB is a'. In this output, language is the top related term and programming is also mentioned. This again shows further

possibility of harvesting information from Google. The use of the “is\_a” term shows potential for harvesting a hierarchical relationship from Google.

Using Google as a resource provides a useful option for extracting information and concepts related to a certain term. However, if a Wikipedia article already exists for the given term then it would also be possible to extract concepts from the Wikipedia page. Wikipedia provides a vast corpus of knowledge that is always updated and maintained by a community. The relationships between concepts entered into Wikipedia are created by users and show links between two concepts that an author has deemed important enough to include. Although the credibility of Wikipedia is sometimes debated (Denning et al. 2005) (Chesney 2006) it can help in suggesting relationships that might exist to the ontology developer.

Figure 36 shows the Concept Cloud generated from the Wikipedia entry for visual basic.



**Figure 36 – A Concept Cloud generate from the Wikipedia “Visual Basic” entry**

Whilst generating a Concept Cloud from the entire Wikipedia article content provides a useful overview of its content, Wikipedia already contains a method for linking concepts that already exist in its system. When someone edits a Wikipedia page they have the ability to create hyperlinks to other concepts that already exist as a page in Wikipedia. Figure 37 shows a Concept Cloud generated from the links within the visual basic entry of Wikipedia. Figure 38 shows a similar cloud for the programming language, Ruby, that highlights the differences found even when searching for quite similar or related terms.



Figure 37 – Wikipedia Page Concept Cloud from the OntoFarm System for “Visual Basic”



Figure 38 - Wikipedia Page Concept Cloud from the OntoFarm System for “Ruby”  
highlighting the differences discovered between similar terms

The Concept Clouds generated from the Wikipedia links show a very high relation to the original concepts and highlight a large number of related concepts and technologies. The key issue with using Wikipedia is it may not contain the concepts that are necessary when building the ontology. It can however be used alongside Google to aid wherever possible.

The idea of harvesting concepts could also make use of a corporate intranet where available. The corporate intranet could provide an invaluable resource with information that is specifically relevant to the company. The company intranet may contain domain specific but also potentially more private information that is

not available in the public domain and that might make a welcome addition to an ontology to improve search performance.

#### **7.4.3 Initial Testing of the Harvesting Concept**

In order to test the concept of harvesting, a test was performed using a number of students studying ontology creation as part of their MSc in Information and Knowledge Management. The students were asked to design an ontology around the subject of camping.

Camping was chosen as the domain to model, as it would be something that everyone would have some basic knowledge of and would not cause a significantly difficult starting point for most. It was assumed that none of the users were really experts at camping. It would not have been an issue if some users were as the idea was to think about the concepts that would be placed into the ontology. In order to simplify the experiment, the users were also given the properties that they could use within the ontology. The properties were simply 'is a' and 'related to'. The phrase 'is a' was used rather than 'subClassOf' to aid the students in determining where that property would fit.

The 24 students were split into three groups. The first group created their ontologies without any external resource to help them. The second group was allowed to use the Internet, but could only use the Google search engine and only the results returned from the initial search query. They were not allowed to click through to the hyperlinks from the initial search results. The third group were only allowed to use Wikipedia to help them generate their ontologies. The ontologies were drawn on a piece of paper and students were given 30 minutes to create their ontologies. Thirty minutes was chosen because it would show how much could be created in a short space of time, which is directly related to the requirements outline earlier in the chapter. Any longer and perhaps too long would be devoted to the ontologies and a picture of the early stages of development would not be gathered. Secondary to this, students would be more likely to devote their full attention to the task if it was only 30 minutes long.

Although the students were not given a great amount of time to create the ontologies, the results from the study showed that the students who had being

creating the ontologies using the Internet had created much broader ontologies. They had found a significant number of related concepts and in many cases had not moved away from the central concept. The students who had been creating the ontologies without these references had created much smaller ontologies, but had begun to add sub concepts and their ontologies followed much more tree-like structures.

Although this experiment was extremely limited, it showed the potential of harvesting information from the Internet in order to aid ontology development. Following this initial investigation, it was determined that harvesting information from the Internet and presenting it to the users rather than attempting to provide a fully automated system might provide better results. As a result, the concept of semi-autonomous harvesting would be incorporated into the design of an ontology creation system.

#### **7.4.4 Development & Implementation**

Given the success in harvesting information from the Internet, it was decided that an ontology harvesting and creation tool should be created. Given that the tool harvested information and created ontologies it was decided that the system would be called OntoFarm.

There were a number of reasons why the ontology harvesting and a creation tool was developed as a web based tool. One reason was it allowed multiple users to work on the ontology at the same time, preventing difficult synchronisation issues that might otherwise occur. Using a web-based system also makes it simple to embed another web page into the tool so that the user may make use of that page whilst creating the ontology.

OntoFarm was developed using Ruby on Rails. The choice of Ruby on Rails was due to familiarity with the environment and the fact that rails allows very fast creation of prototypes and web content. Due to many existing systems within the target organisation already run on Java and because many other OWL and ontology related libraries are available in Java it was decided that JRuby would be used rather than Ruby. JRuby is an implementation of a Ruby compiler that compiles to Java byte code rather than the standard C-based implementation called MRI or

Matz's Ruby Interpreter. There are several benefits to the JRuby. JRuby runs on top of the java virtual machine providing numerous benefits such as efficient garbage collection. The key advantage, however, is the full integration and interoperability between Java and Ruby when using JRuby. It is possible to call any Java library from Ruby and Ruby from Java. Another factor in the choice of programming language was the integration of AJAX. AJAX allows updates and calls to the server from a web page without the need to refresh the entire page. This would prove extremely useful in the interface of the system. Rails makes use of a model, view, controller based architecture and thus the system was designed with this in mind.

There were two key components to the OntoFarm system and some other supplementary pages. The first part of the tool is the search view. In order to try and reduce the opportunity for duplications or entries into the system that might already exist in some form or cause confusion, all entries into the system begin with the search view. The search view allows a user to search for a concept before adding a new one into the system. As the user types into the search box, the list of concepts, shown in Figure 39, is refined to show any matching concepts.



## Create concept

format: namespace:Concept it should not matter how you type these

## Search Concepts

**test:Java**

---

Java,

**test:JRuby**

---

Jruby,

**test:MRIRuby**

---

The default and original ruby interpreter created by Matz

*Mri ruby, MRI, Matz ruby interpreter,*

Figure 39 - OntoFarm search view

A search can be performed with or without a namespace but can also be restricted. For example to force searching within the test namespace for a concept named ruby the user may type "test:ruby" and the concept shall be refined. In order to search any namespace the user may simply enter "ruby". Partial word matches shall also occur so "test:ru" would find the concept "ruby" within the test namespace. The system will also search the description of any concept in order to ensure that all related concepts are found. If no concept is found, the user may enter the namespace and concept name in order to create a new concept. Figure 40 shows the browse view searching for "ruby".

## Create concept

format: namespace:Concept it should not matter how you type these

## Search Concepts

### **test:Ruby\_Programming\_Language**

---

Ruby is a programming lanaguage

Ruby, Ruby programming language,

### **test:Ruby\_Interpreter**

---

Ruby interpreter,

### **test:JRuby**

---

Jruby,

### **test:YARV**

Figure 40 - OntoFarm search for Ruby

Once a concept has been found or a new concept is created, the user is directed to the concept view. The concept view is reachable via its own unique URL. This makes linking to the concept quite simply. New concepts can also be created simply by entering the URL containing the namespace and concept name if desired. For example to create the concept "ruby on rails" in the "test" namespace the following URL may be entered [http://localhost:3000/concept/test:Ruby\\_On\\_Rails](http://localhost:3000/concept/test:Ruby_On_Rails). Allowing a URL to be used is a simple method of creating concepts used by many online systems such as Wikipedia. This makes it easier for the user and increases the familiarity of the system, as it is similar to systems already used before.

Creating the search based system was an important decision, differing from many ontology development tools because it makes the user search before any action can be taken on the ontology. This search-first approach would then be assessed during the focus group in order to determine if users preferred this to the traditional method of browsing concepts in their hierarchies. The focus group

would also assess whether users felt that this would help to reduce the likelihood of duplication within the system.

The second key part of the system was the concept view. The concept view has a number of parts, which are highlighted in Figure 41.

### Concept Label and Description

The screenshot shows a web interface for a concept named 'test:Ruby\_On\_Rails'. At the top, there is a title bar with the name and a description: 'Ruby on Rails is an open source web application framework for the Ruby programming language. It is often referred to as "Rails" or "RoR".'. Below this is a 'Harvester' section with a search bar and filters. The main content area is titled 'Ruby on Rails' and includes a Wikipedia snippet, a 'Contents' list, and a 'Welcome aboard' message. At the bottom, there is a 'Harvest View' section showing a list of related terms and their relationships.

Harvest View

### Properties

The screenshot shows the 'Properties' section for the concept. It is divided into three main sections: 'Subject of Relations', 'Object of Relations', and 'Lexical Representations'. Each section contains a table of related concepts and their relationships.

| Subject            | Predicate | Object             | Delete |
|--------------------|-----------|--------------------|--------|
| test:Ruby_On_Rails | has       | test:Ruby_On_Rails | Delete |

| Subject            | Predicate | Object             | Delete |
|--------------------|-----------|--------------------|--------|
| test:Ruby_On_Rails | is a      | test:Ruby_On_Rails | Delete |

| Lexical Representation | Delete |
|------------------------|--------|
| test:Ruby_On_Rails     | Delete |

Lexical Representations

Figure 41 - The concept view

The first element of the concept view is the label and description section. This element allows the user to state the namespace and class names. The information is automatically formatted and inserts underscores and alters case according to preset rules. These rules were defined during the development of the system to ensure that all concepts entered into the system follow the same naming convention. The naming convention used by the system derives from the RDF naming convention, the only difference being that words that make up as class name are separated by underscores, whereas in RDF there is no separation of words. It was important to separate the words so that the system would be able to include spaces in any harvesting searches and so that the boundaries of different words could be interpreted by the system. During the export process the underscores are removed.

The formatting is done by an addition to the string class in ruby so that “string.conceptify” may be called at any point in time. A description may also be added, although this description is not necessarily exported it allows users to see the intended usage of this class name. If for example two similar classes exist such as the “Oracle\_DBMS” or the “Oracle\_Corp”, which symbolise the Oracle database management system or the Oracle Corporation exist, it prevents any misunderstanding and aids users when working collaboratively. Although it may be bad practice, both of these examples may be entered into the system in different namespaces simply as “Oracle” and the description could be used to differentiate between the two.

The next element of the system is the property section or relations as they are termed within the system. Relations show all of the properties that the concept is either the ‘subject of’ or the ‘object of’ and actually show the entire triple. The predicate can be chosen from a list of pre-defined predicates created by the key ontology master. In order to aid the ontology creator, as the user begins to type the name of a concept into the subject or object box, all existing concepts are suggested along with their descriptions. This allows the user to insert any existing concepts. If the user wishes to insert a new concept, they simply enter a concept that has never been entered previously and it is added to the system. Suggestions are filtered based upon both the namespace and concept parts, so having the namespace present means that the search engine will search within that namespace. Partial namespace titles and concept titles are also supported. Figure 42 shows an example of the auto-complete search for concepts within the OntoFarm system.

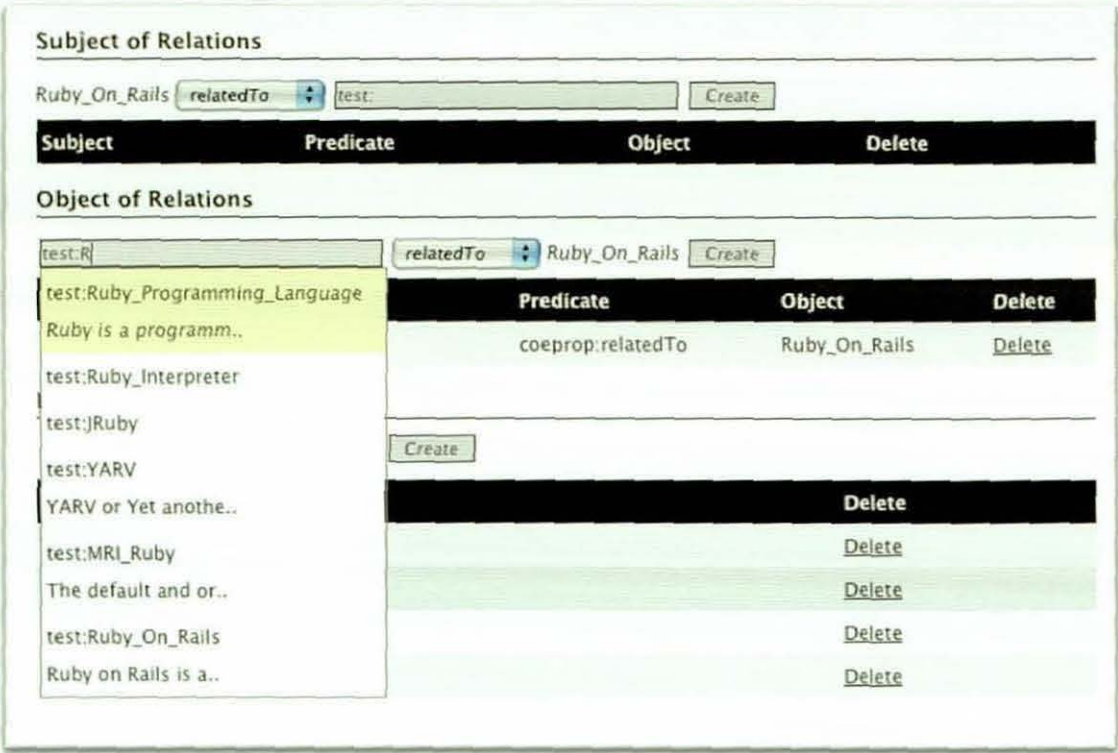


Figure 42 – Auto-complete for subjects or objects

The third section of the OntoFarm system is the lexical representation entry system. Lexical representations are automatically entered based upon a concept title and whenever the concept title is changed. Figure 43 shows the lexical representations entered for the example ruby on rails class.

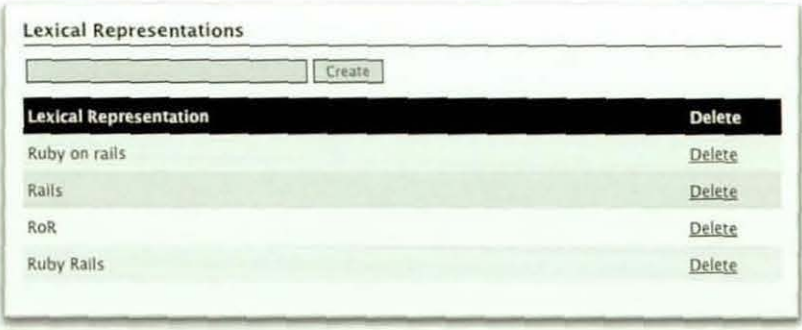


Figure 43 - Lexical representations of Rails

7.4.4.1 Developing the Harvesting System

The most challenging part of the system to create was the harvesting section. The harvest system presents a web page within the OntoFarm system. This web page is then processed and a Concept Cloud is displayed for the web page along with a list of concepts extracted from that page. The concepts extracted are either the links

that exist within a Wikipedia page or terms that most frequently occur on that page for Google and other sites. The harvest system starts with a search bar allowing the user to enter any search term. The default search term is the name of the concept with spaces rather than underscores. The user may then press one of the search buttons in order to search that site and harvest the resultant page. Figure 44 shows the harvester on the Wikipedia entry for ruby on rails.

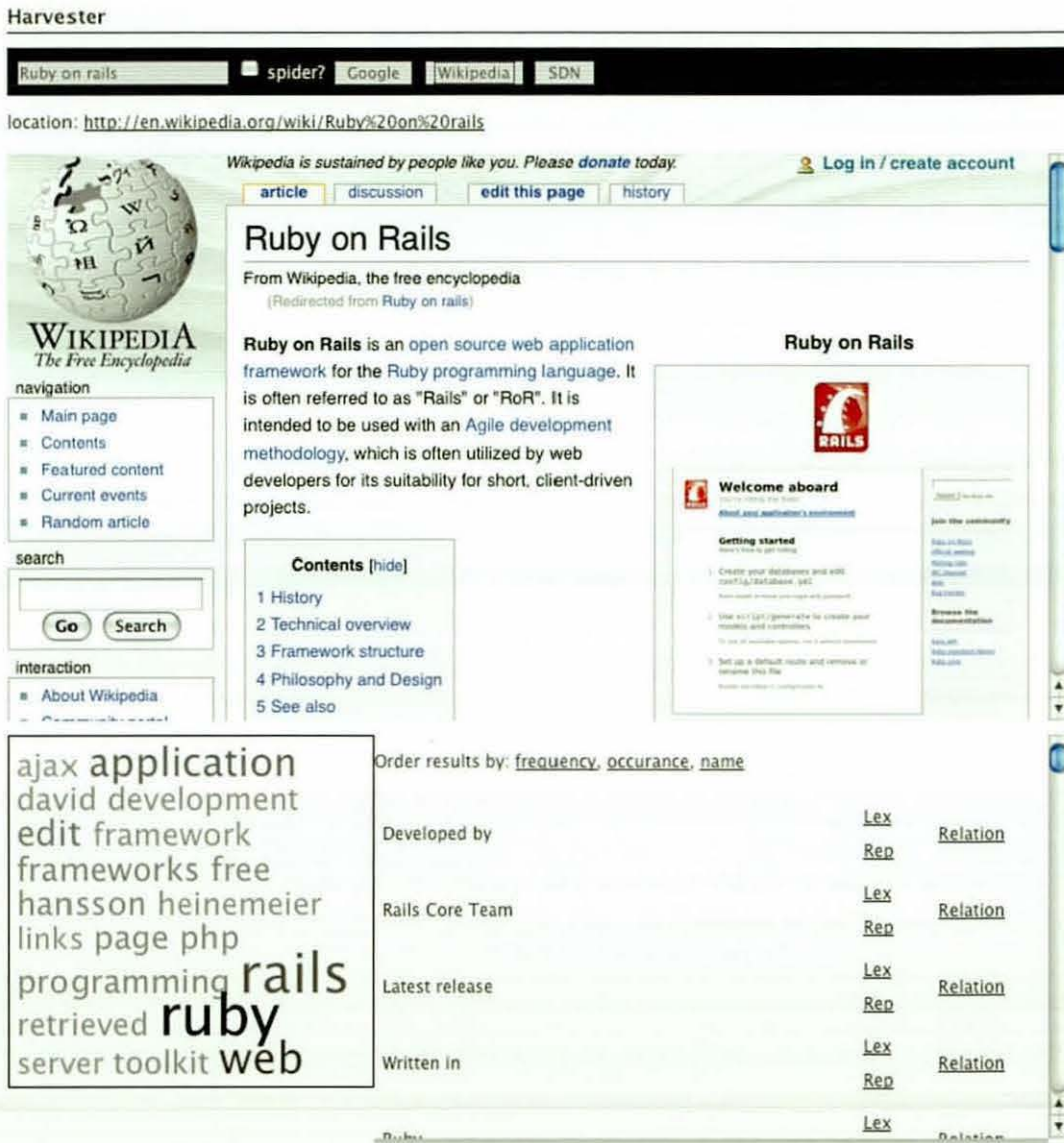


Figure 44 - The harvester searching Wikipedia

The harvester also allows the user to follow all of the links on the resulting page in order to collect the terms and create a Concept Cloud from all of the content found. This can be activated using the spider feature although it is extremely resource intensive, especially on secure sites. One of the key features of the embedded

browser is that it allows the user to navigate to any page and a Concept Cloud and list of terms are extracted for that page.

The cross-domain JavaScript restriction enforced by most modern browsers prevents a site from calling JavaScript on a page or IFrame, such as the one used to create the embedded browser, from a different domain name. This created a significant challenge as in order to determine the page that the user had navigated to in the browser, a JavaScript call to that IFrame would be required. In order to solve this issue a complete proxy server was created and implemented using ruby on rails and the hpricot html parser for ruby. The proxy navigates to and downloads the html for a page the hpricot parser. The hpricot parser then parses the page and alters all the links so that they referred to the same location, but navigate through the proxy. CSS or cascading style sheet files were also modified so that any images or imported styles would be available. Any place that a link could exist is parsed by the engine and modified so that the proxy is used. This allows the system to always know the current page and update the Concept Clouds and related terms accordingly. It also allows the system to display a link to the current page so that the user may open it in a larger, separate browser window if required.

The secondary benefit of this proxy server is that in future work, it will allow injection of content alongside the existing content that a user is browsing. This may allow the system to do things such as scroll to the part of the article where a concept was found if the user clicks the concept in the Concept Cloud.

The harvest view also presents a number of links for each extracted concept which when clicked will either add the concept as a lexical representation of a word or will fill in that concept as the subject or object of a relation. It will also allow the user to choose the predicate before saving the property. The system allows namespaces to be modified and created by an overall administrator. These namespaces can either be local and as such have a local URL or can be a remote URL. The remote URL feature can be used, for example, to import the RDF namespace. The namespaces can also be exported to OWL for importing into reasoning and related ontology systems.

There are a number of disadvantages of the system being created in its current format. The first, is the proxy server is quite resource intensive. Whenever a page is loaded into the harvester, it must be processed by the server. This processing can take considerable time when multiple accesses occur simultaneously. This is especially true when an SSL encrypted page is viewed. One of the contributing factors to this is that ruby on rails is currently not thread safe and although multiple instances of the server can be started and load balanced, it is still not an ideal solution. The current development version of ruby does however contain native threading rather than green threading and it is expected that rails will eventually become thread safe. Secondly, one of the disadvantages of the system comes from one of its advantages. The simplification of the system does impose a number of restrictions. One of these issues is that predicates for the relationships or properties that are created by the end user may not be modified by anyone other than an administrator. The advantages provided by these systems and methods were deemed by the researcher and SoftwareCo to outweigh the disadvantages.

## 7.5 Assessment

In order to assess and evaluate the performance of the OntoFarm system, two evaluations were performed. The first involved demonstrating the system and completing a questionnaire regarding the OntoFarm tool. This was conducted by undergraduate students that were familiar with ontologies. The second evaluation involved a focus group, which was held at SoftwareCo and was followed by a telephone interview with one of the ontology masters. The telephone interview was conducted because the employee was unable to be present during the focus group, but had used the system in production for a number of months to create an ontology. The stages used to assess the performance of the Harvesting method and OntoFarm system are shown within Figure 45.





**Figure 45 - Evaluating the Harvesting method and OntoFarm System**

### 7.5.1 First Evaluation – Demonstration and Questionnaire

An initial questionnaire was created to gather feedback on the OntoFarm tool before employees of the target organisation SoftwareCo were surveyed.

The students were undergraduate students on the Information Management and Retrieval module. As part of their course the students had been taught in the field of ontologies. The students were given a presentation about the OntoFarm tool. The presentation covered all of the features of the OntoFarm system and highlighted some of the benefits of the system. Following the presentation, the students designed an ontology together as a group. The ontology was based around the concept of music and involved the students working together. The students decided on the concepts within the ontology and reasoned with each other where concepts should be placed in an interactive session. The OntoFarm system was projected onto the screen and was used to capture the ideas of the students. It was also used at a number of stages to harvest information and show possible concepts that could be added to the system.

Following the interactive session and demonstration of the OntoFarm system, the students were asked to complete a questionnaire relating to the OntoFarm system that they had just seen in use. The questionnaire (See Appendix Four – OntoFarm Questionnaires) asked a number of questions relating to the OntoFarm tool. The questions were centred on the tool's ability to create an ontology, especially the initial bootstrapping of an ontology and its efficiency.

The questions in the questionnaire were based on three key areas. The first was the overall ability of the OntoFarm tool and contained six questions. This section

asked general questions relating to the ease of use and understanding of the tool, the layout of information and the lexical representations feature of the system. The second section contained questions relating to the benefits of the OntoFarm system over alternatives. The section asked questions relating to the speed at which concepts could be added to the system, the highlighting of concepts that already existed in the system and the ability for more than one person to use the system at a time. The third and final section contained six questions around the harvesting part of the tool. The questions were related to areas such as how adequate the returned concepts by the harvesting system were and to evaluate the overview provided by the Concept Cloud. Every question in this questionnaire related to a feature that had been implemented as part of the requirements identified for this system or related to the ease of use of benefits of the system.

52 students were asked to fill in the questionnaire. Of the 52, 27 responses were received. Three of the questionnaires were discarded because they contained only neutral answers throughout the entire questionnaire. These three responses would appear to be students who had no feelings towards the OntoFarm system and since the results were all neutral they would not affect the results. With the three removed there were 24 sets of responses remaining. Of the 24 there are a number that are still questionable but are still included as it would be impossible to tell whether the answers were chosen randomly or because the students felt that these were the correct answers.

When asked to rate the ease of use of the OntoFarm tool 82% of participants stated that the system was easy to use. Creating a concept and finding an existing concept were also seen as an easy thing to do. The layout of information was also seen as easy to understand with nearly 95% either agreeing or strongly agreeing. The students also agreed that the different lexical representations could be adequately represented.

Of those who responded 92% of participants agreed that the system allowed users to add concepts in a quicker manner than previous systems they had seen. Eighty-five percent stated that the system help highlight concepts that may already exist

in the system. The students were then asked if the highlighting of concepts already in the system prevented duplication. All of those who answered agreed that it did.

Almost 90% of participants felt that a clear separation of namespaces was produced by the system and 82% either agreed or strongly agreed that the system provides improved search functionality over traditional systems.

All of the participants who answered this question stated that they agreed the system allows more than one person to work on the ontology at the different times. Respondents stated (75%) that the system allowed them to work on an ontology at the same time.

When questioned about the harvesting tool 94% agreed that a good list of concepts were provided by the system and 27% stated that they strongly agreed. The participants also appeared to be pleased with the Concept Cloud system with 93% feeling that the Concept Cloud gives a good overview of the page that the harvester is currently displaying.

The majority of participants felt the list of concepts generated for a page was not too long, with only 20% indicating it was. Ninety percent also agreed or strongly agreed that the tool made it easier to add concepts from the list provided by the harvester.

Many of the participants stated the ease of use, the ease of expanding a concept through searching and the prevention of duplication as being strengths during free text questions. One of the key areas that participants stated could be improved was the layout of the system and its user interface. The layout of the system though was not one of the priorities listed during the requirements of this system. It was felt that first the concept must be proved and then if the system was successful then the layout and user interface could be improved at a later date. These comments need to be taken into consideration in any future development.

Overall the system would appear to have met many of its requirements. The complexities of ontology development have been reduced and a simplified view is provided for the user allowing users to quickly and easily create an initial ontology. Participants stated that they felt they could create an ontology faster

than before. The participants seemed satisfied that an ontology could be created quickly and effectively and were impressed by the functionality allowing them to enter multiple lexical representations. Participants also stated that the system produced a clear separation of namespaces.

Further investigation with a focus group within target organisation, SoftwareCo formed the second evaluation.

### 7.5.2 Focus Group

In order to further analyse the tool and assess its strengths and weaknesses a focus group was conducted within the SoftwareCo organisation. The focus group comprised of the same employees that completed the initial questionnaire relating to ontologies in general in section 7.3.

The focus group contained ten members all of which had some form of experience with ontology development tools and were familiar with OWL. Before the focus group participants were asked how they rated their experience with ontologies. Four of the employees stated that they only had a limited experience with ontologies. The four participants with limited experience with ontologies were also given a prior and unbiased introduction to some of the alternative tools available, such as protégé and TopBraid composer. The participants that had limited experience were given time to create a number of ontologies and experience those tools before the focus group to enable fair comparisons to be made.

The focus group received a demonstration of the Concept Cloud system and was invited to use the system and explore its full potential. The focus group began with a discussion surrounding ontologies. Following this, the OntoFarm tool was shown and a discussion took place. The focus group focused on the tool's strengths and potential within the organisation then the weaknesses and areas that the tool could be improved. Following this discussion, the focus group was asked to answer a number of questions within a questionnaire (See Appendix Four – OntoFarm Questionnaires). Finally, after the questionnaire had been completed, a final discussion took place to incorporate questions or thoughts that may have arisen from the thought provoking questionnaire.

The questionnaire centred on several topic areas. The requirements of the tool identified in Section 7.4.1 were used as a basis for the questions to determine how well the tool met these requirements. The questionnaire therefore asked questions in the following areas:

- The ability of the tool to help the 'bootstrapping' of an initial ontology
- Allowing concurrent access
- *Highlighting concepts that already exist in the system*
- Minimising the complexities associated with developing an ontology
- Increasing the speed at which initial ontologies can be created but still allow detail to be added later
- Allowing the entry of different lexical representations of a concept
- *The tool's system of restricting what can be entered into the system including the properties that can be entered.*
- The tool's support for modularity through namespaces
- The search based approach of the tool in contrast to traditional browsing of an ontology.

The questions comprised of a number of multiple-choice questions and some free text questions asking participants to give more detailed descriptions. The multiple-choice questions gave participants five options on an ordinal scale and included a neutral answer. Employees were not forced to answer all questions and could leave questions blank if they wished to.

The focus group was not recorded or transcribed. This was due to privacy concerns raised by SoftwareCo and the employees that took part in the focus group. As also used in the previous focus group, participants were asked to write down on a grid the strengths, weaknesses, opportunities and threats of the system. These four options were explained as follows

- Strengths - The advantages of the tool (i.e. what does the tool do well?)
- Weaknesses - The disadvantages of the tool (i.e. what does the tool do badly?)
- Opportunities - The advantages of the tool to the organisation (i.e. how will the listed strengths benefit you and/or the organisation)

- Threats - The disadvantages of the tool to the organisation (i.e. how will the listed strengths or weaknesses threaten you and/or the organisation)

These factors were recorded in place of a transcription by the participants themselves and allowed the participants to phrase their views how they wished and record as little or as much as they wanted.

After the focus group had taken place, a telephone interview was also conducted with an employee who had used the OntoFarm system extensively since its deployment within the organisation. The participant was unable to attend the focus group but had many interesting points. These points shall be included at the end of the focus group results.

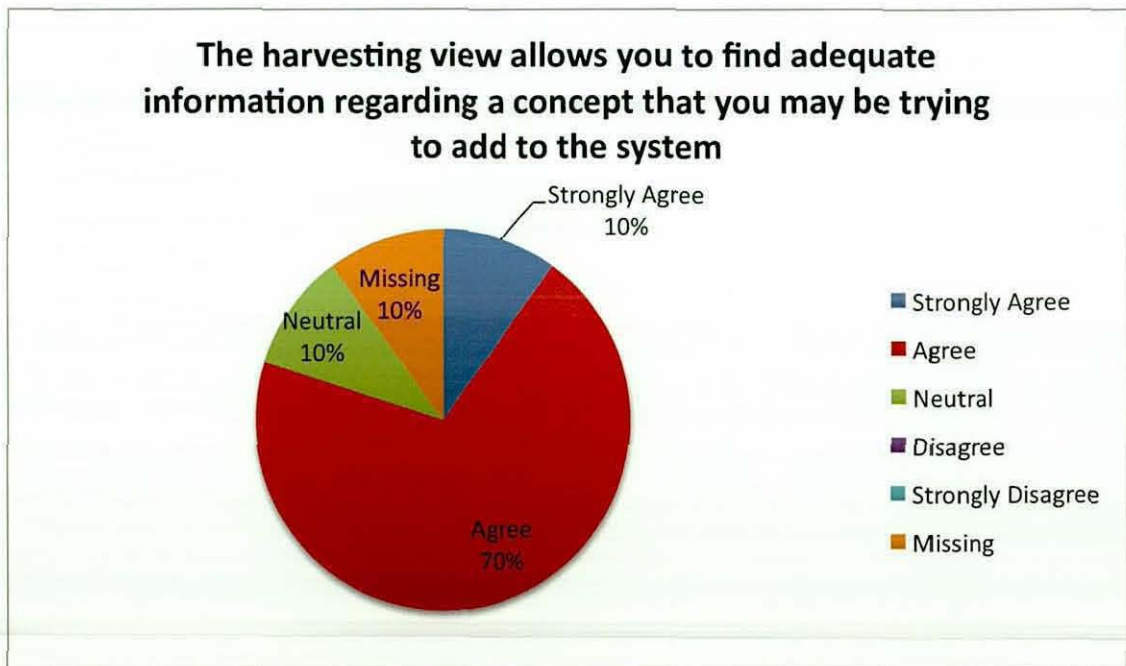
#### **7.5.2.1 Results**

Participants of the focus group were first shown the OntoFarm search view, which was widely accepted. Participants of the focus group felt that it would allow them to find concepts far more easily, especially if other people had added the concepts. All of the participants appeared to feel that the search system made things easier *and that it was better than having to browse a hierarchical structure. The* questionnaire results confirmed this. Employees were asked if the search based approach makes it easy to discover concepts in the system to which participants unanimously agreed, 100% of the participants felt that it would.

After looking at the search view, participants were shown the concept page that included the harvesting system as shown in Figure 41. Participants first examined the harvesting system and its approach. During the focus group, the key area that participants appeared impressed with was the Concept Cloud view of the page. Many of the participants felt that this alone would help to prompt them when creating an ontology. Participants felt that they should already have an understanding of an area when they were creating an ontology. They felt that the ontology creation should not be left to those users that did not understand the subject area they were describing. The Concept Cloud would therefore help to remind these users of the concepts that should be added to the ontology. The focus group participants also appeared to appreciate the web browser being built into the system. However, the fact that they could open the link in a new window was of

more interest. Participants stated that they preferred to see the page in its entirety instead of within the small window of the OntoFarm system. The browser window was only really of use to find the correct page for the harvesting system to harvest.

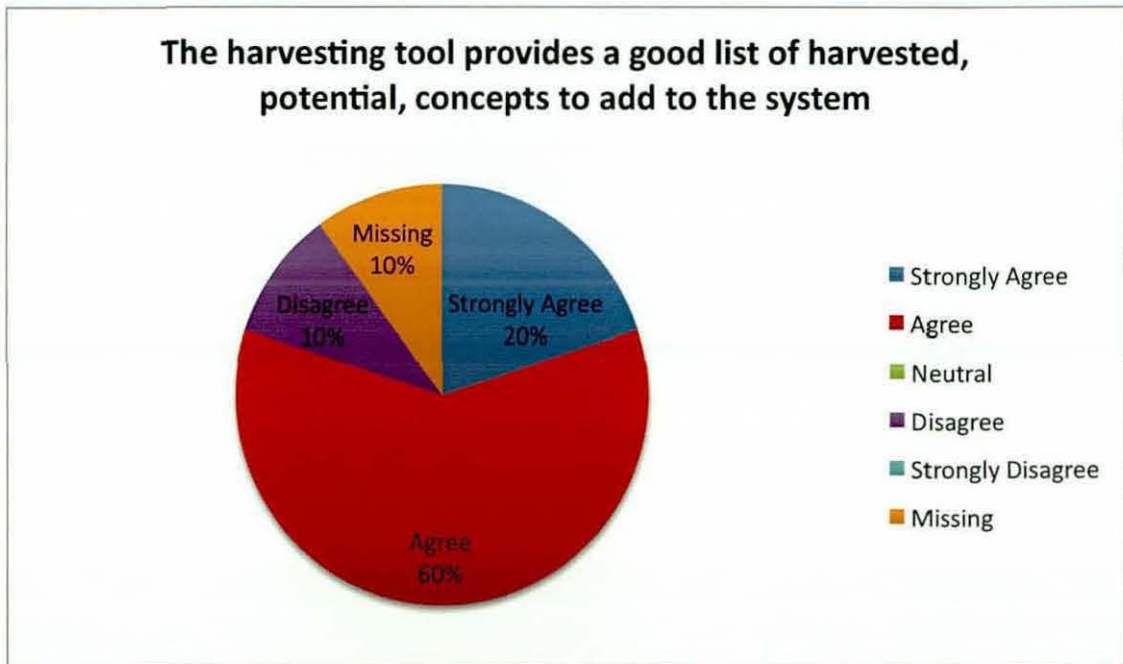
The questions relating to the harvesting system also highlighted its potential with 89% (of those participants that answered) of the participants thinking that the harvesting system worked well, providing a good list of potential concepts to add to the ontology. The same participants (89%) also agreed that the harvesting view allowed them to find adequate information regarding the concept they were adding to the system. The questionnaire also confirmed the participants' appreciation for the Concept Cloud system with 89% of the participants that answered agreeing that the Concept Cloud gave a good visual overview of the page that the system was currently displaying. This demonstrates that participants were happy with the visualisation and its ability to demonstrate the content of a page along with the harvesting system in general.



**Figure 46 – Adequate information from the harvest view**

Once harvesting was complete and it came to actually adding concepts to the system, all of the participants appeared content that the system made it easy to add concepts from harvested list to the ontology. The questionnaire results showed all of the participants stated that adding concepts to the system was easy.

The same participants in the focus group were impressed when they looked through the list of retrieved concepts surrounding a topic. Many of the focus group participants chose a subject that they were interested in and then looked for words that they would have suggested in the list of results returned. In almost all cases they were pleased to find the words that they wished to find. Those that did not find the terms they expected appreciated why the system would not find that result as the term did not occur commonly on the Internet but was of specialist interest to them. The questionnaire was used to verify these findings. As previously mentioned the questionnaire results found that 89% of participants felt that the system provided a good list of potential concepts.

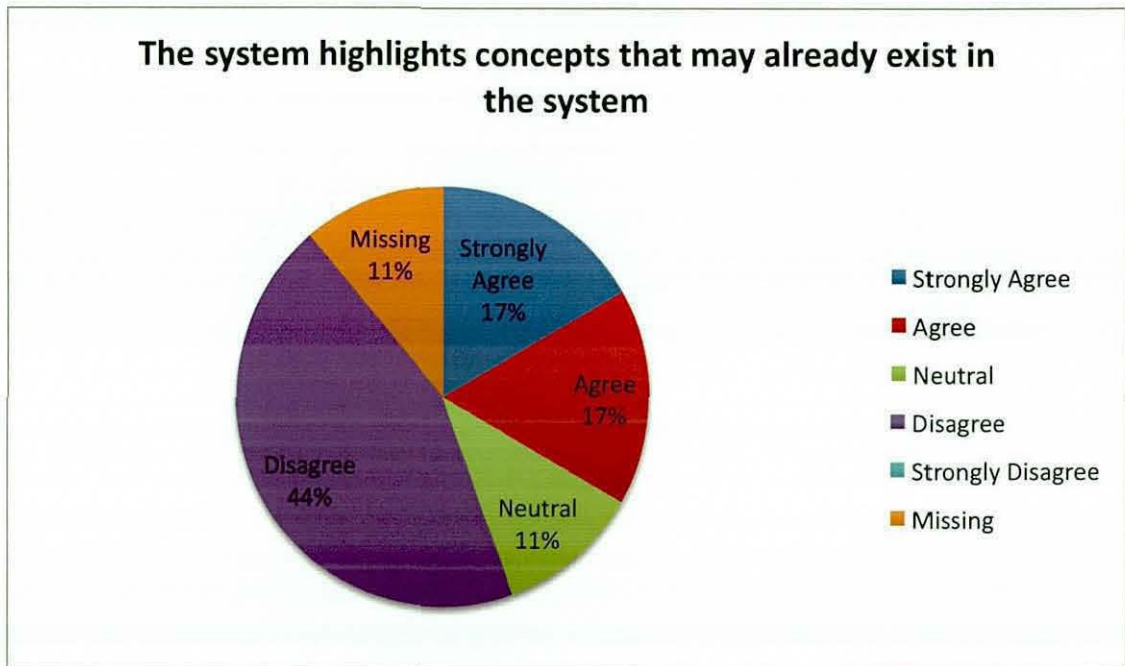


**Figure 47 – Harvested concepts**

One of the original intentions of the system was to use Ajax in order to highlight concepts that might already exist in the system when adding new concepts to relationships. The focus group did not touch too much upon the Ajax functionality or the ability to prevent ontology masters from adding concepts that might already exist to the system. The questionnaire however asked a number of questions around this subject. A number of participants felt that the system highlighted concepts that may already exist in the ontology, 25% of the participants that answered this question strongly agreed with this and 50% agreed. With one user (13%) disagreeing and one user (13%) had a neutral opinion. Two participants did



not answer this question. This shows that 75% of the participants felt that the system would help users to find concepts that already existed in the ontology. Highlighting these concepts would allow users to see concepts that already existed in the system and allow users to link to these concepts from the one that they were creating.



**Figure 48 - The system highlights concepts that already exist**

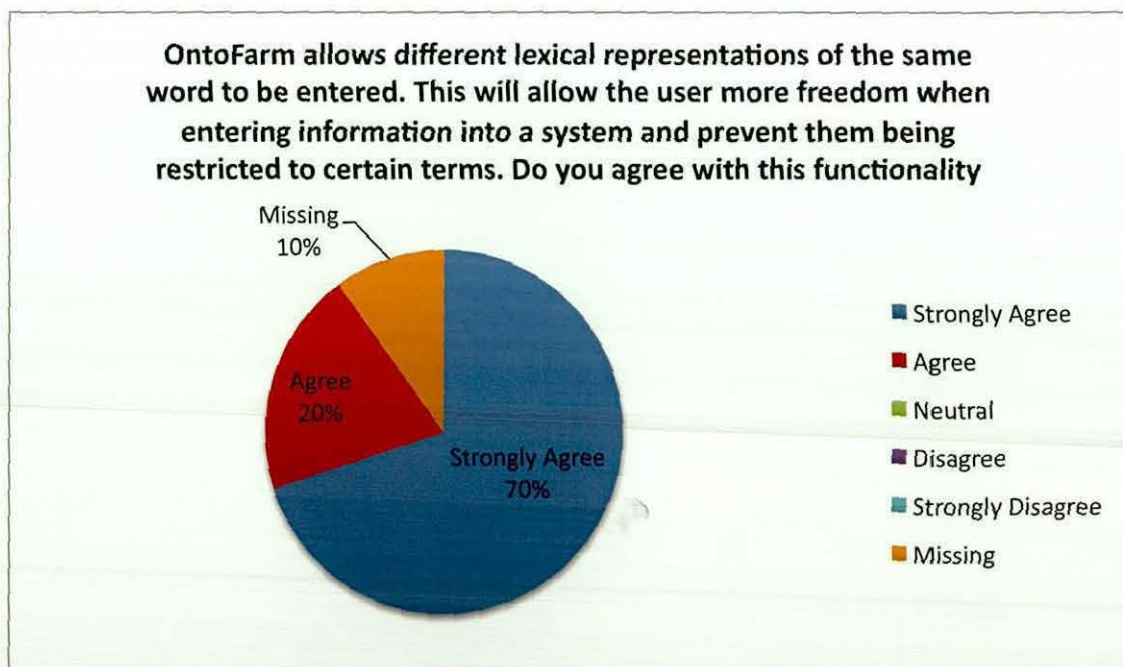
The main aim of this feature was to help to prevent the participants from creating concepts that already existed in the system. If the participants created duplicate concepts then the ontology would be disjointed. The reasoning performed on such an ontology may therefore be incomplete or incorrect.

Questionnaire participants were asked whether simply highlighting existing concepts would help the system to prevent participants from entering duplicate concepts. Although 38% of participants who answered agreed with this statement, 38% of participants that answered had a neutral opinion and 25% actually disagreed with this statement. Two users did not answer this question. When questioned about this in the focus group, some of the participants stated that they felt the system only highlighted concepts that had a similar name, set of lexical representations or whose descriptions contained similar words. Participants stated that the system makes no attempt to highlight concepts that are perhaps

synonyms of each other or that may describe the same concept but in a different way. It is arguable that this is the job of the ontology masters, however when more than one ontology master is working with the ontology difficulties may occur if extremely good communication is not present. Further research into this area may be necessary.

Many participants felt that the system allowed adequate division of namespaces, with 78% either agreeing or strongly agreeing. Two participants did not answer this question. The two (22%) participants that gave a neutral answer further explained that the system allowed separation of namespaces however, modularity is a design issue rather than something that the tool can provide.

Participants were extremely happy with the system when it came to lexical representations of concepts. Within the focus group they stated that they felt that one of the key features of the tool was the lexical representation system and that this would prove extremely useful by itself even if the other features did not exist. Seven participants (78% of those who answered) strongly agreed with the lexical representation system, two (22%) agreed and one user did not answer this question.



#### Figure 49 - Lexical representations of concepts

The questionnaire asked if it was important that the system enabled more than one person to work with the ontology at the same time. Opinions were quite split with the majority (67% of those who answered) of participants strongly agreed or agreed but there were two neutral opinions (22%) and a disagreement from one user (11%). One user did not answer this question. Again most participants felt that restricting the predicates that may be entered into the system was a good idea and made it easier to user, but one user disagreed. Most participants also agreed that the system allowed participants who were not necessarily experts when working with ontologies to enter items into the ontology. Two participants (22% of those who answered) strongly agreed, 6 participants (67%) agreed, one user (11%) disagreed and one user did not answer this question.

The simple idea of adding a description to concepts was also favoured in the most part with two participants (22% of those who answered) strongly agreeing that it helps to prevent duplication of concepts and misunderstanding. Six (67%) participants agreed with this system and only one (11%) disagreed. One person did not answer this question. The person who disagreed in the questionnaire did not voice any concerns with the descriptions being presented at the focus group.

Overall most participants felt that the system was easy to understand. When the questionnaire asked if the system took a long time to understand, one participant did not answer, three (33% of those who answered) had a neutral opinion, four disagreed (44%) and two strongly disagreed (22%). This was apparent within the focus group also with most participants quickly understanding the system and using it without problems.

The questionnaire showed that participants felt that the OntoFarm system would take less time to create an initial ontology than traditional systems such as protégé or than simply using a text editor. All but one (86% of those who answered) participant of the seven that answered this question also felt that the OntoFarm system would take less time to maintain an ontology once it had been created. The one participant (14%) felt that when maintaining an ontology, protégé would be

quicker. The same user felt that protégé would provide a better-structured ontology.

Following the focus group, an interview with an employee who heavily used the OntoFarm tool for ontology generation within the target organisation took place via the telephone. This employee was not present during the focus group. He stated that when entering concepts, he had already decided the concepts that should be added. Where the harvesting tool was of benefit was in determining the relationships that should exist to other concepts and within the lexical representation field. Once concepts are added he would search around that concept finding different lexical representations and any relationships that might exist. In this area, the auto-complete system also proved extremely useful. The idea of lexical representations also helps considerably as often, for example, products have even changed name over time, although they still refer to the same concepts. The ability to jump to a concept by simply clicking its name was also appreciated. One small worry from this power user was that jumping from concept to concept may lead to a slightly disjointed ontology if considerable thought was not placed into its creation. Overall he said that the system was "Very easy to use, but very powerful" which was an extremely positive comment.

## 7.6 Conclusions

Ontologies present a method that can add context and understanding to information in a way that makes it easier to search and interpret the information. In turn this context and reasoning can provide benefit when attempting to discover relevant information and filter out the irrelevant information, in turn helping to reduce the information overload problem.

This chapter has given an introduction to the need for an easy to use ontology creation system. The research has shown that the majority of ontology creation systems are quite cumbersome and take some time to develop an ontology. As part of this chapter, requirements were developed in order to create an ontology system that would be an improvement on existing systems, a summary of the main requirements are:

- Users should be able to quickly bootstrap and create an initial ontology containing a number of concepts whilst being given as much help as possible.
- The system must allow concurrent access.
- The system should aid users in determining if a concept already existed in the ontology and thus help reduce the number of duplicated concepts entered into the system.
- The system should be designed to minimise the complexities of creating an ontology wherever possible.
- The system should be able to create something that enables a very fast creation of a basic ontology.
- The system should allow different lexical representations of a concept to be entered into the system.
- Restrictions should be allowed to be imposed on what may be entered into the system
  - Control over properties available to those who work with the ontology should be added so that users are not free to enter any properties they choose and are confined to those already entered into the system.
  - Instances should not be allowed from the tool
- Modularity should feature in the system by allowing users to create a number of namespaces and place concepts into those namespaces.
- The system should allow users to search for a concept rather than have to find the concept in the existing hierarchy.

Before the system was developed a number of approaches were trialled. One of the concepts was harvesting. This involves using existing material on the web to create a Concept Cloud that can be presented to the user. To test this approach, a

proof of concept was developed and the potential of semi-automated harvesting system was assessed.

With the potential for harvesting highlighted, an *ontology development tool* was created based on the requirements. The construction of the tool and its functionality was detailed within the chapter and included the reasoning behind developing a proxy server in order to bypass the *cross-domain JavaScript* restriction in place in modern web browsers.

The ontology tool and its potential was assessed. Firstly, a questionnaire was given to a group of participants who were studying ontologies as part of their course. Secondly, a *focus group* was held at *SoftwareCo* and they also completed the questionnaire. The focus group provided more detail and discussed the benefits and disadvantages of the system. The results from both studies showed positive results. Although there are areas where the tool can be improved there were significant advantages of using tool. The key advantages highlighted included the search-based approach that it took, the ability to easily and quickly discover and add *lexical representations for a concept* and the ability to quickly discover and determine relationships that might exist. The *Concept Clouds* included in the harvesting system were also praised for being able to give an impressive overview of the page and thus other concepts related to a particular concept or topic area.

In summary, the semi-autonomous ontology development method has been a success. The *OntoFarm* system itself has met the requirements and has provided a solution that is both quick and easy to use when it comes to building ontologies. The system has the potential to save organisations significant cost during the development and maintenance of ontologies. Not only will it save employee time in ontology construction, but also has the potential to produce a richer ontology that will aid information retrieval throughout the organisation. The research detailed in this chapter shows that there is a cost effective way of creating an ontology that can ultimately help employees by enhancing their search for information. This is shown by Figure 50, which also shows how an ontology can cut across the majority of information sources to aid employees in their quest for useful information.

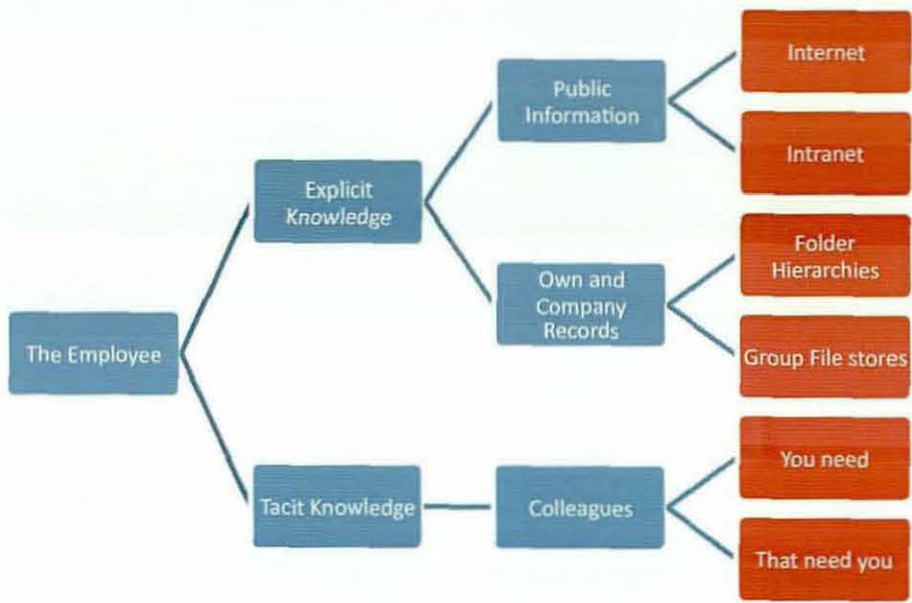


Figure 50 - Information Sources that can make use of Ontologies

## 8 Conclusions and Recommendations Framework

### 8.1 Preface

This chapter concludes the research that has been detailed in this thesis and provides recommendations that can be used to optimise information retrieval within organisations. The chapter also reflects on how the aim and objectives have been achieved, the limitations of the research and provides areas for future research.

### 8.2 Introduction

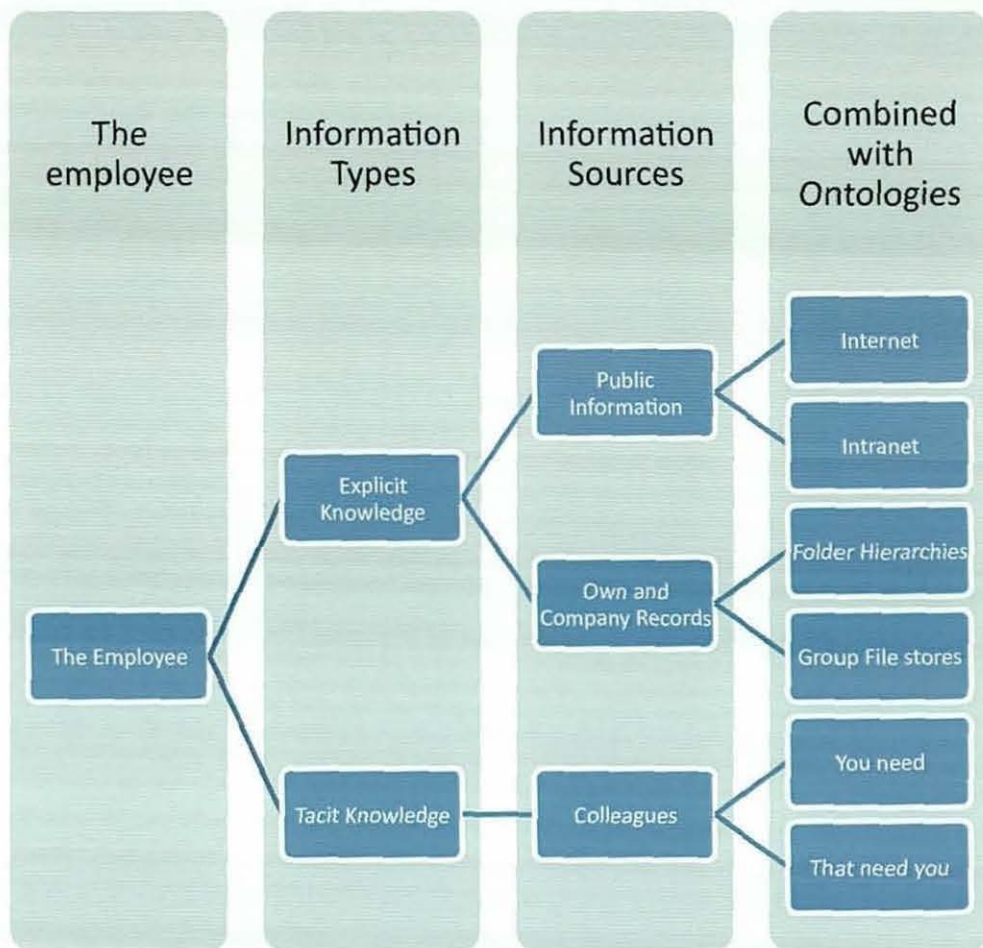
The literature review and Chapters 4 to 7 have identified ways to help overcome the barriers to the discovery of relevant information with a view to reducing information overload. This section combines all of the research and provides a framework that can be used to help organisations reduce the barriers to obtaining relevant information and help reduce information overload.

The research contained within this thesis has focused on information that can be obtained from three key sources.

- Public information from a company intranet or the larger Internet;
- Information from an employee's own or company record stores;
- and information that can be passed on to and obtained from colleagues directly.

The sources of information, identified within the literature review, which formed the basis for this research, are presented in Figure 51.





**Figure 51 - The sources of Information addressed by this research**

In addition to the information sources shown in Figure 51 the research also identified the potential benefits of ontologies. Ontologies can be used throughout an organisation to help structure and store information and aid the retrieval of that information. This in turn can significantly improve a user's ability to obtain relevant information.

This research has shown that, independently, through the reduction of the barriers to each of the information sources presented, it might be possible to improve an employee's access to relevant information. However, if these methods are used appropriately and with the aid of ontologies it may be possible for an organisation to dramatically reduce the problem of information overload, and potentially save the organisation significant resources.

### **8.3 Introduction to the Recommendations Framework**

The literature review identified that knowledge sharing barriers could affect organisations in different ways, in addition the literature review further found that different organisations were affected by information overload to varying degrees. The following sections discuss a Recommendations Framework, shown by Figure 52, that can be used to establish where an organisation should focus its efforts when it comes to improving information retrieval. If the organisation is able to determine where it must focus its efforts then it can implement solutions to overcome the barriers described in this thesis.

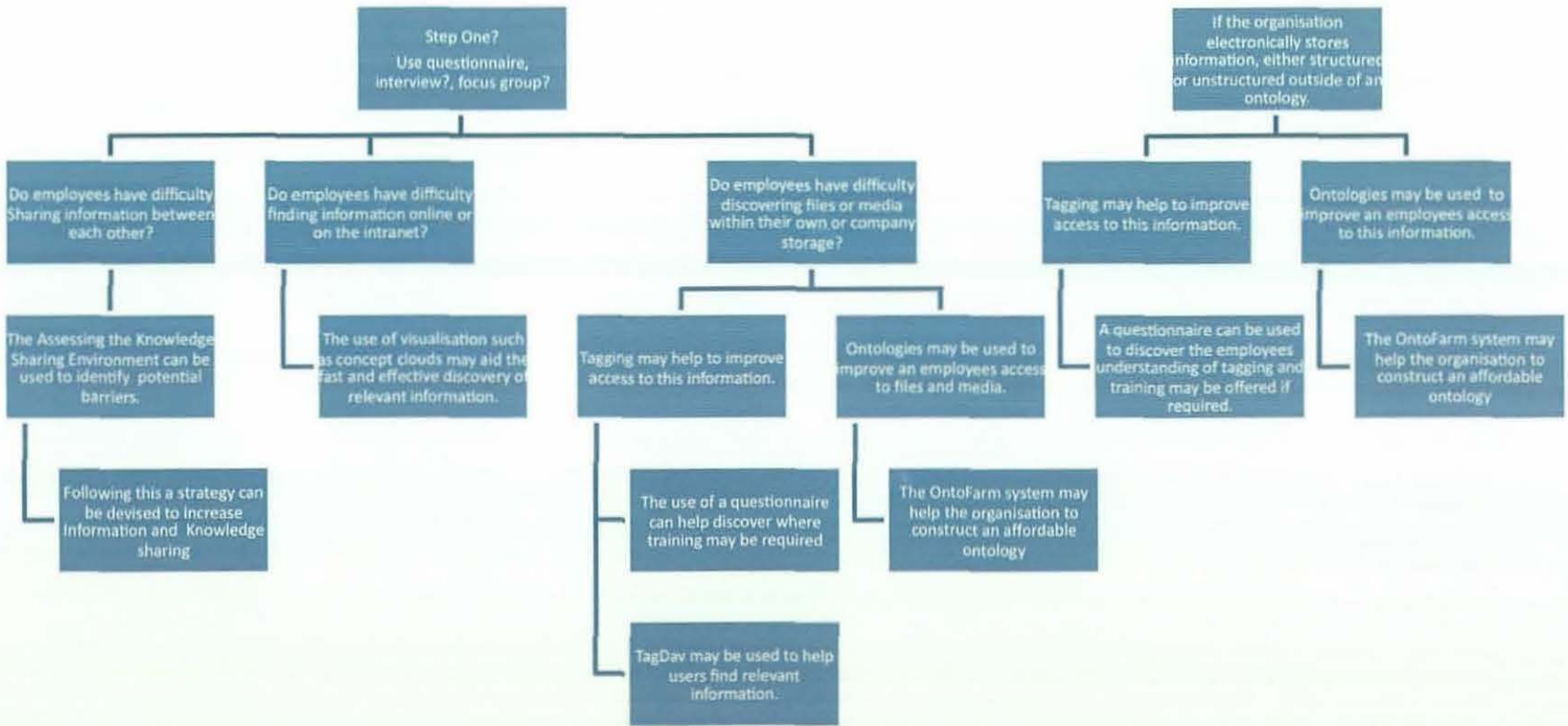


Figure 52 - Recommendations Framework

The solutions provided in Figure 52 aim to reduce the barriers that exist to each of the approaches and improve the ability for an employee to discover and interact with relevant information and prevent information overload.

## **8.4 Breakdown of the Recommendations Framework**

The first step is to try and gain an understanding of the problems that might exist within a particular organisation. Once the organisation has a grasp of the problems that exist within the organisation the framework can be used to help reduce those problems.

This section shall now take a more detailed look at each of the potential problem sources that employees may face when trying to discover relevant information and how the research from this thesis may be used to help overcome these issues.

### **8.4.1 Difficulty sharing Information from Colleagues**

If an organisation discovers that its employees struggle to both discover and share relevant information with their colleagues or the information is relevant information is held back due to knowledge sharing concerns then the following approach can be taken.

The literature review identified a number of potential barriers to the success of knowledge sharing between colleagues within organisations (Riege 2005). The literature further identified that each organisation and its employees would differ when it came to the facilitation of knowledge sharing and the knowledge sharing that occurred (Argote, Ingram 2000). Authors such as Reige (2005) argued that it is difficult to determine the extent that knowledge sharing is taking place within an organisation. Despite these difficulties Reige (2005) did identify 36 key knowledge sharing barriers that organisations may face. Further to this Riege also offered a number of possible solutions.

Chapter 4 of this research took the study by Reige further by allowing an organisation to determine the extent to which these barriers were present within that organisation. The chapter presented an approach to determining the severity of each of the barriers through the use of a questionnaire. Organisations may use

this questionnaire to determine the key areas to improve knowledge sharing between colleagues.

Two case study organisations have used the questionnaire to determine how the potential barriers affected their organisation. Following the use of the questionnaire if organisations make use of the 'traffic light system' presented in section 4.6.2 they can quickly gain an overview of the areas that must be addressed to improve knowledge sharing within that organisation. With knowledge sharing improved then an environment open to the sharing of relevant information is likely to follow.

Chapter 4 showed how the two case-study organisations made use of the traffic light system, after responding to the questionnaire, to identify their potential barriers and highlighted the differences between these two organisations and the barriers present. Once these barriers are identified the organisation can work to reduce their impact and improve the knowledge sharing environment and ensure that access to relevant information is optimised and the amount of irrelevant information an employee is exposed to is reduced.

#### **8.4.2 Difficulty discovering Information from intranets and the Internet**

Chapter 5 presented a visualisation called Concept Clouds. The Concept Cloud visualisation may be used by organisations that have difficulty discovering relevant information on the Internet and intranets. If employees are having difficulty discovering relevant information through this medium then they can make use of the Concept Clouds visualisation to allow them to discover relevant information more quickly. The Concept Cloud visualisation is a complementary system, to be used in conjunction with traditional search engine results and was found to benefit both the time it took to find relevant information and reduce the inaccuracy of results.

In an experiment, documented in Chapter 5, users showed between a two to twelve percent decrease in the time it took them to discover relevant information using the Concept Cloud system. In addition the system saw the number of incorrect answers given by participants was 28% less for those using the Concept Cloud system. This could constitute a considerable saving for organisations aiming

to improve access to relevant information. Although the visualisation was implemented before query relevant text was as prevalent as it is today, it may still provide help to users attempting to find relevant information.

#### **8.4.3 Difficulty locating relevant documents from personal or group stores**

The diagram in Figure 52 shows two potential avenues that can benefit an organisation where employees struggle to find relevant information personal or group document stores. The first of the potential options presented within this research is the use of tagging. The second solution will be discussed in section 8.4.4. Through tagging keywords can be assigned to relevant documents to help users retrieve documents using the keywords at a later date. Tagging has been shown to benefit organisations hoping to improve their ability to retrieve relevant information.

There were two key components to the use of tagging. Firstly, before tagging can be effective it is important to ensure that users understand the benefits of tagging and how to make effective use of this potential solution. The literature review found that many users would assume different standard practices when it came to tagging. Improving these practices, such as being aware of listing alternative lexical representations of concepts could help to ensure that different or returning users find more relevant information.

The questionnaire given to employees of SoftwareCo, as detailed in Chapter 5, can be used to help determine the barriers that may exist preventing users from making effective use of tagging. By identifying the barriers it may be possible to train users to make more effective use of tagging. Once the barriers to tagging have been reduced then tagging may be used to help users discover relevant information through any system that makes use of tagging. In addition to this, Chapter 5 also introduced TagDav. TagDav is a filing system, based upon the concept of tagging, which can be used for the organisation of both a user's personal and group documents. The system presented allowed users to make use of tagging for the storage and retrieval of all of their documents through normal file system interfaces and applications. This is something previously unavailable to

organisations and could help the organisation to further reduce their information overload problem by increasing access to relevant documents and information.

The TagDav system was well received by members of SoftwareCo. A number of employees at SoftwareCo stated that they often or sometimes had difficulty discovering files because they did not know which directory they were stored in. They stated that using the TagDav system would improve their access to information and on average they felt the system would allow them to save 40 minutes per day searching for information. In addition, ontologies can be used to help discover relevant information from personal and group document stores.

#### **8.4.4 Storing and Retrieving Information**

If information is stored within an organisation without the use of ontologies then both tagging and ontologies may be of use. In the literature review this research showed that ontologies could provide significant benefits to an organisation aiming to improve its ability to store and retrieve relevant information (Noy, McGuinness 2001) and that ontologies have become one of the key methods for representing knowledge as information within knowledge management applications (Brewster 2002).

The literature review also showed that although ontologies can provide significant benefits to organisations, they can be overly resourced and cost intensive, making them too expensive for many. This research provided a potential solution to help overcome the barriers to the creation of ontologies. Chapter 6 presented OntoFarm, a tool that makes use of a new approach to the creation of ontologies. The harvesting approach used by OntoFarm was introduced in Chapter 6 and can enable users to develop ontologies through the use of company portals as well as sites such as Google and Wikipedia. Employees at SoftwareCo felt that the OntoFarm tool would make it easier to discover concepts to add to an ontology and felt that the harvesting system would allow them to create ontologies in a far easier manner than before. The harvesting concept and OntoFarm was used by SoftwareCo to successfully develop a domain specific ontology in a greatly reduced timeframe compared to traditional approaches.

The OntoFarm system and semi-automated approach allows the creation of ontologies less resource than may previously have been possible and helps remove the barriers to ontology creation. With the barriers to the creation of ontologies reduced, organisations have a greater ability to make use of ontologies. Through improved use of ontologies the organisation can improve users' access to relevant information and help reduce the problem of information overload.

### **8.5 The Recommendations Framework Summary**

This framework provides recommendations on how access to each of the potential information sources can be improved and how this research can be adopted by organisations to help improve their access to relevant information. The framework can help an organisation to choose which information sources they wish to focus upon and then details how the organisation may attempt to improve the employees' access to relevant information.

Through implementing the potential solutions detailed within this thesis and summarised in the recommendations framework, it is possible to improve the discovery of relevant information. It will also aid an organisation in reducing the information overload problem faced by employees, providing better performance and a more positive information rich environment.

The information framework presented in this chapter itself addresses objective 7, "Establish an *information overload framework to provide direction and solutions to the information overload problem experienced by information workers.*"

### **8.6 Meeting the Aims and Objectives**

Section 8.4 has discussed each of the information sources identified by the framework. The information sources and the associated implications relating to information overload were the subject of the objectives of this research. Each chapter has proposed recommendations to allow organisations to reduce the problem of information overload.

summarises the chapters of this thesis and the objectives that each of the chapters fulfils.



**Table 23 - Objectives of this research and the chapter that meet these objectives**

| Chapter   | Objective No. | Objective Details  |
|---|---------------|--|
| 2 - Literature Review   | 1             | Critically review literature on information overload and other information defects and the effect it has upon information workers.   |
| 4 - Assessing the Knowledge Sharing Environment               | 2             | To establish through the use of a questionnaire the extent that multi-faceted barriers hinder information and knowledge sharing.   |
| 5 - Alternative Search Visualisation – Concept Clouds         | 3             | To determine how information overload can be reduced through the investigation and development of summarisation techniques.  |
| 6 - Using Tagging to Discover Networked and Local Information | 4             | To develop and assess alternative approaches to storing information to improve information retrieval and reduce information overload.  |
| 7 - Ontology Development                                      | 5, 6          | To establish the role ontologies can play in the retrieval of relevant information and reduction of information overload, the complexities of ontology development and the barriers to their use.<br><br>Investigate alternative approaches to traditional ontology development tools that may be used by subject experts rather than ontology specialists to aid in the creation of ontologies that can help the discovery of relevant information. |
| 8 - Conclusions and Recommendations Framework                 | 7             | Establish an information overload framework to provide direction and solutions to the information overload problem experienced by information workers.   |

## 8.7 Limitations of research and potential future work

There were a number of limitations to this research. This section shall take each of the key information source chapters along with the ontology's chapter and investigate the limitations of the research and the research approaches taken. The chapter section shall then present potential future work following the limitations.

### 8.7.1 Assessing the Knowledge Sharing Environment

Chapter 4 presented the questionnaire and traffic light system in order to assess the knowledge sharing that took place within an organisation. The organisations

that were chosen were both large multinational organisations and thus the potential exists to conduct the same study with a larger number of companies including smaller organisations.

In addition given that only two organisations have used the questionnaire and the difficulty in measuring the success of knowledge sharing it has not been possible to validate the questionnaire. The questionnaire was derived from work in the literature review. Further work could be performed to attempt to measure the benefits within an organisation once the questionnaire presented in Chapter 4 has been used and any recommendations put in place. This further work could also further validate that the questions themselves accurately discover impact of each of the potential barriers. Along with this with more time and resource available, along with organisations willing to take part it may be possible to establish whether other survey methods such as interviews and focus groups could benefit an organisation in determining the barriers that exist and reducing those barriers to knowledge sharing.

### **8.7.2 Extending Concept Clouds**

One of the key limitations with the research into information from intranets and the Internet, in Chapter 5, came from the advancement of search engines after the research was conducted. The query relevant text that is selected to show a snippet of an article within search results provides many of the benefits of the Concept Cloud. In order to determine the benefit of the Concept Cloud visualisation approach it would be necessary to reassess the system with the query relevant text present within search results. Although many corporate search systems still only present the documents title and description these are quickly being updated with information relevant to the search itself.

In addition the experiments within Chapter 5 were performed with undergraduate students. Although it was felt that these students would sufficiently generalise as they would be the information knowledge workers of the future validating the system within real organisations may also benefit the results.

Concept Clouds themselves offer a wide scope for future expansion. These future possibilities exist around a number of different aspects of the system. The first

option is to look at the words that are placed into the system and work upon them. The second area exists around the visualisation itself and how that may be changed. The third and final possibility exists around feedback and whether or not it is possible to learn from the system. The possibilities shall be discussed below.

Currently no effort is made to detect and combine plural and singular versions of words. During the literature review relating to tagging, this issue was discussed and the difference between the words 'apple' and 'apples' was given as an example. Whilst 'apples' most frequently relates to having more than one 'apple' the word 'apple' may mean the fruit or the Apple Corporation. There are also other grammatical options relating to this such as the handling of "Apple's". Although in tagging, the suggestion is that the different forms may have different meanings within a Concept Cloud. Where only an overview of an article is needed, converting the words so they all occur in the singular form may be of benefit. This is especially true since other words within the visualisation would also hint towards the intended meaning of the word. An investigation into this area would be necessary to see if any benefits could be found.

Similarly, the idea of stemming also occurred within the literature review. An investigation into whether terms such as 'snow', 'snowing', 'snowy', 'snows' could all be taken back to the word 'snow'. Research into whether or not a benefit would exist when doing this could yield interesting results.

Key phrase extraction rather than simple word-based tokenisation may also be a useful addition to the Concept Cloud system. In tagging based solutions manual tags are added to content multiple words may be expressed in a number of ways. However, in the Concept Clouds system words are tokenised by spaces. Although currently limited support exists for finding known phrases and replacing spaces with underscores, further investigation into key phrase extraction could form a welcome addition to the Concept Cloud system.

As stated previously, some options are available relating to the actual presentation of the word list. The first option comes from the improvements mentioned by Hassan-Montero and Herrero-Solana (2006). Hassan-Montero discussed improvements to tag clouds and those same improvements could be incorporated

into Concept Clouds. Similar concepts are grouped and placed upon the same line of the visualisation allowing the user to quickly see the topic areas of the content. Figure 53 shows an example of the suggested improvements.



**Figure 53 - The tag cloud after Hassan-Montero and Herrero-Solana's improvements**

Another study mentioned in the literature review by Sebrechts et al. (Sebrechts et al. 1999) mentioned that the use of colour coding similar concepts was the most frequently used and accepted feature of their interface. This idea could also be incorporated into the Concept Cloud system so that a different colour could be given to related groups of concepts.

The final set of optimisations or possibilities for expansion centre round the idea of feedback. The first idea is that users could click upon a word within the concept cloud in order to either add this word to the current search term or refine the search or to create a new search. This would allow users to quickly identify new keywords that could be used to perform additional searches and produce more accurate results. The other option surrounding feedback is to monitor the search terms that users are entering and then look at the content of the documents and the concept cloud that people choose in order to intelligently generate the concept cloud. The concept cloud could make use of statistics and related technologies in order to provide a more personalised and accurate concept cloud.

### 8.7.3 Tagging and TagDav Future Work

The research approach within Chapter 6 took two forms. The first was the use of questionnaires and the second the use of a focus group. The key issue was that the organisation would not allow transcription or recording of the focus group and it

was not possible to find a large enough sample to fully satisfy the requirements of the questionnaire. There is a large scope to further assess the potential of both tagging in general and the TagDav system.

The research into tagging in general and whether training and increasing awareness of the potential issues associated with tagging could be performed through the use of either surveys or experiments. Assessing a users understanding and knowledge and the way in which they tag both before and after training.

In order to more fully understand the TagDav system the study would need to be repeated. In addition if an organisation willing to try the TagDav system could be found then the system be fully implemented within an organisation and tested using experimental methods to assess the ability of a user to retrieve relevant information.

There are a number of opportunities for further exploration of the TagDav work explored in Chapter 6. One of the problems with this system, although it also presented opportunities, was that the system required users to upload files using a web page. The opportunities that this provided would allow integration of a number of systems to prevent users from adding tags that were malformed. The upload system could quite easily be modified to alert users when they entered plural tags, for example, show users how many times a tag had been used before. This system would allow for more efficient use of tags and help to solve some of the issues associated with a lack of training. The ability to add content to the system without having to upload via a web page would make an interesting area of further work. The WebDav protocol supports the ability to create files and folders in much the same way that a normal directory structure allows. Research into this area might be able to discover whether it would be possible to allow users to upload a file simply by placing it into a folder created for that file. For example, the user could create a new folder that used a comma to separate the list of tags that would be applied to the file and then the file could be placed directly into this folder. Research could investigate the best method to allow this type of file creation and tagging, and compare this to existing web-based upload procedures.

Another area that offers significant possibility for further development comes from one of the key issues identified within the system. The users of the focus group identified that different users may use different tags to describe the same concept. This would result in an inability to find content once it had been tagged. During the literature review, the concept of ontologies was shown in contrast to tagging, but there is opportunity for the two systems to work in synergy. There are a number of proponents within the target organisation for the concept of a hybrid built from both tagging and the use of ontologies. This concept has also been discussed in the literature (Barbosa 2008). It may be possible to expand the tag-based systems through the use of a domain specific ontology that has information relating to the tags. A further understanding of the tags that were chosen by the user could be used to automatically add tags to a file or even simply suggest more tags that could be used by looking at the relationships that might exist within ontology. This would require a full ontology to exist before this could be achieved. Other databases of concepts that are related to each other also exist such as those used by <http://www.semantichacker.com/> that may present an alternative approach.

#### 8.7.4 **OntoFarm Future Work**

Unfortunately Chapter 7 suffered from the same key limitation as Chapter 6. Although the system was demonstrated to undergraduate students who completed a questionnaire its exposure to industry could not be transcribed or recorded.

Repetition of this study both within large organisations and within smaller ones that may not have previously been able to make use of ontologies would be highly beneficial. Since the OntoFarm tool has also been used within an organisation to construct a full domain specific ontology in use within SoftwareCo it may also be possible to further examine the potential of the tool within organisations with lower barriers to entry.

With the popularity of the OntoFarm tool, shown in Chapter 7, there is significant potential for further development of the tool itself. One of the key concepts that arose was that many of the relationships could be extracted automatically. Work already exists within the field of automatically harvesting relationships from the World Wide Web (Ponzetto, Strube 2007), however this work takes a fully

automated approach. In order to ensure accuracy and completeness perhaps integrating automated relationship harvesting into the tool, to provide a semi-automated approach, may produce enhanced results. Using this natural language based parsing, there is significant opportunity to expand the system and really improve the quality of the concepts discovered and the relationships between these concepts. This relationship detection could also be of great use once the system already knows that a relationship exists between two concepts, but doesn't know the exact nature of the relationship. Natural language processing could also help to detect concepts in the form of phrases rather than single words that have simply been tokenised using spaces.

Although it was already mentioned previously, the proxy offers a number of great opportunities to embed content into the pages that the user is viewing when using the harvesting system. Further investigation into the use of the proxy server to enhance the content could be of great interest. For example, when selecting a concept that has been discovered from the page, the system could jump to the section of the page that the word was found and highlight that section.

Spell checking was never intended to be embedded into the OntoFarm system. It is possible to use the spell-checking feature of the browser and with a small modification, Firefox can be made to spell check both single and multi-lined textboxes. The issue is there is limited control over the spell checking facility. With spell checking built into the system there is a greater benefit to the end user. The spell checker could also help when searching to ensure the concepts that are being linked to, are not duplicated, because they are spelt incorrectly. An investigation into whether there is an advantage to including misspelt words as lexical representations of concepts may also be of benefit.

## **8.8 Final Summary**

In summary this research has provided a number of recommendations and approaches to enhance an organisations ability to reduce the problem of information overload. These recommendations and approaches have been derived from the research into knowledge sharing barriers in Chapter 4, visualisation approaches in Chapter 5, tagging in Chapter 6. In addition Chapter 7 provided a

solution to help increase the accessibility of ontologies within organisations. Along with the benefits to organisations each of the chapters has added to existing academic research. These benefits shall be outlined below along with the benefits to organisations.

The research investigated three key information sources. Information from colleagues, information from intranets or the Internet and finally information from users' own or group file stores.

Chapter 4 presented a method to assess and summarise the barriers to sharing knowledge within an organisation. The method was successfully used within two organisations and can help an organisation to identify where it should focus its efforts to improve knowledge sharing to increase the quality of information shared between colleagues. In addition, the chapter provided a method of being able to quickly identify these issues through the use of a traffic light system. Although previous research identified that organisations would differ with regards to the effect each potential barrier to knowledge sharing has upon the organisation, research did not exist in determining how organisations differed or which barriers affected specific organisations.

Chapter 5 presented Concept Clouds. The chapter furthered existing research into visualisation systems by providing a novel application of the Tag Cloud visualisation to create Concept Clouds. The Concept Cloud system presented a familiar visualisation to a different data source. By using a visualisation familiar to those exposed to the TagCloud the visualisation would improve adoption and provide a familiar interface. The visualisation can be used to increase the ability of a user to discover relevant information on a corporate intranet or Internet. Users of the Concept Cloud system achieved a two to twelve percent decrease in the time it took to discover relevant information and gave far fewer incorrect responses to questions helping to filter the irrelevant information.

Chapter 6 followed this theme to provide an alternative method for the discovery of a user's documents. The chapter presented TagDav a system that made use of tagging to provide a novel approach to the discovery of relevant documents. Users could tag documents and make use of these tags to retrieve the documents later



without any change to their existing workflows. The TagDav system was well received by members of the focus group within SoftwareCo who felt that the system could be of value and could potentially save significant amounts of time. Chapter 6 provided two key aspects for academia. The first is a concise list of potential pitfalls to avoid when using tagging. This was produced from a literature review and then a questionnaire was created to help identify and thus allow the prevention of these pitfalls. The second aspect provided by Chapter 7 was the TagDav file system. TagDav is a new form of file system that allows the use of tagging from within existing file systems and applications. This could allow a much broader range of research relating to tagging because the system does not change the users existing working environment.

In addition many of the approaches presented could benefit from the addition of ontology. The literature review showed that, although ontologies can be beneficial to organisations, the creation process can be difficult and resource intensive. Chapter 7 took research relating to Ontologies further by presenting a hybrid methodology for semi-autonomous ontology development. Fully automated approaches towards ontology creation also existed however the inaccuracy of these systems often made them unfeasible. The semi-automated approach helps by speeding up the development process whilst still allowing human influence to control the development leading to a less resource intensive ontology creation process. The tool that followed, OntoFarm, made use of this methodology to allow ontologies to be created quickly and easily based upon concepts harvested from the Internet and corporate websites. The OntoFarm tool was used to successfully develop a domain specific ontology within SoftwareCo and make ontology development less resource intensive for the organisation.

Finally Chapter 8 presented a framework that combined each of the approaches discovered within the previous chapters to present a method to reduce the information overload problem. Overall the research has provided a number of potential solutions to save employees time and improve their productivity. Although the research was only conducted within two organisations its implications are thought to be of value to any organisation wanting to reduce the

information overload problem and has contributed a number of findings with each of the relevant sections to academia.

## 9 References

- Ackoff, R.L. 1989, "From Data to Wisdom", *Journal of Applied Systems Analysis*, vol. 16, no. 3, pp. 9.
- Ackoff, R.L. 1967, "Management Misinformation Systems", *Management Science*, vol. 14, no. 4, pp. 147-156.
- Alavi, M. & Leidner, D.E. 2005, "Review: Knowledge Management and Knowledge Management Systems: Conceptual Foundations and Research Issues", *Knowledge Management*.
- Alavi, M. & Leidner, D.E. 1999, "Knowledge management systems: issues, challenges, and benefits", *Communications of the AIS*, vol. 1, no. 2es.
- Allen, T.J. 1978, *Managing the flow of technology*, MIT press Cambridge, MA.
- Argote, L. & Ingram, P. 2000, "Knowledge transfer: A basis for competitive advantage in firms", *Organizational behavior and human decision processes*, vol. 82, no. 1, pp. 150-169.
- Barbosa, D. 2008, *The taxonomy folksonomy cookbook*, Dow Jones Client Solutions.
- Bechhofer, S., van Harmelen, F., Hendler, J., Horrocks, I., McGuinness, D.L., Patel-Schneider, P.F. & Stein, L.A. 2004, "OWL Web Ontology Language Reference", *W3C Recommendation*, vol. 10, pp. 2001-2006.
- Begelman, G., Keller, P. & Smadja, F. 2006, "Automated Tag Clustering: Improving search and exploration in the tag space", *WWW2006, May*, pp. 22-26.
- Bellinger, G., Castro, D. & Mills, A. 2004, *Data, Information, Knowledge, and Wisdom* [2008, 09/01].
- Berners-Lee, T. & Fischetti, M. 1999, *Weaving the Web*, Orion Business Books.
- Bozsak, E., Ehrig, M., Handschuh, S., Hotho, A., Maedche, A., Motik, B., Oberle, D., Schmitz, C., Staab, S. & Stojanovic, L. 2002, "KAON—Towards a large scale semantic web", *E-Commerce and Web Technologies*, pp. 231-248.
- Brewster, C. 2002, "Techniques for automated taxonomy building: Towards ontologies for knowledge management", *Proceedings CLUK Research Colloquium* Citeseer.
- Brin, S. & Page, L. 1998, "The anatomy of a large-scale hypertextual Web search engine", *Computer Networks and ISDN Systems*, vol. 30, no. 1-7, pp. 107-117.
- Broder, A. 2002, "A taxonomy of web search".
- Bryman, A. & Bell, E. 2003, *Business research methods*, Oxford University Press Oxford; New York.

- Chen, H. & Dumais, S. 2000, "Bringing order to the web: automatically categorizing search results", *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 145-152.
- Chesney, T. 2006, "An empirical examination of Wikipedia's credibility", *First Monday*, vol. 11, no. 11.
- Christiaens, S. 2006, "Metadata mechanisms: From ontology to folksonomy... and back" in *On the Move to Meaningful Internet Systems 2006: OTM 2006 Workshops* Springer .
- Cornford, T. & Smithson, S. 1996, *Project Research in Information Systems: A Student's Guide*, Macmillan.
- Cross, R., Parker, A., Prusak, L. & Borgatti, S.P. 2001, "Knowing what we know:- Supporting knowledge creation and sharing in social networks", *Organizational dynamics*.
- Dalkey, N.C. 1969, "The Delphi method: An experimental study of group opinion",.
- Davenport, T.H. & Prusak, L. 1998, *Working Knowledge: How Organizations Manage What They Know*, Project Management Institute.
- De Kunder, M. 2008 , *WorldWideWebSize.com* [2008, 06/16].
- Denning, P., Horning, J., Parnas, D. & Weinstein, L. 2005, "Wikipedia risks", *Communications of the ACM*, vol. 48, no. 12, pp. 152-152.
- Dew, J.R. 1996, "Are You a Right-Brain or Left-Brain Thinker?", *Quality Progress Magazine*, Apr, pp. 91-93.
- Ding, L., Kolari, P., Ding, Z. & Avancha, S. 2007, "Using ontologies in the semantic web: A survey", *Ontologies*, pp. 79-113.
- Dixon, N.M. 2000, *Common Knowledge: How Companies Thrive by Sharing what They Know*, Harvard Business School Press.
- Dubie, D. 2006, Time spent searching cuts into company productivity, *Network World* [Online]. Available: <http://www.networkworld.com/news/2006/102006-search-cuts-productivity.html> [2010, 05/10]
- Eliot, T.S. 1934, *The Rock: a pageant play, written for performance at Sadler's Wells Theatre, 28 May-9 June 1934, on behalf of the forty-five churches fund of the diocese of London*, Faber & Faber.
- Eppler, M.J. & Mengis, J. 2003, "A framework for information overload research in organizations", *report, Universita Della Svizzera, Italiana* .
- Farhoomand, A.F. & Drury, D.H. 2002, "Managerial Information Overload", *Communications of the ACM*, vol. 45, no. 10, pp. 127.

- Farquhar, A., Fikes, R. & Rice, J. 1997, "The ontolingua server: A tool for collaborative ontology construction", *International Journal of Human-Computers Studies*, vol. 46, no. 6, pp. 707-727.
- Firestone, J.M. 2001, "Key Issues In Knowledge Management", *Knowledge and Innovation, a Journal of Knowledge Management Consortium International*, vol. 1, no. 3.
- Flickr 2006, , *Flickr*. Available: <http://www.flickr.com/photos/tags/> [2006, 12/01].
- Fricke, M. 2009, "The knowledge pyramid: a critique of the DIKW hierarchy", *Journal of Information Science*, vol. 35, no. 2, pp. 131.
- Galliers, R. 1992, *Information Systems Research: Issues, Methods, and Practical Guidelines*, Blackwell Scientific.
- Garshol, L.M. 2004, "Metadata? Thesauri? Taxonomies? Topic maps! Making sense of it all", *Journal of Information Science*, vol. 30, no. 4, pp. 378.
- Gazzaniga, M.S. & Sperry, R.W. 1967, "Language after section of the cerebral commissures", *Brain*, vol. 90, no. 1, pp. 131-148.
- Giustini, D. 2006, "How Web 2.0 is changing medicine", *British medical journal*, vol. 333, no. 7582, pp. 1283.
- Golder, S.A. & Huberman, B.A. 2006, "Usage patterns of collaborative tagging systems", *Journal of Information Science*, vol. 32, no. 2, pp. 198.
- Good, B.M., Tranfield, E.M., Tan, P.C., Shehata, M., Singhera, G.K., Gosselink, J., Okon, E. & Wilkinson, M. 2006, "Fast, cheap and out of control: A zero curation model for ontology development", *Pacific Symposium on Biocomputing*, pp. 128-139.
- Gruber, T. 2007, "Ontology of folksonomy: A mash-up of apples and oranges", *International Journal on Semantic Web & Information Systems*, vol. 3, no. 1, pp. 1-11.
- Gruber, T.R. 1995, "Toward principles for the design of ontologies used for knowledge sharing", *International Journal of Human-Computer Studies*, vol. 43, no. 5/6, pp. 907-928.
- Gruber, T.R. 1993, "A translation approach to portable ontology specifications", *Knowledge Acquisition*, vol. 5, no. 2, pp. 199-220.
- Guardian Newspapers Limited 2006, , *What is the Folksonomic Zeitgeist?*. Available: <http://blogs.guardian.co.uk/global/whatisfz.html> [2006, 12/01].
- Gulli, A. & Signorini, A. 2005, "The indexable web is more than 11.5 billion pages", *International World Wide Web Conference*, pp. 902-903.
- Halpin, H., Robu, V. & Shepherd, H. 2007, "The complex dynamics of collaborative tagging", *Proceedings of the 16th international conference on World Wide WebACM*, pp. 220.

- Hammond, T., Hannay, T., Lund, B. & Scott, J. 2005, "Social Bookmarking Tools (I)", *D-Lib Magazine*, vol. 11, no. 4, pp. 1082-9873.
- Hassan-Montero, Y. & Herrero-Solana, V. 2006, "Improving Tag-Clouds as Visual Information Retrieval Interfaces", *Merida, InSciT2006 conference* .
- Hayman, S. & Lothian, N. 2007, "Taxonomy Directed Folksonomies".
- Hearst, M.A. & Pedersen, J.O. 1996, "Reexamining the cluster hypothesis: scatter/gather on retrieval results", *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 76-84.
- Hemp, P. 2009, "Death by information overload.", *Harvard business review*, vol. 87, no. 9, pp. 82.
- Hepp, M., Bachlechner, D. & Siorpaes, K. 2006, "Harvesting Wiki Consensus-Using Wikipedia Entries as Ontology Elements", *First Workshop on Semantic Wikis*.
- Hepp, M. 2008, "Ontologies: State of the art, business potential, and grand challenges", *Ontology Management*, pp. 3-22.
- Herrmann, N. 1996, *The Whole Brain Business Book*, McGraw-Hill.
- Hey, J. 2004, "The Data, Information, Knowledge, Wisdom Chain: The Metaphorical link".
- Hölscher, C. & Strube, G. 2000, "Web search behavior of Internet experts and newbies", *Computer Networks*, vol. 33, no. 1-6, pp. 337-346.
- Horrocks, I. & Patel-Schneider, P. 2004, "Reducing OWL entailment to description logic satisfiability", *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 1, no. 4, pp. 345-357.
- Horrocks, I., Patel-Schneider, P.F. & van Harmelen, F. 2003, "From SHIQ and RDF to OWL: the making of a Web Ontology Language", *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 1, no. 1, pp. 7-26.
- Hu, J. 2005, "Yahoo buys photo-sharing site Flickr", *Cnet News.March*, vol. 20.
- Jacso, P. 2007, "Vivismo, Central Search, TIME Magazine, and the Open Directory Project", *ONLINE-WESTON THEN WILTON-*, vol. 31, no. 1, pp. 58.
- Jansen, B.J. & Spink, A. 2006, "How are we searching the World Wide Web? A comparison of nine search engine transaction logs", *Information Processing and Management*, vol. 42, no. 1, pp. 248-263.
- Janssen, R. & de Poot, H. 2006, "Information overload: why some people seem to suffer more than others", *Proceedings of the 4th Nordic conference on Human-computer interaction: changing roles*ACM, pp. 400.
- Joachims, T. 2002, "Optimizing search engines using clickthrough data", *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*ACM New York, NY, USA, pp. 133.

- Johnson, B. & Christensen, L.B. 2007, *Educational research: Quantitative, qualitative, and mixed approaches*, Sage Publications, Inc.
- Jones, Q., Ravid, G. & Rafaeli, S. 2004, "Information overload and the message dynamics of online interaction spaces: A theoretical model and empirical exploration", *Information Systems Research*, vol. 15, no. 2, pp. 194-210.
- Keller, P. 2008,, *deli.ckoma.net Del.icio.us stats*. Available: [http://deli.ckoma.net/stats#tab\\_tags](http://deli.ckoma.net/stats#tab_tags) [2008, 06/09].
- Kirsh, D. 2000, "A few thoughts on cognitive overload", *Intellectica*, vol. 1, no. 30, pp. 19-51.
- Kobayashi, T., Misue, K., Shizuki, B. & Tanaka, J. 2006, "Information gathering support interface by the overview presentation of web search results", *Proceedings of the Asia Pacific symposium on Information visualisation-Volume 60*, pp. 103-108.
- Krueger, R.A. & Casey, M.A. 2000, *Focus Groups: A Practical Guide for Applied Research*, Sage Publications Inc.
- Landeta, J. 2006, "Current validity of the Delphi method in social sciences", *Technological forecasting and social change*, vol. 73, no. 5, pp. 467-482.
- Leouski, A.V. & Croft, W.B. 1996, "An evaluation of techniques for clustering search results", *Technical Report IR-76, Department of Computer Science, University of Massachusetts, Amherst*.
- Macgregor, G. & McCulloch, E. 2006, "Collaborative tagging as a knowledge organisation and resource discovery tool", *Library Review*, vol. 55, no. 5, pp. 291-300.
- Madden, A.D., Eaglestone, B., Ford, N.J. & Whittle, M. 2006, "Search engines: a first step to finding information: preliminary findings from a study of observed searches.", *Information Research: an international electronic journal*, vol. 12.
- Mathes, A. 2004, "Folksonomies-Cooperative Classification and Communication Through Shared Metadata", *Computer Mediated Communication, LIS590CMC (Doctoral Seminar), Graduate School of Library and Information Science, University of Illinois Urbana-Champaign, December*.
- McFedries, P. 2006, "Technically Speaking: Folk Wisdom", *Spectrum, IEEE*, vol. 43, no. 2, pp. 80-80.
- McGuinness, D.L. 2002, "Ontologies come of age", *Spinning the semantic web: bringing the world wide web to its full potential*, pp. 171-192.
- McNabb, D.E. 2004, *Research Methods for Political Science: quantitative and qualitative methods*, ME SHARPE INC.
- Microsoft Microsoft Proposes Acquisition of Yahoo! for \$31 per Share-last update, *Microsoft Proposes Acquisition of Yahoo! for \$31 per Share*. Available:

<http://www.microsoft.com/presspass/press/2008/feb08/02-01CorpNewsPR.msp> [08, 05/21].

- Miles, M.B. & Huberman, A.M. 1994, *Qualitative Data Analysis: An Expanded Sourcebook*, Sage.
- Mukherjea, S. & Hara, Y. 1999, "Visualizing World-Wide Web search engine results", *Information Visualization, 1999. Proceedings. 1999 IEEE International Conference on*, pp. 400-405.
- Myers, M.D. & Avison, D.E. 2002, *Qualitative research in information systems*, SAGE London.
- Nahapiet, J. & Ghoshal, S. 2005, "Social capital, intellectual capital, and the organizational advantage", *Sumantra Ghoshal On Management: A Force For Good*.
- Nelson, M.R. 1994, "We have the information you want, but getting it will cost you!: held hostage by information overload.", *Crossroads*, vol. 1, no. 1, pp. 11-15.
- Nelson, R.R. & Winter, S.G. 1982, *An Evolutionary Theory of Economic Change*, Harvard University Press.
- Newman, B.D. & Conrad, K.W. 2000, "A Framework for Characterizing Knowledge Management Methods, Practices, and Technologies", *Proceedings of the Third International Conference on Practical Aspects of Knowledge Management. Basel, Switzerland*, pp. 1-16.
- Nonaka, I. & Takeuchi, H. 1995a, "Four Modes of Knowledge Conversion" in *The Knowledge-Creating Company: How Japanese Companies Create the Dynamics of Innovation* Oxford University Press, pp. 62-72.
- Nonaka, I. & Takeuchi, H. 1995b, *The Knowledge-Creating Company: How Japanese Companies Create the Dynamics of Innovation*, Oxford University Press.
- Northedge, R. 2007, "Google and beyond: information retrieval on the World Wide Web", *The Indexer*, vol. 25, no. 3, pp. 192-195.
- Noy, N.F. & McGuinness, D.L. 2001, "Ontology Development 101: A Guide to Creating Your First Ontology".
- Open Directory Project, *ODP - Open Directory Project*. Available: <http://dmoz.org/> [2007, 01/26].
- O'Reilly III, C.A. 1980, "Individuals and information overload in organizations: is more necessarily better?", *Academy of Management Journal*, pp. 684-696.
- O'Reilly, T. 2005, "What is Web 2.0: Design Patterns and Business Models for the Next Generation of Software".
- Orlikowski, W.J. & Baroudi, J.J. 1991, "Studying Information Technology in Organizations: Research Approaches and Assumptions", *Information Systems Research*, vol. 2, no. 1, pp. 1-28.



- Paek, T., Dumais, S.T. & Logan, R. 2004, "WaveLens: A new view onto internet search results", *Proceedings on the ACM SIGCHI Conference on Human Factors in Computing Systems*, pp. 727-734.
- Pan, J. & Horrocks, I. 2003 "Web ontology reasoning with datatype groups", *The SemanticWeb-ISWC 2003*, pp. 47-63.
- Peters, I. & Stock, W.G. 2007, "Folksonomy and Information Retrieval", *Proceedings of the 70th Annual Meeting of the American Society for Information Science and Technology*, vol. 45.
- Polanyi, M. 1967, "The Tacit Dimension", *New York* .
- Ponzetto, S.P. & Strube, M. 2007, "Deriving a Large Scale Taxonomy from Wikipedia", *Proceedings of the 22nd National Conference on Artificial Intelligence*, pp. 22-26.
- Rainie, L. 2007, *28% of Online Americans Have Used the Internet to Tag Content*.
- Remenyi, D. & Money, A. 2004, *Research supervision: for supervisors and their students*, Academic Conferences.
- Riege, A. 2007, "Actions to overcome knowledge transfer barriers in MNCs The Authors", *Journal of Knowledge Management*, vol. 11, no. 1, pp. 48-67.
- Riege, A. 2005, "Three-dozen knowledge-sharing barriers managers must consider", *Journal of Knowledge Management*, vol. 9, no. 3, pp. 18-35.
- Ruiz-Casado, M., Alfonseca, E. & Castells, P. 2006, "From Wikipedia to Semantic Relationships: a Semi-automated Annotation Approach", *First Workshop on Semantic Wikis*.
- Ryle, G. 1946, "Knowing How and Knowing That", *Proceedings of the Aristotelian Society*, vol. 46, pp. 1-16.
- Savolainen, R. 2007, "Filtering and withdrawing: strategies for coping with information overload in everyday contexts", *Journal of Information Science*, pp. 0165551506077418v1.
- Schick, A.G., Gordon, L.A. & Haka, S. 1990, "Information overload: A temporal approach", *Accounting, Organizations and Society*, vol. 15, no. 3, pp. 199-220.
- Schmitz, P. 2006, "Inducing ontology from flickr tags", *Collaborative Web Tagging Workshop at WWW2006, Edinburgh, Scotland, May*.
- Sebrechts, M.M., Cugini, J.V., Laskowski, S.J., Vasilakis, J. & Miller, M.S. 1999, "Visualization of search results: a comparative evaluation of text, 2D, and 3D interfaces", *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 3-10.
- Shadbolt, N., Berners-Lee, T. & Hall, W. 2006, "The Semantic Web Revisited", *IEEE INTELLIGENT SYSTEMS*, pp. 96-101.
- Sharma, N. 2004,, *The Origin of the Data Information Knowledge Wisdom Hierarchy*.

- Sivula, P., Van den Bosch, F. & Elfring, T. 2001, "Competence-Based Competition: Gaining Knowledge from Client Relationships", *Knowledge Management and Organizational Competence*, pp. 63-76.
- Smith, G. 2004, "Atomiq: Folksonomy: social classification", *Information Architecture*, vol. 3.
- Smith, S., Jackson, T. & Adelman, H. 2007, "Concept Clouds - Improving Information Retrieval", *8th European Conference on Knowledge Management*.
- Sood, S., Hammond, K.J., Owsley, S.H. & Birnbaum, L. 2007, "TagAssist: Automatic Tag Suggestion for Blog Posts", *International Conference on Weblogs and Social Media*.
- Spender, J.C. 1996, "Making Knowledge the Basis of a Dynamic Theory of the Firm", *Strategic Management Journal*, vol. 17, pp. 45-62.
- Sperry, R.W. 1984, "Consciousness, personal identity and the divided brain", *Neuropsychologia*, vol. 22, no. 6, pp. 661-673.
- Spink, A., Wolfram, D., Jansen, M.B.J. & Saracevic, T. 2001, "Searching the web: The public and their queries", *Journal of the American Society for Information Science and Technology*, vol. 52, no. 3.
- Tedmori, S., Jackson, T.W., Bouchlaghem, D., Adelman, H. & Nagaraju, R. 2006a, "Building a Tool for Expertise Discovery", *Emerging Trends and Challenges in Information Technology Management* Idea group Publishing, pp. 1053.
- Tedmori, S., Jackson, T.W., Bouchlaghem, N.M. & Nagaraju, R. 2006b, "Expertise Profiling: Is Email Used to Generate, Organise, Share or Leverage Knowledge", *11th International Conference on Computing in Civil and Building Engineering*, eds. H. Rivard, E. Miresco & H. Melham, , pp. 179.
- Teevan, J., Alvarado, C., Ackerman, M.S. & Karger, D.R. 2004, "The perfect search engine is not enough: a study of orienteering behavior in directed search", *Proceedings of the SIGCHI conference on Human factors in computing systems* ACM New York, NY, USA, pp. 415.
- Tvarozek, M. & Bielikova, M. 2008, "Personalized view-based search and visualization as a means for deep/semantic web data access".
- Vallet, D., Fernández, M. & Castells, P. "An Ontology-Based Information Retrieval Model", *2nd European Semantic Web Conference (ESWC 2005). Lecture Notes in Computer Science* Springer, pp. 455.
- Vivísimo 2006, , *Vivísimo Search Engine*. Available: <http://Vivísimo.com> [2006, 12/01].
- Vogel, D.R. & Wetherbe, J.C. 1984, "MIS research: a profile of leading journals and universities", *ACM SIGMIS Database*, vol. 16, no. 1, pp. 14.
- Walsham, G. 1995, "The emergence of interpretivism in IS research", *Information systems research*, vol. 6, no. 4, pp. 376-394.

- Walsham, G. 1993, *Interpreting information systems in organizations*, John Wiley & Sons, Inc. New York, NY, USA.
- Wang, X.H., Zhang, D.Q., Gu, T. & Pung, H.K. 2004, "Ontology based context modeling and reasoning using OWL", *Proceedings of the Second IEEE Annual Conference on Pervasive Computing and Communications Workshops* Citeseer, .
- Weisberg, H.F., Krosnick, J.A. & Bowen, B.D. 1996, *An introduction to survey research, polling, and data analysis*, Sage.
- Wen, J.R., Nie, J.Y. & Zhang, H.J. 2001, "Clustering user queries of a search engine", *Proceedings of the tenth international conference on World Wide Web*, pp. 162-168.
- Wilson, T.D. 2002, "The nonsense of knowledge management", *Information Research*, vol. 8, no. 1, pp. 8-1.
- Wilson, T.D. 2001, "Information overload: implications for healthcare services", *Health Informatics Journal*, vol. 7, no. 2, pp. 112.
- Wiza, W., Walczak, K. & Cellary, W. 2004, "Periscope: a system for adaptive 3D visualization of search results", *Proceedings of the ninth international conference on 3D Web technology*, pp. 29-40.
- Wurman, R.S. 1989, *Information anxiety*, Doubleday New York.
- Xie, X., Poshyvanyk, D. & Marcus, A. 2006, "3D visualization for concept location in source code", *International Conference on Software Engineering*, pp. 839-842.
- Zamir, O. & Etzioni, O. 1999, "Grouper: A Dynamic Clustering Interface to Web Search Results", *WWW8 / Computer Networks*, vol. 31, no. 11-16, pp. 1361-1374.
- Zeng, H.J., He, Q.C., Chen, Z., Ma, W.Y. & Ma, J. 2004, "Learning to cluster web search results", *Proceedings of the 27th annual international conference on Research and development in information retrieval*, pp. 210-217.
- Zhang, D. & Dong, Y. 2004, "Semantic, Hierarchical, Online Clustering of Web Search Results", *Proceedings of the 6th Asia Pacific Web Conference (APWEB)*, Hangzhou, China, April.
- Zmud, R.W. 1998, "Conducting and Publishing Practice-Driven Research", *Conference Proceedings of IFIP WG8*, pp. 10.

**10 Appendix One - Assessing the Knowledge Sharing  
Environment**

## 10.1 Questionnaire in PharmaCo

## Assessing the Knowledge Sharing Environment

Assess the Knowledge Sharing Environment of an Organisation.

### Basic Details

The following details are not used to identify you they are simply used to allow us to analyse the results

How old are you

Please choose \*only one\* of the following:

- Under 25  
 25-30  
 31-40  
 41-50  
 51-60  
 Over 60

How long have you worked at your current company

Please choose \*only one\* of the following:

- Less than a Year  
 1 Year  
 2-4 Years  
 5-9 Years  
 10 Years or more

### Technology

How would you rate yourself as a computer user

Please choose \*only one\* of the following:

- Expert  
 Experienced  
 Some Experience  
 Novice

How would you rate yourself with regards to using technology in general (e.g. a video recorder)

Please choose \*only one\* of the following:

- Expert  
 Experienced  
 Some Experience  
 Novice

When you are given a new piece of technology do you

Please choose \*only one\* of the following:

- Look forward to using it  
 Use it only when required  
 Become apprehensive about using it

How adequate do you feel the training you have received in using the software/technology you are required to use in your daily work is

Please choose \*only one\* of the following:

- 10%  
 25%  
 50%  
 75%  
 90% or More

Is sufficient training given when a new system is introduced?

Please choose \*only one\* of the following:

- Always  
 Often  
 Sometimes  
 Rarely

Do you feel the benefits of a new system over the old are clearly explained?

Please choose \*only one\* of the following:

- Always  
 Often  
 Sometimes  
 Rarely

Do you think that current IT tools and business processes are well integrated

Please choose \*only one\* of the following:

- Always  
 Often  
 Sometimes  
 Rarely

Are you given sufficient opportunity to give feedback on the suitability of Information Technology and Tools provided by the company?

Please choose \*only one\* of the following:

- Yes  
 No

Is there sufficient technical support available for the applications you use

Please choose \*only one\* of the following:

- Yes  
 No

Answer this question if you answered 'No' to question '9'

If NO the please explain why

Please write your answer here:

Do newly implemented systems live up to your expectations

Please choose \*only one\* of the following:

- Yes  
 No

By answer this question if you answered 'No' to question '10' ]

If NO please give examples

Please write your answer here:

Do you suffer from the lack of compatibility between IT systems?

Please choose \*only one\* of the following:

- Yes  
 No

By answer this question if you answered 'Yes' to question '11' ]

If YES please give examples

Please write your answer here:

Do you find it difficult to actually capture knowledge and know where to store information and knowledge

Please choose \*only one\* of the following:

- Yes  
 No

Please explain your answer, include any tools you may use

Please write your answer here:

### Organisational Factors

Do you feel you receive sufficient credit when sharing knowledge

Please choose \*only one\* of the following:

- Always  
 Often  
 Sometimes  
 Rarely

Please explain your answer

Please write your answer here:

By answer this question if you answered 'Sometimes' or 'Rarely' to question '1' ]

If Rarely or Sometimes does this make you reluctant to share knowledge in future

Please choose \*only one\* of the following:

- Yes  
 No

Are you given enough time to share knowledge

Please choose \*only one\* of the following:

- Yes  
 No

Do you feel you can record 'Knowledge Sharing' in your timesheets

Please choose \*only one\* of the following:

- Yes  
 No

By answer this question if you answered 'No' to question '3' ]

If you have difficulty please give suggestions

Please write your answer here:

Are you given enough time to meet and identify colleagues that have the knowledge YOU SEEK

Please choose \*only one\* of the following:

- Yes
- No

Please explain your answer giving examples or suggestions where necessary

Please write your answer here:

Are you given enough opportunity to meet and identify colleagues with a need for YOUR knowledge

Please choose \*only one\* of the following:

- Yes
- No

Please explain your answer giving examples or suggestions where necessary

Please write your answer here:

Which methods and/or tools do you use to identify people with the appropriate knowledge

Please write your answer here:

Have you benefited through sharing knowledge with others (including receiving knowledge from others)

Please choose \*only one\* of the following:

- Always
- Often
- Sometimes
- Rarely

Please give examples of any systems which aided this knowledge sharing for example discussion forums, email

Please write your answer here:

In your opinion what are the downsides of knowledge sharing

Please write your answer here:

Please enter any further issues

Please write your answer here:

Are there currently knowledge capture tools available within your organisation

Please choose \*only one\* of the following:

- Yes
- No

Answer this question if you answered 'Yes' to question '10'

If yes please describe any problems you may have with them

Please write your answer here:

### Daily Routine

How often do you make use of the web for work

Please choose \*only one\* of the following:

- All the time
- Hourly
- Once per day
- Once per week

Please enter some of the pages you visit most often and find most useful



Please write your answer here:

How frequently do you read the content on the company portal

Please choose \*only one\* of the following:

- Always
- Often
- Sometimes
- Rarely

Why answer this question if you answered 'Sometimes' or 'Rarely' to question '2' ]

If rarely or sometimes why do you not use it?

Please write your answer here:

How many emails do you send or receive

Please choose \*only one\* of the following:

- Less than 5 per day
- Around 5 - 10
- 10-25 per day
- 25+ per day

How frequently is Microsoft Outlook open in your working day

Please choose \*only one\* of the following:

- Always
- Often
- Sometimes
- Rarely

How frequently do you use Microsoft Word

Please choose \*only one\* of the following:

- Always
- Often
- Sometimes
- Rarely

How frequently do you use Microsoft Excel

Please choose \*only one\* of the following:

- Always
- Often
- Sometimes
- Rarely

Please list any other applications you frequently use

Please write your answer here:

### Organisational Sharing

Where do you believe knowledge is currently shared within your department

Please write your answer here:

Where do you believe knowledge is currently shared within your organisation

Please write your answer here:

Do you share knowledge outside your team

Please choose \*only one\* of the following:

- Yes
- No

Does your company make its Knowledge Sharing goals clear

Please choose \*only one\* of the following:

- Yes
- No

How regularly are you encouraged to share knowledge by your management

Please choose \*only one\* of the following:

- Always
- Often
- Sometimes
- Rarely

Do you share knowledge outside your team or group part of your work process

Please choose \*only one\* of the following:

- Yes
- No

Do you find it easy to actually share knowledge

Please choose \*only one\* of the following:

- Yes
- No

Are there enough formal (e.g. within meetings) and informal (e.g. coffee rooms) places to share, generate and reflect on new knowledge

Please choose \*only one\* of the following:

- Yes
- No

Do you feel you are given sufficient opportunity to interact with colleagues outside your immediate job, for example at conferences

Please choose \*only one\* of the following:

- Yes
- No

### Rewards and Recognition

Do you know of any reward schemes present to encourage the sharing of knowledge within your organisation

Please choose \*only one\* of the following:

- Yes
- No

By answer this question if you answered 'Yes' to question '1'

If yes do you feel these schemes offer sufficient reward to encourage Knowledge Sharing

Please choose \*only one\* of the following:

- Yes
- No

If no to either of the above questions please give suggestions

Please write your answer here:

Do you feel you are in competition with other people within your department

Please choose \*only one\* of the following:

- Yes
- No

Does your organisational reporting structure hinder Knowledge Sharing, for example knowledge is only shared between yourself and your manager

Please choose \*only one\* of the following:

- Yes
- No

Knowledge Management and Sharing were included within a yearly review process would you spend more time developing your skills in 'Knowledge Sharing'

Please choose \*only one\* of the following:

- Yes
- No

**Submit Your Survey.**  
Thank you for completing this survey..

## 10.2 Questionnaire in SoftwareCo

# Assessing the Knowledge Sharing Environment of the SoftwareCo Dept

A survey to assess the knowledge sharing within the SoftwareCo Dept

## Basic Details

The following details are not used to identify you they are simply used to allow us to analyse the results

How old are you

Please choose \*only one\* of the following:

- Under 25  
 25-30  
 31-40  
 41-50  
 51-60  
 Over 60

How long have you been affiliated with the SoftwareCo Dept

Please choose \*only one\* of the following:

- Less than a Year  
 1 Year  
 2-4 Years  
 5-9 Years  
 10 Years or more

Are you an employee of SoftwareCo

Please choose \*only one\* of the following:

- Yes  
 No

Are you a coach

Please choose \*only one\* of the following:

- Yes  
 No

What percentage of your time do you work for the SoftwareCo Dept

Please choose \*only one\* of the following:

- 10%  
 25%  
 50%  
 75%  
 90% or More

## Technology

When you are given a new piece of technology do you

Please choose \*only one\* of the following:

- Look forward to using it  
 Use it only when required  
 Become apprehensive about using it

How adequate do you feel the training or general information you have received in using the software/technology you are required to use in your daily work is

Please choose \*only one\* of the following:

- 10%  
 25%  
 50%  
 75%  
 90% or More

Is sufficient training or general information given when a new system is introduced?

Please choose \*only one\* of the following:

- Always  
 Often  
 Sometimes  
 Rarely

Do you feel the benefits of a new system over the old are clearly explained?

Please choose \*only one\* of the following:

- Always  
 Often  
 Sometimes  
 Rarely

Do you think that current SoftwareCo Dept tools and your business processes are well integrated

Please choose \*only one\* of the following:

- Always  
 Often  
 Sometimes  
 Rarely

Do you feel you are given sufficient opportunity to give feedback on the suitability of Information Technology and Tools provided by the SoftwareCo Dept?

Please choose \*only one\* of the following:

- Yes  
 No

Is there sufficient technical support available for the SoftwareCo Dept applications you use

Please choose \*only one\* of the following:

- Yes
- No

answer this question if you answered 'No' to question '9']

If NO the please explain why

Please write your answer here:

Do newly implemented systems live up to your expectations

Please choose \*only one\* of the following:

- Yes
- No

answer this question if you answered 'No' to question '10']

If NO please give examples

Please write your answer here:

Do you suffer from the lack of compatibility between IT systems?

Please choose \*only one\* of the following:

- Yes
- No

answer this question if you answered 'Yes' to question '11']

If YES please give examples

Please write your answer here:

Do you find it difficult to actually capture knowledge and know where to store information and knowledge

Please choose \*only one\* of the following:

- Yes
- No

Please explain your answer, include any tools you may use

Please write your answer here:

### Organisational Factors

Do you feel you receive sufficient credit when sharing knowledge in general

Please choose \*only one\* of the following:

- Always
- Often
- Sometimes
- Rarely

Please explain your answer

Please write your answer here:

answer this question if you answered 'Sometimes' or 'Rarely' to question '1']

Rarely or Sometimes does this make you reluctant to share knowledge in future

Please choose \*only one\* of the following:

- Yes
- No

Do you given enough time to share knowledge

Please choose \*only one\* of the following:

- Yes
- No

Do you feel you can record 'Knowledge Sharing' in the activity recorder

Please choose \*only one\* of the following:

- Yes
- No

answer this question if you answered 'No' to question '3']

if you have difficulty please give suggestions

Please write your answer here:

Are you given enough opportunity to meet and identify colleagues that have the knowledge YOU SEEK

Please choose \*only one\* of the following:

- Yes  
 No

Please explain your answer giving examples or suggestions where necessary

Please write your answer here:

Are you given enough opportunity to meet and identify colleagues with a need for YOUR knowledge

Please choose \*only one\* of the following:

- Yes  
 No

Please explain your answer giving examples or suggestions where necessary

Please write your answer here:

Which methods and/or tools do you use to identify people with the appropriate knowledge

Please write your answer here:

Have you benefited through sharing knowledge with others (including receiving knowledge from others)

Please choose \*only one\* of the following:

- Always  
 Often  
 Sometimes  
 Rarely

Please give examples of any systems which aided this knowledge sharing for example discussion forums, email

Please write your answer here:

In your opinion what are the downsides of knowledge sharing

Please write your answer here:

Are there currently knowledge capture tools available within the SoftwareCo Dept

Please choose \*only one\* of the following:

- Yes  
 No

Answer this question if you answered 'Yes' to question '10'

If yes please describe any problems you may have with them

Please write your answer here:

### Daily Routine

How often do you make use of the world wide web for work

Please choose \*only one\* of the following:

- All the time  
 Hourly  
 Once per day  
 Once per week

Please enter some of the pages you visit most often and find most useful

Please write your answer here:

How frequently do you read the content on the SoftwareCo Dept portal

Please choose \*only one\* of the following:

- Always
- Often
- Sometimes
- Rarely

If you answered 'Sometimes' or 'Rarely' to question '2', please answer this question if you answered 'Sometimes' or 'Rarely' to question '2'. If rarely or sometimes why do you not use it?

Please write your answer here:

How many emails do you send or receive

Please choose \*only one\* of the following:

- Less than 5 per day
- Around 5 - 10
- 10-25 per day
- 25+ per day

How frequently is Microsoft Outlook open in your working day

Please choose \*only one\* of the following:

- Always
- Often
- Sometimes
- Rarely

How frequently do you use Microsoft Word

Please choose \*only one\* of the following:

- Always
- Often
- Sometimes
- Rarely

How frequently do you use Microsoft Excel

Please choose \*only one\* of the following:

- Always
- Often
- Sometimes
- Rarely

Please list any other office applications you frequently use

Please write your answer here:

### Organisational Sharing

How do you believe knowledge is currently shared within your labs

Please write your answer here:

How do you believe knowledge is currently shared within the SoftwareCo Dept

Please write your answer here:

Do you share knowledge outside your capability area

Please choose \*only one\* of the following:

- Yes
- No

Has the SoftwareCo Dept made its Knowledge Sharing goals clear

Please choose \*only one\* of the following:

- Yes
- No

How regularly are you encouraged to share knowledge by your management

Please choose \*only one\* of the following:

- Always
- Often
- Sometimes
- Rarely

sharing knowledge outside your labs part of your work process

Please choose \*only one\* of the following:

- Yes
- No

do you find it easy to actually share knowledge

Please choose \*only one\* of the following:

- Yes
- No

are there enough formal (e.g. within meetings) and informal (e.g. coffee rooms) places to share, generate and reflect on new knowledge

Please choose \*only one\* of the following:

- Yes
- No

do you feel you are given sufficient opportunity to interact with colleagues outside your immediate job, for example at conferences

Please choose \*only one\* of the following:

- Yes
- No

### Rewards and Recognition

do you know of any reward schemes present to encourage the sharing of knowledge within the SoftwareCo Dept

Please choose \*only one\* of the following:

- Yes
- No

only answer this question if you answered 'Yes' to question '1'

if yes do you feel these schemes offer sufficient reward to encourage Knowledge Sharing

Please choose \*only one\* of the following:

- Yes
- No

if no to either of the above questions please give suggestions

Please write your answer here:

do you feel you are in competition with other people within the SoftwareCo Dept

Please choose \*only one\* of the following:

- Yes
- No

does your organisational reporting structure hinder Knowledge Sharing, for example knowledge is only shared between yourself and your resource coach

Please choose \*only one\* of the following:

- Yes
- No

Knowledge Management and Sharing were included within a yearly review process would you spend more time developing your skills in 'Knowledge Sharing'

Please choose \*only one\* of the following:

- Yes
- No

Submit Your Survey.

Thank you for completing this survey. Please submit by 2008-09-01.



## 11 Appendix Two – Assessing the Knowledge Sharing

### Environment Full Results

This section gives a more detailed investigation into the results of the questionnaire presented in Chapter 4

#### 11.1 Organisation One – PharmaCo

The results of the survey for PharmaCo have been presented under the same headings as the questionnaire, as detailed in section 4.4.1. The sections are Technology, Organisational Factors, Daily Routine, Organisational Sharing and Rewards and Recognition.

##### 11.1.1.1 Technology

The first key area examined was technology. As stated previously in section 4.4.1 participants were initially asked questions relating to their competence with technology in general.

Reluctance to use IT systems due to a lack of familiarity or experience is an issue for some employees (Riege 2005), however, in an organisation where the majority of employees feel that they have experience with computers one might expect that this would not be a problem.

Users were asked to state how experienced they were as a computer user. Two thirds stated that they were experienced as a computer user. Whilst 20% said that they had some experience and 12 said that they were experts. Only one employee (2%) stated they felt that they were a novice when it came to using a computer and the remaining 66% stated they were experienced. The employees' opinions of their experience with technology in general, followed a similar pattern. However, slightly more employees felt they only had some experience. When asked 11% said they were experts, 52% felt they were experienced and 36% felt that they had some experience.

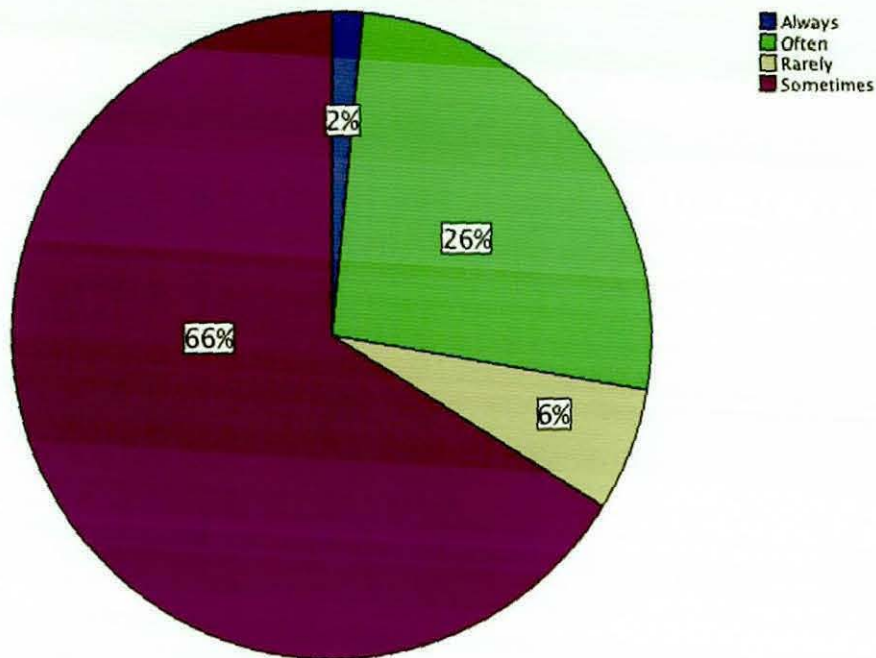
What is also quite positive is that 49% of the participants actually became excited about the prospect of something new. Only one employee felt apprehension about

the prospect of using a new piece of technology. The remaining 49% of users simply said they would use a new piece of technology when they were required to do so.

Employees were also generally positive about the amount of training they had received to perform tasks associated with their daily work. Twelve percent of them felt that their training was 90% or more adequate and 45% felt that their training was 75%+ adequate. This still left 30% stating that their training is only halfway towards being enough for them to complete tasks relating to their daily work. Leaving nine employees stating that the training they required to do their daily jobs was far from adequate. Further to these findings 40% of users felt that when a new system was introduced, only sometimes sufficient training was given, and almost 50% felt that sufficient training was often given.

With the assumption that all users had received the same level of training when systems were introduced, this may indicate that increased reflection and feedback is required. This would allow those who are perhaps slower at picking up the new technology or do not feel comfortable with the training, to obtain more references or training, whilst allowing those who are content with the training to get on with their job.

**Do you feel the benefits of a new system over the old are clearly explained**

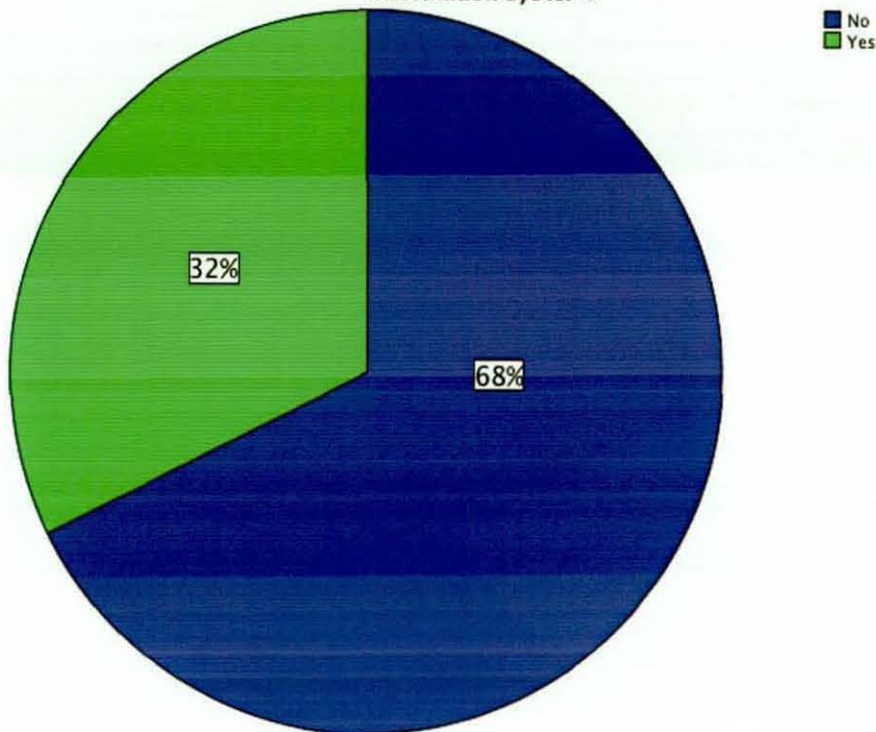


**Figure 54 - Benefits of a new system**

Another key issue mentioned by Riege (2005) is that not demonstrating all of the advantages of a new system over an old can cause negativity towards a system. This has also been identified as a problem within the PharmaCo with 6% saying that benefits are rarely explained and 66% saying that they are only sometimes made clear.

There was a feeling that information systems tools and business processes are not very well integrated with 61% saying that only sometimes they are well integrated and a further 20% stating that rarely were they integrated well. Building upon the lack of integration, two thirds of employees felt they were not given the opportunity to give feedback on the suitability of the information system provided. This perceived lack of integration along with the feeling that there is no opportunity to give feedback on the suitability of systems provided presents a serious issue that must be investigated further by the organisation. Research has shown that if users are not a part of the process when systems are created and that if systems are not well integrated then there will certainly be a negative reaction from users (Riege 2005).

**Are you given sufficient opportunity to give feedback on the suitability of Information Systems**



**Figure 55 - IS Feedback**

Two thirds of employees felt that sufficient technical support was available if needed. When asked why technical support was lacking, many employees felt that the technical support available was very impersonal, coming from call centres which appeared anonymous, rather than face-to-face help being available. Employees also reported that support was often not available when they needed to perform tasks within the applications they had to use. Many of the responses related to the concept of training rather than support being available to one user. Participants stated that, whilst sufficient technical support was available, if, for example, there was a problem with their computer, they could not get any help relating to a specific application. There appeared to be a strong feeling that what was actually required was 'top-up' training or even simply more training for the applications that those employees needed to use. One employee commented that learning was dependent upon one's self.

Fewer than 48% of employees also responded saying that newly implemented systems did not live up to expectations giving such reasons as systems being too complicated, not integrated well and still containing many errors. Some of the

comments suggested that people might be expecting too much from systems. For example, 'all systems should be completely integrated and should be tailored to suit their exact needs'. Whilst this represents an ideal solution it may be one that is simply not feasible. Perhaps a lack of understanding or communication could be the root cause in this case. It may be possible to address these issues through increased communication rather than trying to solve too many problems at once.

#### Do newly implemented systems live up to your expectations

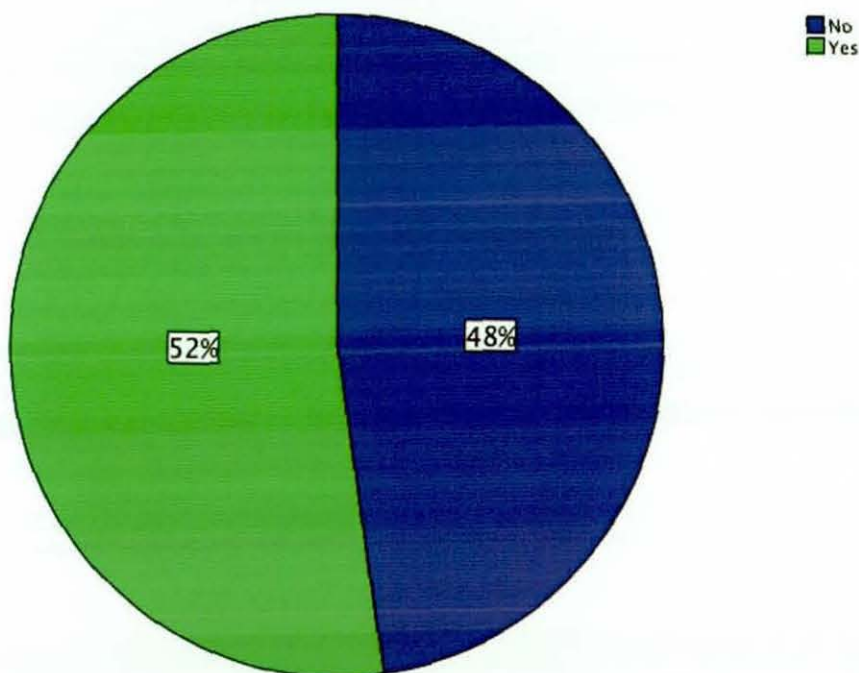


Figure 56 - Do newly implemented systems live up to expectations

#### 11.1.1.2 Organisational Factors

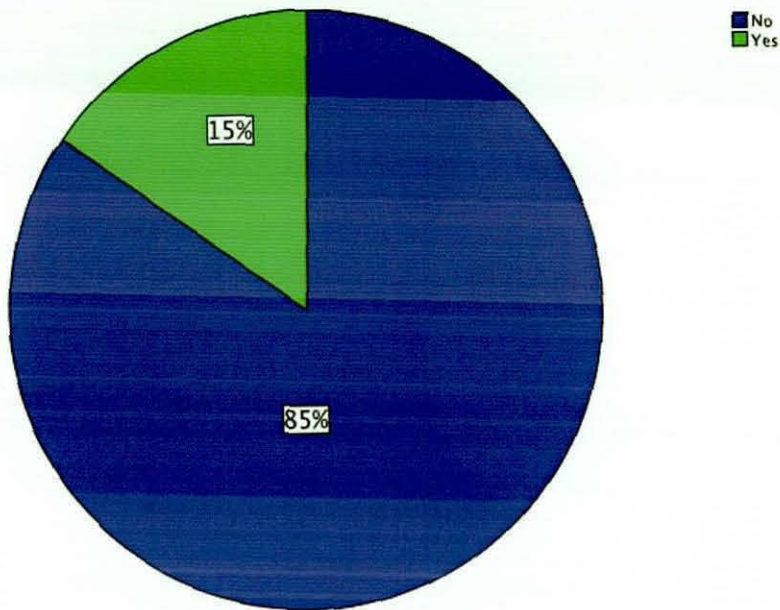
It has been stated in literature (Riege 2005) that employees can take "ownership of intellectual property because they do not feel they are given sufficient credit when sharing knowledge" (Riege 2005). This will obviously cause employees to be reluctant to share knowledge and cause them to keep knowledge to themselves, and only sharing what is truly necessary. Almost 50% of employees said they only sometimes receive credit and 25% said that they rarely receive credit for their knowledge sharing efforts.

It is important that employees still feel an emotional attachment to the knowledge they create or share even if the knowledge is being used by the team rather than the individual to whom it originally belonged. If they do not then the risk of them taking ownership of this information is increased. This may then lead to the users not sharing knowledge in future. Methods for giving credit are often difficult to implement because ideas and comments can come from a number of different sources and be used in a variety of ways. It is often impractical to attribute information to a specific employee throughout the lifecycle of that information. However, if managers and other team members make sufficient effort to credit employees, the efforts will often be greatly appreciated.

Only a small number of people mentioned malicious intent stating for example "it is not unusual to find that someone has run off and passed off as their own whatever it is that you provided". Most of the participants focused on the fact that although they were thanked by the individual who they shared their knowledge with, the organisation itself or a collective group did not acknowledge or thank the individual for their parted knowledge. Other employees acknowledged that knowledge transfer was often informal and that it would be impractical to credit someone for everything that they ever shared, or that if you were seen as someone who was frequently sharing knowledge by colleagues or the community as a whole, that it would be reflected in your appraisal and therefore would be acknowledged by the organisation.

What is quite surprising is although there was a lack of credit for knowledge sharing, 85% of participants stated this did not make them reluctant to share knowledge in the future. Although it appears that not receiving credit will not have a negative impact on the organisation, it may have an impact on the employees themselves. It seems that not only could the organisation make its goals clearer but also if regular sharers of information were highlighted or rewarded and information was credited whenever possible then perhaps employees would feel happier sharing their knowledge. The employee would feel confident that they would receive credit for the knowledge they have shared and would be far happier sharing knowledge. This again enforces the need for the rewards and recognition questions which were asked later on in the questionnaire.

**If Rarely or Sometimes does this make you reluctant to share knowledge in future**



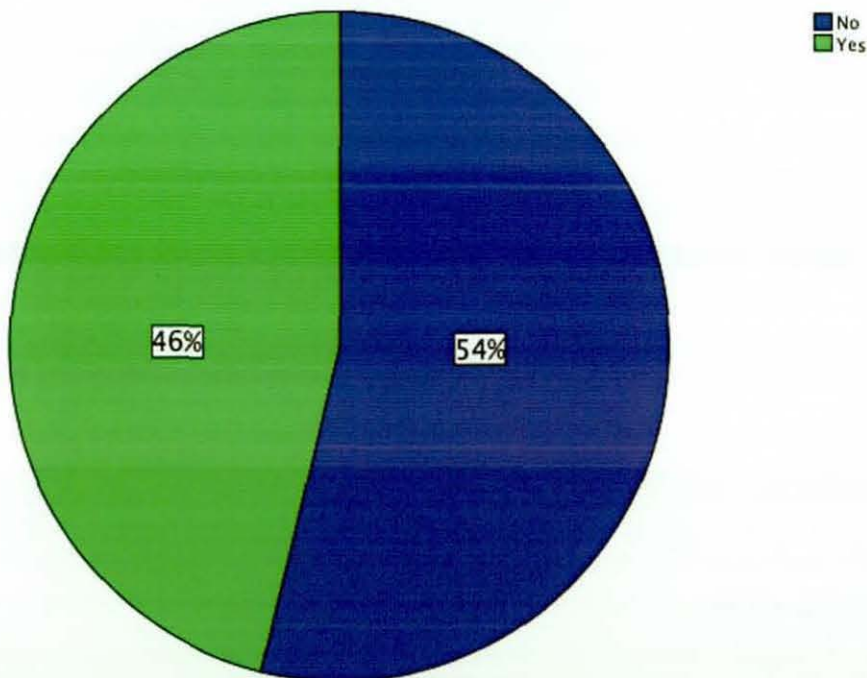
**Figure 57 - Does the lack of credit received make you reluctant to share knowledge in future**

A lack of time to share knowledge has been highlighted by employees with 12% saying they rarely received enough time to share knowledge and 56% said that only sometimes did they get enough time to share knowledge. Only two employees (3%) felt that they always had enough time to share knowledge. Further to this point over three-quarters of employees felt they could not record time that they had spent sharing knowledge in their timesheets. 23% of employees felt that it was possible.

Identifying employees to share knowledge with and employees who need your knowledge is also important. Benefits have been shown both through sharing within your organisation and across multiple organisations (Sivula, Van den Bosch & Elfring 2001). Fifty-three percent of employees said that they did not have time to identify employees who have knowledge that they require. A slightly higher number of respondents (61%) said that they did not have time to identify employees who may require their knowledge. Many employees quoted a simple lack of time as being the problem, although some actually stated other reasons for the problem, such as an inability to find people efficiently; insufficient tools

available to help them find other employees within the organisation; currently extremely difficult to know who was involved in which part of a particular project. Some employees did state that if they really took the time to search then it was usually possible to find the people that they required. This indicates that there is not necessarily a lack of expertise available but an inability to find expertise efficiently. When asked how employees identify other people with appropriate knowledge, employees were given the opportunity to write in an open text box with their comments. The majority of participants responded saying that they relied on 'word of mouth' and contacts they already knew to find others.

**Are you given enough time to meet and identify colleagues that have the knowledge YOU SEEK**



**Figure 58 - Graph – Are you given enough time to identify colleagues with the knowledge that you seek**

Literature also states that users must see the benefit in sharing knowledge (Riege 2005). If they do not see the benefits associated with sharing knowledge then users are less likely to share knowledge. Although this seems like a simple principle if someone has not benefited from sharing knowledge or even feels that sharing knowledge is not going to benefit them, then there is no incentive to do so. In this study 24% of employees felt that they had always benefited from sharing



knowledge and a further 49% felt that they often benefited from sharing knowledge. This did leave over one quarter of employees who felt that only sometimes or rarely did they benefit from sharing knowledge from others.

When asked about the downsides of sharing knowledge very few participants responded. Those that did respond chose a lack of trust by others taking credit for their work or a lack of recognition as the key problems. A fear of job security was not included as an issue from participants.

#### **11.1.1.3 Daily Routine**

In order to provide suggestions for methods and locations that information retrieval systems could be integrated, the daily processes and applications used by employees was assessed. If any possible development could integrate knowledge sharing into the existing tasks and work processes of employees, then adaptation time and disruption might be considerably reduced.

When asked how often employees made use of the web for work, 30% of participants stated all the time. A further 16% stated hourly, over one third stated once or twice per day and almost 11% stated once or twice per week. This shows that well over two thirds of employees used the World Wide Web at least once or twice per day.

When asked how often employees read the content on the company portal only one respondent stated always. Almost 28% stated that they would read it often and almost 48% would read it only sometimes. Many respondents stated that they simply did not have time to read the content on the portal. A wide variety of reasons were given for this ranging from the portal was not regularly updated or that the content did not appear to be relevant. Some users also felt that the portal was very difficult to navigate successfully. The portal would indicate a logical place to embed any information system that is to be developed by the organisation. However, the portal may not signify a prominent enough position and more questions may need to be asked in order to determine how the information system could be promoted more successfully than simply making it available from the corporate portal.

As was initially expected, 81% of employees said that they received 25 or more emails per day and 92% stated that they left Microsoft Outlook running all the time. Microsoft Word and Excel were also often used with 26% of employees stating they used Word always and 60% stating they used it often. Excel was used slightly less but the overall results showed that over 60% said they used it always, often or sometimes.

#### How frequently is Microsoft Outlook open in your working day

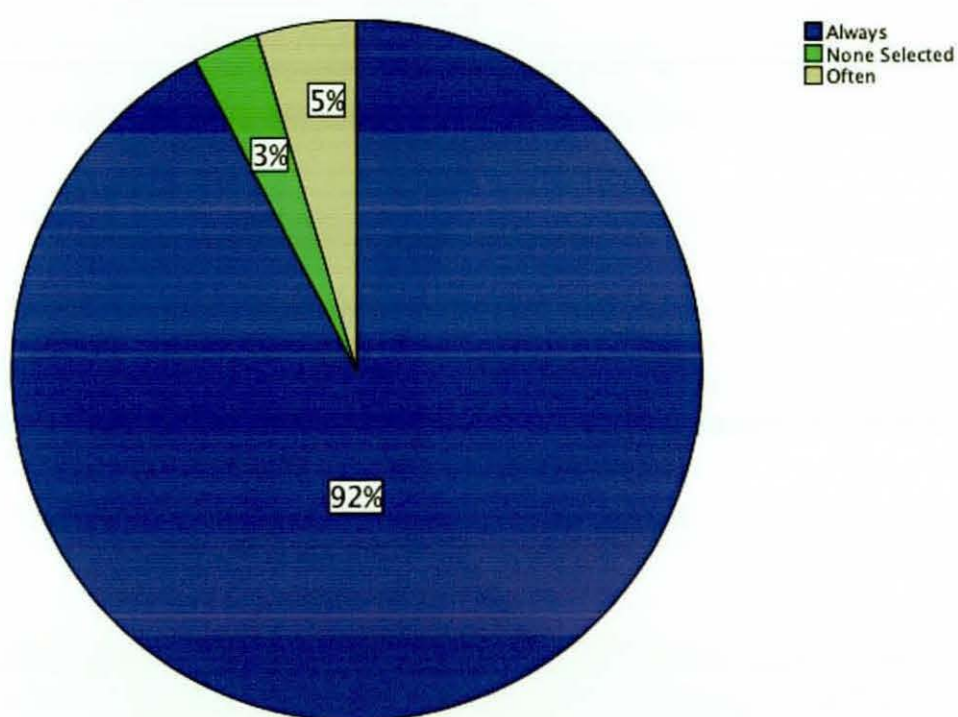


Figure 59 - Microsoft Outlook usage

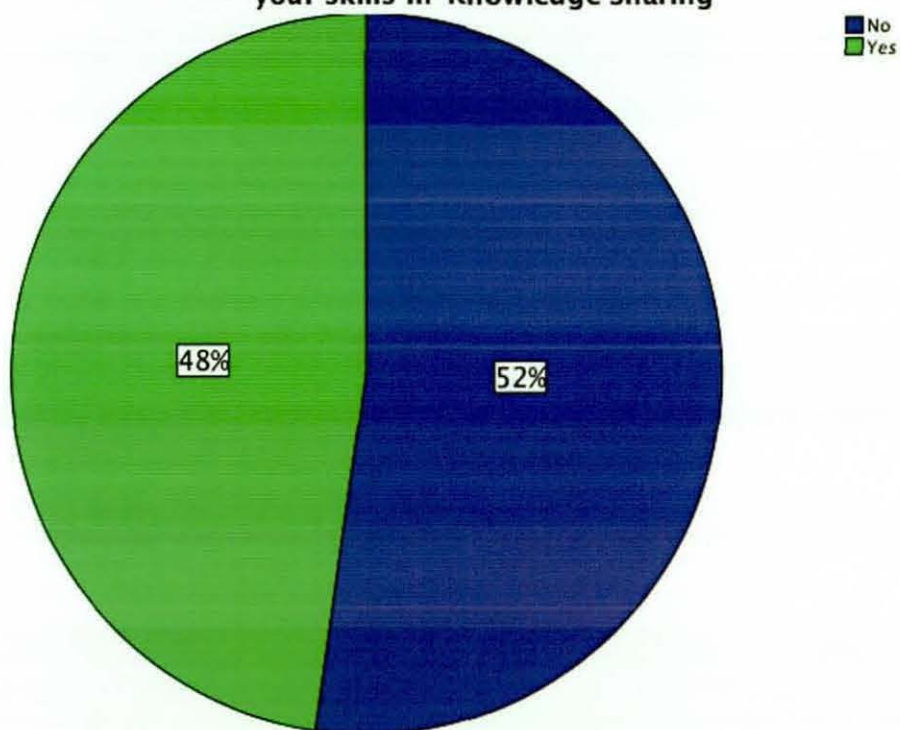
These MS Office applications appear to offer suitable locations to embed potential future information retrieval systems. These applications are in heavy use throughout the working day of most employees within the organisation and would offer a location that is always in use and accessible to employees.

#### 11.1.1.4 Organisational Sharing

The organisation structure and systems made available to employees can hinder potential sharing. This section looks at the results from these two areas.

When asked if knowledge was shared outside a participant's team, 81% said that it was, and over half felt that sharing knowledge outside their team was part of their job. Only 47% were aware of the company wide goals for knowledge sharing. Often the "Integration of KM strategy and sharing initiatives into the company's goals and strategic approach is missing or unclear" (Riege 2005). With almost half of the employees within this organisation unaware of the company's knowledge sharing goals and objectives, it may be difficult for a unified approach to succeed.

**If Knowledge Management and Sharing were included within the performance review process would you spend more time developing your skills in 'Knowledge Sharing'**



**Figure 60 - Companies knowledge sharing goals**

Whilst 40% of employees stated that they are often encouraged to share knowledge by their superiors and 13% always encouraged, this still leaves 32% who are only sometimes encouraged and 14% who felt that they rarely received encouragement to share knowledge. Over 56% stated that sharing knowledge outside of their team or group was something that they did as part of their job on a daily basis, leaving close to 43% who did not share knowledge outside of their team. Whilst for some employees there is no necessity to share outside of their team, there are benefits to be seen from doing so in many cases.

Sixty-five percent said that they found sharing knowledge easy. Almost 60% felt that there were an adequate number of places to interact formally and informally to share knowledge with colleagues, for example, within meetings and coffee rooms. Only one-third of respondents felt that they were given the opportunity to interact with colleagues outside of their immediate job, for example at conferences.

Also only 22% of employees felt that their organisational structure prevented them from knowledge sharing. For example, they felt that knowledge was only shared between themselves and their direct manager. Whilst this is not necessarily a real problem for the company as a whole, for those individuals, the sharing of knowledge is being suppressed.

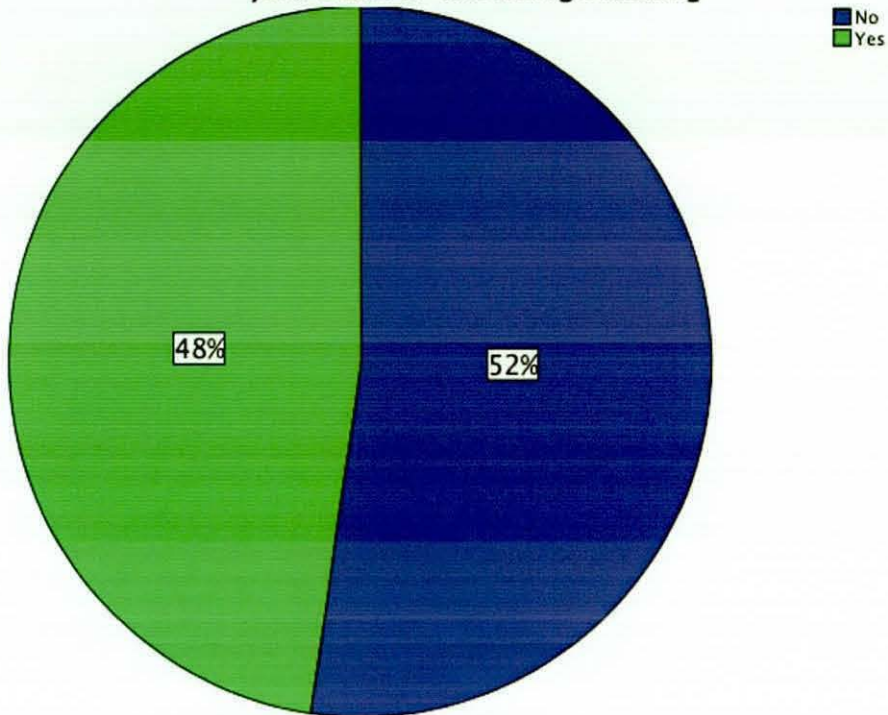
Only a small percentage of employees (35%) felt that they were in competition with employees both within and outside their department. This indicates that 65% do not feel they are in competition, which is a good outcome for knowledge sharing. However, competition can also be useful to successfully motivate employees.

#### ***11.1.1.5 Rewards and Recognition***

The final key indicator has been used to determine the mindset of employees within the department to whether they require a reward and recognition system, or if they currently use one and how it affects their work. It is clear that the majority (90%) of employees do not know of any reward schemes that currently run within the organisation. Yet the small minority that do (10%) feel that the scheme offers sufficient reward for knowledge sharing.

Finally, almost 50% of employees stated they would be encouraged to share knowledge if it were incorporated into their yearly review process. Whilst there was a reward scheme present within the organisation, an award available for sharing knowledge was clearly not advertised well or perhaps only known to a few people who were heavily involved with knowledge sharing.

**If Knowledge Management and Sharing were included within the performance review process would you spend more time developing your skills in 'Knowledge Sharing'**



**Figure 61 - If knowledge management and sharing and the review process**

## **11.2 Organisation Two – SoftwareCo**

Results from SoftwareCo have been divided into the same categories as the previous organisation, PharmaCo. The sections are Technology, Organisational Factors, Daily Routine, Organisational Sharing and Rewards and Recognition.

### **11.2.1.1 Technology**

Whilst participants in the survey from PharmaCo rated themselves with regards to their skills with computers and with technology in general, this question was not asked of participants from the second organisation, SoftwareCo. SoftwareCo employed experts within their fields in IT and programming. It was felt that asking this question might have been somewhat insulting to the participants. It was assumed that all users could be regarded as experts or at least have a high understanding of technology.

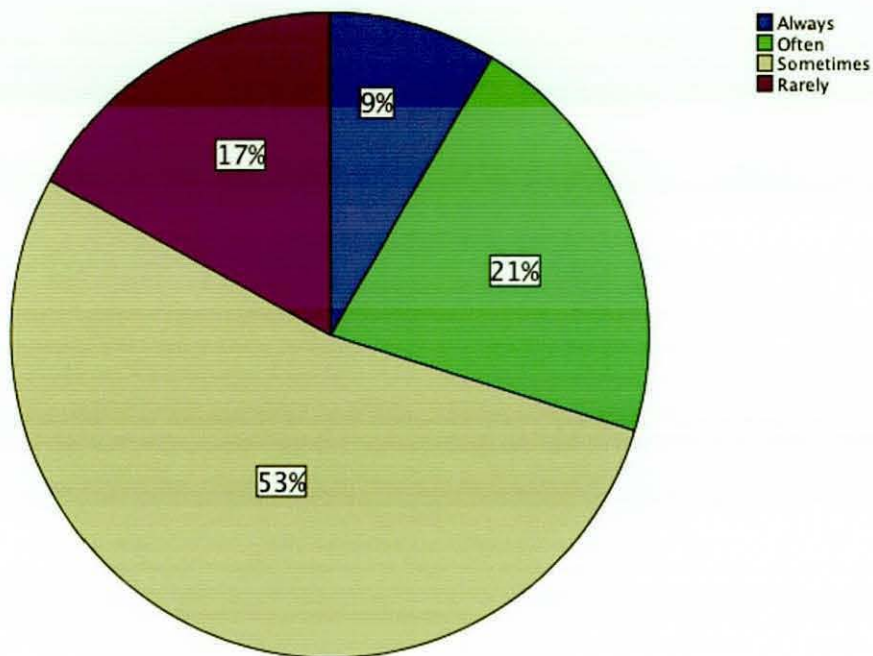
This high interest and understanding of technology was also reflected within a number of the questions within this section. One such example is that users were

asked how they would react to being given a new piece of technology. In this organisation 74% stated that they looked forward to using a new system, 22% stated that they would use it only when required and the remaining 4% became apprehensive about using new technology. The 4% means that of those asked, only two out of the 50 employees who answered this question felt apprehensive about trying a new piece of technology. This clearly highlights an environment more suited to accepting new technology. This small apprehension perhaps came from previous systems not living up to their expectations. It is important that any new system lives up to user's expectation in order to ensure its success. Within this organisation almost 76% stated that newly implemented systems did live up to their expectations. This is a surprisingly high number and is a very positive finding. Further to this, almost 68% stated that there was no lack of compatibility between systems.

Interestingly in this organisation only 11% of employees felt the training they were given to perform tasks with the software and technology associated with their daily work was 90% or more adequate. Twenty-nine percent stated training was 75% adequate, and 41% stated that training was 50% adequate. Clearly employees see a difference between the training that they feel they need and the training that they have received. The trend does show that more employees are happy with their training than those that feel they have not received sufficient training. One thing to note is that in the IT industry it can often be the case that employees are expected to learn certain elements as they work, as technology can move so quickly that it is hard to provide training that can keep up with the pace of change.

This is reflected in the answers given to whether sufficient training was given when a new system is introduced. Only 9% stated that sufficient training was always given, and 21% stated that it was often sufficient. The majority of employees stated that only sometimes sufficient training was given (53%) and 17% stated that rarely was sufficient training given. This highlights that over 70% of employees felt that normally the training was only sometimes or rarely sufficient. There is clearly a feeling within this organisation that training is insufficient, especially as new systems are introduced.

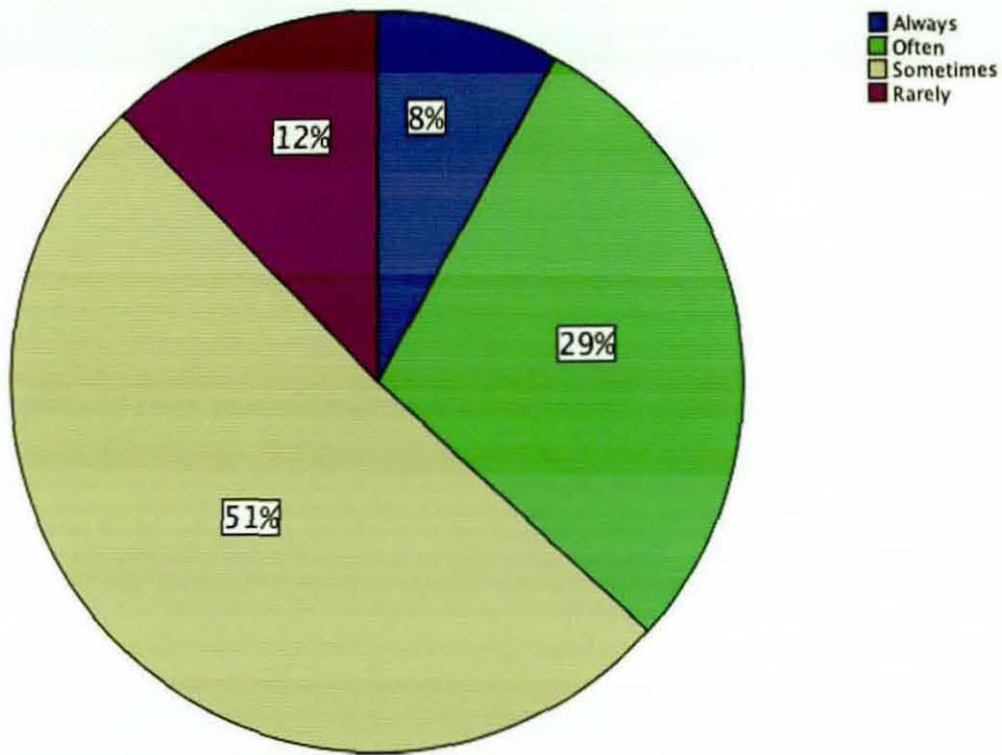
**Is sufficient training or general information given when a new system is introduced?**



**Figure 62 - Graph – Sufficient training for new systems**

As mentioned within PharmaCo's results, it has been stated that not clearly demonstrating the advantages of the new system over the old can cause negativity towards any new system (Riege 2005). Only 8% felt that the benefits of a new system were always explained and almost 29% felt that they were often explained. This left 51% of employees stating that benefits were only sometimes explained and 12% saying that they were rarely explained. This highlights that the benefits of new systems were not sufficiently explained within this organisation and this can have a negative impact on employees.

**Do you feel the benefits of a new system over the old are clearly explained?**



**Figure 63 - New system benefits**

The questionnaire found that business processes and the current tools available within the organisation were integrated well. There is still work to be done in integrating these tools, as tools being rarely integrated and sometimes integrated were both chosen by 12% of participants. However, 43% felt that tools were often integrated and almost 9% stated that they were almost integrated

These results may, in a large part, be due to the fact that people within the organisation felt that they did have sufficient opportunity to feedback upon information technology and tools provided by the organisation. Seventy percent of employees felt that this was the case. This still leaves 30% of participants that did not feel that this was the case and there is certainly room for improvement, but it is a positive result. Almost all employees (just over 91%) also felt that there was sufficient technical support available should they need it, under 9% people actually stated they had issues with technical support.

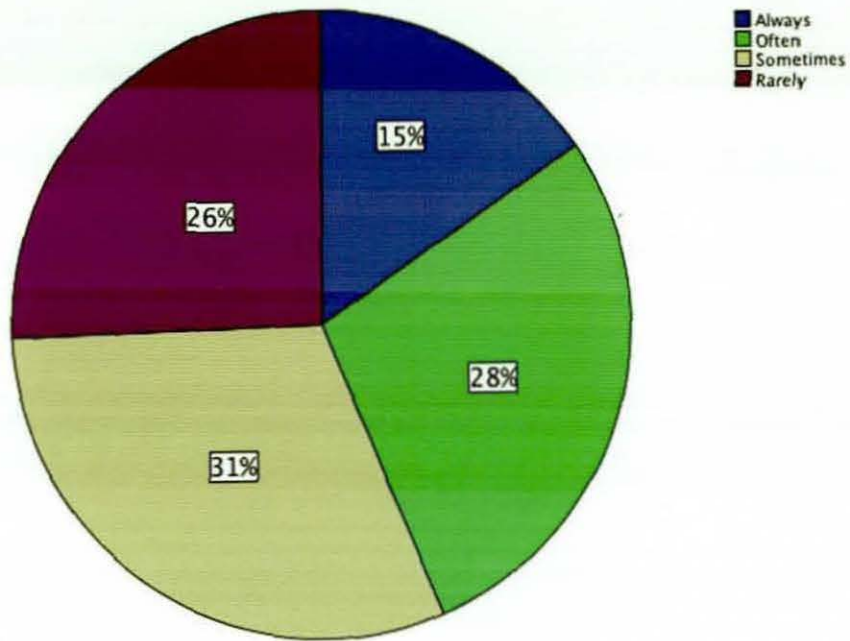


### ***11.2.1.2 Organisational Factors***

When asked whether employees received sufficient credit when sharing knowledge in general 15% stated always, 28% stated often, 31% stated sometimes and 26% stated rarely. This actually shows that over 56% of employees felt that they did not receive sufficient credit for the knowledge that they shared. This highlights a key area that should be addressed. Although only six people, or one third of the employees who stated they did not receive sufficient credit said that this would make them reluctant to share knowledge in the future.

When asked whether they were given enough time to share knowledge 68% answered yes. This meant that only 32% of employees felt that they did not have enough time to share knowledge. When employees were asked whether they felt they could record knowledge sharing within their timesheets, one third of participants answered yes and two thirds no. This shows a clear issue, if people are not able to record the time which they spend sharing knowledge then the time that gets devoted to sharing knowledge is likely to be reduced. One element that must be highlighted is that the organisation had recently introduced an activity-recording piece of time management software. Reviewing the free text results for this question it was apparent that many of the users interpreted this question as relating to this software. This may have skewed the results that were given.

Do you feel you receive sufficient credit when sharing knowledge in general



If Rarely or Sometimes does this make you reluctant to share knowledge in future

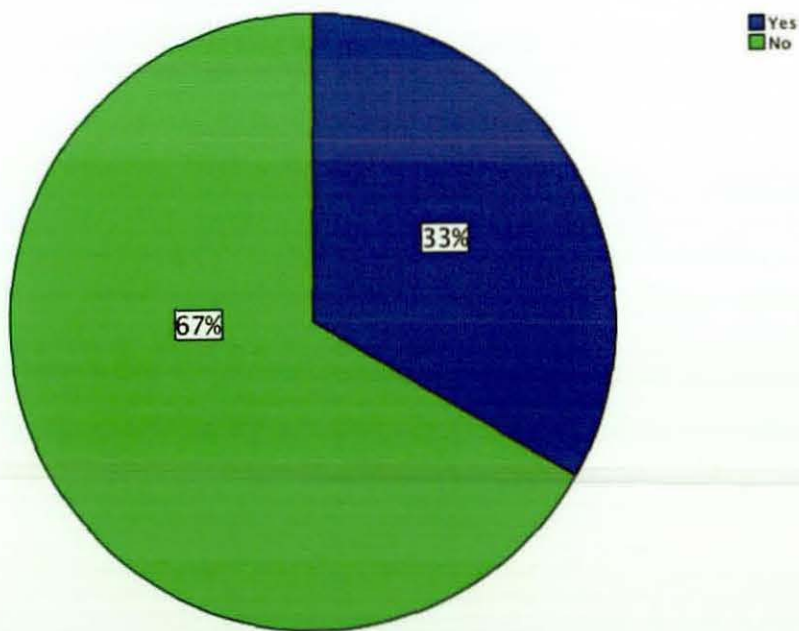


Figure 64 - Sufficient credit and does this make you reluctant to share

Identifying colleagues to share knowledge with and those who require your knowledge is an important part of being able to share knowledge. The majority

(68%) of employees felt that they were given adequate opportunity to identify colleagues with knowledge that they required. Whilst this does mean 32% of employees lacked the opportunity to identify the colleagues that had the knowledge that they were seeking, it is a positive number of employees. When asked if employees were given the opportunity to find colleagues that needed their knowledge, the number was slightly lower, with 58.8% of employees satisfied. Some users stated that they could always ask other colleagues who could put them in touch with others. Generally there were no negative comments from users in this area. One suggestion was to create a more formal structure to achieve the aim of sharing knowledge.

When asked what methods and tools employees used to identify experts within the organisation there were some interesting answers. One of the key answers was a tool specifically created by the organisation to help employees identify others based upon a variety of factors including the project they worked on, their subject area and any skills sets that they chose to list. Private expert networks within employees own minds and asking others also featured highly.

When asked whether employees saw the benefits associated with sharing, 32% stated always and 42% stated often. Not many users could not see the benefits of sharing 18% said only sometimes could they see a benefit and just three people (8%) stated they rarely could. Whilst this is clearly positive result, with just over one quarter stating that they don't regularly see the benefits of sharing knowledge, there are still a large number of employees for which this needs be addressed.

When asked about the downsides of sharing many people simply stated that there were no downsides. A minority quoted that it can be a time consuming process or that there could be a lot of effort with no visible benefit. One participant also mentioned, "too much information can slow decisions and actions" which is an interesting comment.

#### **11.2.1.3 Daily Routine**

The daily processes and applications used by employees were discussed within the questionnaire in order to determine where possible development and integration of any future information retrieval systems could occur in order to minimised

disruption to employees. Participants within this questionnaire were predominantly computer programmers or worked within the field of IT and thus answers given to questions such as how often they made use of the Internet may not reflect those expected in other organisations.

When asked how often employees made use of the World Wide Web for their work, 51% said all the time, 23% said hourly and 18% answered once per day. Only 8% answered once per week. Showing that over 90% of the organisation used the web on a daily basis.

However, when asked whether employees made use of the content on the departmental portal, one quarter said often, 35% said sometimes and 40% said rarely. None of the participants stated that they always read the portal. This is a very negative reflection with over three quarters of users stating that they only sometimes or rarely read the portal site. Participants of the questionnaire were also asked why they did not read the content on the portal. Many employees stated that information was not relevant or was not updated frequently enough or even that it lacked structure. One interesting comment was that they only had to look at the system sometimes because it was rarely updated, which gives a slight implication that should there be more content that employees would spend more time reading it. This does not detract from the fact that some employees stated that information could be irrelevant or outdated.

An unexpected result was that only 33% of employees within the organisation stated that they sent or received more than 25 emails per day. Thirty-nine percent stated they sent or received 1-25 per day and 23% said they sent or received around 5-10 per day. Only two employees, or 5%, sent or received less than five per day. This number was lower than expected but still represents a large number of emails being sent each day. Two possible explanations for the lack of emails being sent are that many of the employees within this organisation would see each other face-to-face on a frequent basis and also that many users also use an instant messaging service to communicate with each other.

### How frequently is Microsoft Outlook open in your working day

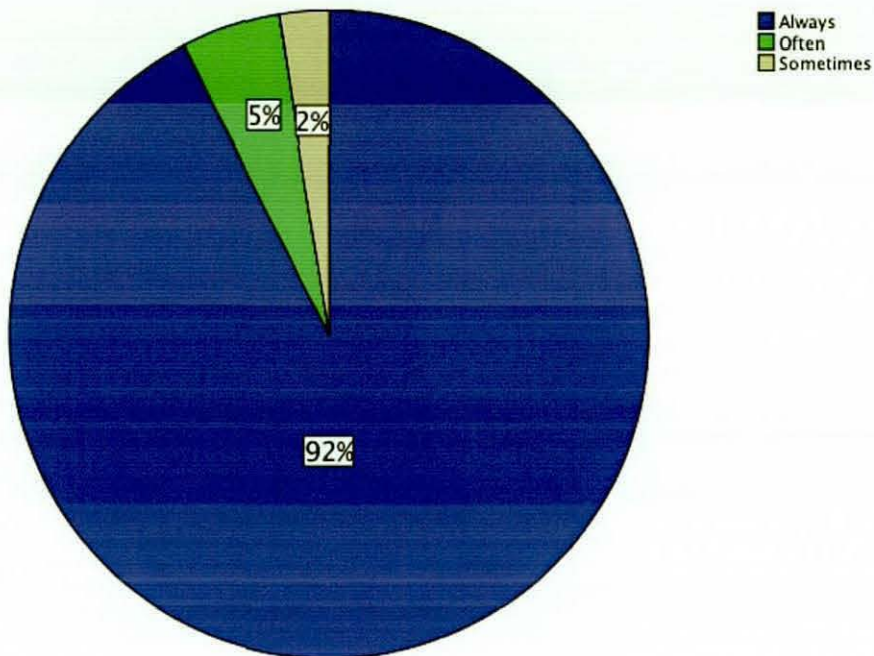


Figure 65 - Use of outlook

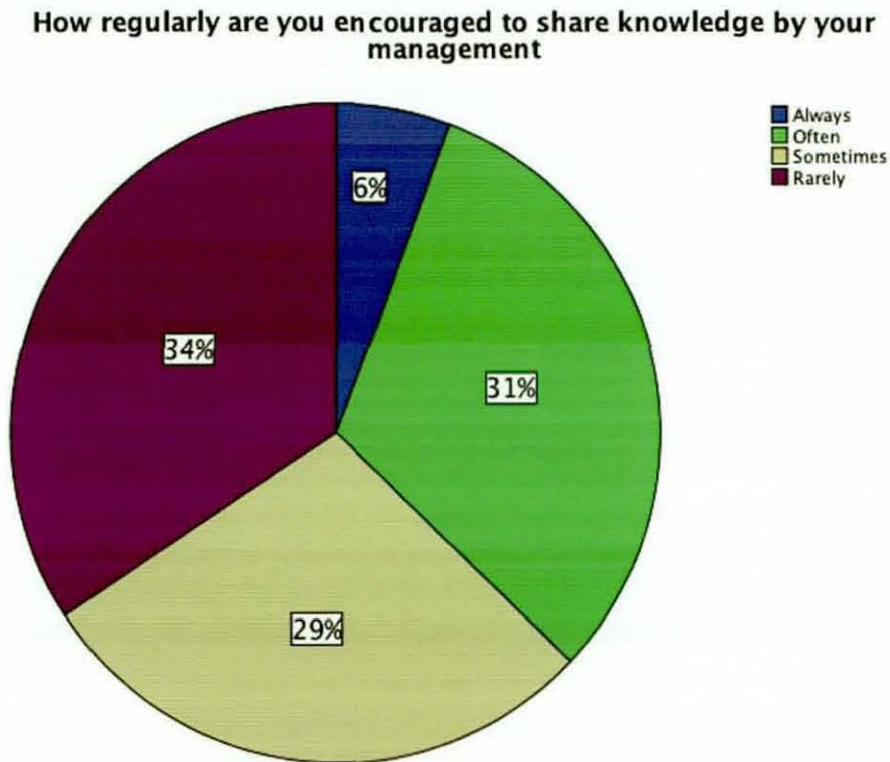
For 93% of users Outlook was open always during their day. A further 5% said often and the final 3% stated sometimes. Like the previous organisation this represents a definite area for further investigation. Microsoft Word and Excel were also in frequent use. Sixty-five percent of employees 65% stated that they always or often used Word and 63% of users stated that they always or often used Excel. Only 7% rarely used Word and 8% rarely used Excel.

#### 11.2.1.4 Organisational Sharing

When asked if knowledge was shared outside a participant's capability area, which is akin to their team in some organisations. Of the participants 79% said yes and exactly half felt that sharing knowledge outside their labs was part of their job. Only 32% were aware of the company wide goals towards knowledge sharing. A total of 28 participants answered this question.

When asked how frequently employees were encouraged to share knowledge by their management, only 6% said always, and 31% stated often. However, 29% and 34% stated sometimes and rarely respectively. This clearly highlights a potential barrier. Half of the participants who answered found it easy to share knowledge, as

63% stated that there were enough formal and informal places to share, generate and reflect on new knowledge. Less than half (45%) of employees felt that they were given sufficient opportunity to interact with colleagues outside their immediate jobs, the example given was at conferences.



**Figure 66 - Management Encouragement**

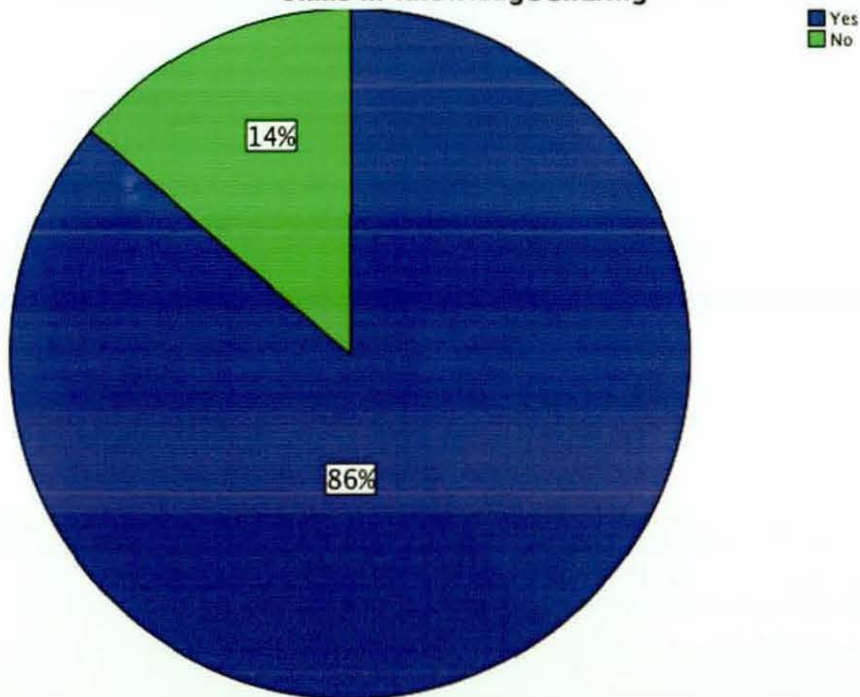
Very few employees felt that the organisational structure hindered knowledge transfer. There is a risk that information, for example, may only be transferred between them and their managers. In SoftwareCo only 14% felt that this was the case and they were hindered by their organisational structure. Only 35 percent of users felt that they were in competition with other employees within the organisation.

#### **11.2.1.5 Rewards and Recognition**

As with the first organisation, PharmaCo, the final key indicator was rewards and recognition. This was used to determine whether knowledge sharing could be affected by rewards or recognition.

When asked whether there were any rewards schemes present 91% stated that there were not. Of those three (9%) employees that felt that there were reward schemes available one stated (33%) that they were sufficient to encourage sharing. One stated (33%) that they were not and one (33%) did not answer the second question. When asked to give suggestions some people mentioned adding knowledge sharing to their key performance indicators or as part of bonus schemes. One employee suggested that it was not necessary, as it was part of employee's job. Another useful suggestion was to create an award for useful information shared or useful whitepapers passed to internal employees.

**If Knowledge Management and Knowledge Sharing were included within a yearly review process would you spend more time developing your skills in Knowledge Sharing**



**Figure 67 - Knowledge sharing as part of a yearly review process**

Finally 86% of the questionnaire participants stated that if knowledge management and sharing were incorporated into their yearly review process they would spend more time developing their skills within this field.

## 12 Appendix Three - TagDav Questionnaire



## Tagging In General (Part One)

1. Have you tagged content before (using sites such as Del.icio.us or Flickr)?
  - a. Always, Often, Sometimes, Rarely, NeverIf never go to question 9
2. How many tags on average will you use to tag content per item?
  - a. 1 - 2, 3 - 4, 5 - 6, 7 - 8, 8+
3. Do you try to re-use tags that **you** have used before?
  - a. Always, Often, Sometimes, Rarely, Never
4. Do you try to re-use tags that **others** have used before?
  - a. Always, Often, Sometimes, Rarely, Never
5. Do you use plural tags, singular tags or a mixture of both?
  - a. Plural, Singular, Mixture
6. Do you enter spaces when tagging content?
  - a. Yes, No, Mixture
7. Do you consider capitalisation of tags important?
  - a. Yes, No
8. Do you enter synonyms of words in the hope that others will be able to use these synonyms when retrieving content that you have tagged?
  - a. Always, Often, Sometimes, Rarely, Never
9. Do you ever have difficulty finding files on your own because you do not know which directory they are in?
  - a. Always, Often, Sometimes, Rarely, Never
10. Do you ever have difficulty finding files other people have created because you do not know where they are stored?
  - a. Always, Often, Sometimes, Rarely, Never
11. Do you ever use the search functionality of a file system to find files?
  - a. Always, Often, Sometimes, Rarely, Never

12. If Yes do you find that this is sufficient to effectively find the files that you want?

a. Always, Often, Sometimes, Rarely, Never

13. Do you ever place the same file in more than one place (or create symbolic links) so that it is easier to find?

a. Always, Often, Sometimes, Rarely, Never

### Tag Based Filing

14. Do you think that tagging files would be of benefit to help find files without the need for search engines?

a. Definatly, Maybe, Sometimes, Rarely, Never

15. Do you think that time can be saved by tagging files and then using these tags to retrieve them?

a. Definatly, Maybe, Sometimes, Rarely, Never

16. Do you have your own storage procedure to make it easier to find files?

a. Always, Often, Sometimes, Rarely, Never

17. Do you store files by year, project, lab, customer, topic (Choose all or add any additional)?

a. Year, Project, Lab, Customer, Topic, Recipient

b. Other \_\_\_\_\_

18. Would you use this system to replace your current way of filing and retrieval?

a. Definatly, Maybe, Sometimes, Rarely, Never

19. Do you think that this system has benefits over a traditional directory based filing system?

a. Definatly, Maybe, Sometimes, Rarely, Never

20. Is browsing a file structure to place the file into harder than simply uploading it and tagging it?

a. Yes, No

21. How much time do you think you could save using this system (Per Week)?

a. \_\_\_\_\_ Minutes

22. Is this system easy to use and understand?

a. Yes, No

23. Do you think training would be required before people can use the system at all?

a. Always, Often, Sometimes, Rarely, Never

24. Do you think training would be required before people can use the system to its full potential?

a. Always, Often, Sometimes, Rarely, Never

## Strengths, Weaknesses, Opportunities and Threats

Please analyse the tool and place your comments into the relevant boxes:

**Strengths:** The advantages of the tool (i.e. what does the tool do well?)

**Weaknesses:** The disadvantages of the tool (i.e. what does the tool do badly?)

**Opportunities:** The advantages of the tool to the organisation (i.e. how will the listed strengths benefit you and/or the organisation)

**Threats:** The disadvantages of the tool to the organisation (i.e. how will the listed strengths or weaknesses threaten you and/or the organisation)

|               |            |
|---------------|------------|
| Strengths     | Weaknesses |
| Opportunities | Threats    |

## 13 Appendix Four – OntoFarm Questionnaires

## 13.1 Student Demonstration and Questionnaire

## Overall

1. Overall I would rate the ease of use of the Onto Farm tool as:
  - a. Very Easy, Easy, Neutral, Difficult, Very Difficult
2. Creating a concept is easy:
  - a. Strongly Agree, Agree, Neutral, Disagree, Strongly Disagree
3. Finding an existing concept is easy:
  - a. Strongly Agree, Agree, Neutral, Disagree, Strongly Disagree
4. The layout of information within the system is easy to understand:
  - a. Strongly Agree, Agree, Neutral, Disagree, Strongly Disagree
5. The ability to export to OWL is:
  - a. Very Easy, Easy, Neutral, Difficult, Very Difficult
6. Differing lexical representations for each system can be entered adequately
  - a. Strongly Agree, Agree, Neutral, Disagree, Strongly Disagree

## Strengths, Weaknesses, Opportunities and Threats

Please analyse the tool and place your comments into the relevant boxes:

**Strengths:** The advantages of the tool (i.e. what does the tool do well?)

**Weaknesses:** The disadvantages of the tool (i.e. what does the tool do badly?)

**Opportunities:** The advantages of the tool to the organisation (i.e. how will the listed strengths benefit you and/or the organisation)

**Threats:** The disadvantages of the tool to the organisation (i.e. how will the listed strengths or weaknesses threaten you and/or the organisation)

|               |            |
|---------------|------------|
| Strengths     | Weaknesses |
| Opportunities | Threats    |

## Benefits

1. The system allows users to add concepts to the system in a faster manor than before:
  - a. Strongly Agree, Agree, Neutral, Disagree, Strongly Disagree
2. The system helps to highlight concepts that may already exist in the system:
  - a. Strongly Agree, Agree, Neutral, Disagree, Strongly Disagree
3. The system helps to prevent duplication of concepts that are already entered into the system:
  - a. Strongly Agree, Agree, Neutral, Disagree, Strongly Disagree
4. The system gives a clear separation of namespaces:
  - a. Strongly Agree, Agree, Neutral, Disagree, Strongly Disagree



5. The system provides an improved search functionality over traditional systems:
  - a. Strongly Agree, Agree, Neutral, Disagree, Strongly Disagree
6. The system allows more than one person to work with the ontology at **different** times:
  - a. Strongly Agree, Agree, Neutral, Disagree, Strongly Disagree
7. The system allows more than one person to work with the ontology at the **same** time:
  - a. Strongly Agree, Agree, Neutral, Disagree, Strongly Disagree
8. The system allows sufficient restriction of the relationships that may be entered into the system:
  - a. Strongly Agree, Agree, Neutral, Disagree, Strongly Disagree

## Harvesting Information

- The harvesting tool provides a good list of concepts to add to the system:
  - Strongly Agree, Agree, Neutral, Disagree, Strongly Disagree
- The harvesting view allows you to find adequate information regarding a concept that you may be trying to add to the system:
  - Strongly Agree, Agree, Neutral, Disagree, Strongly Disagree
- The harvesting tool's concept cloud gives a good overview of the page you are currently viewing:
  - Strongly Agree, Agree, Neutral, Disagree, Strongly Disagree
- The list of concepts collected for a page is too long:
  - Strongly Agree, Agree, Neutral, Disagree, Strongly Disagree
- The list of concepts collected for a page is not long enough:
  - Strongly Agree, Agree, Neutral, Disagree, Strongly Disagree
- The tool makes it easy to add concepts from the list provided
  - Strongly Agree, Agree, Neutral, Disagree, Strongly Disagree

**13.2 Focus Group within SoftwareCo**

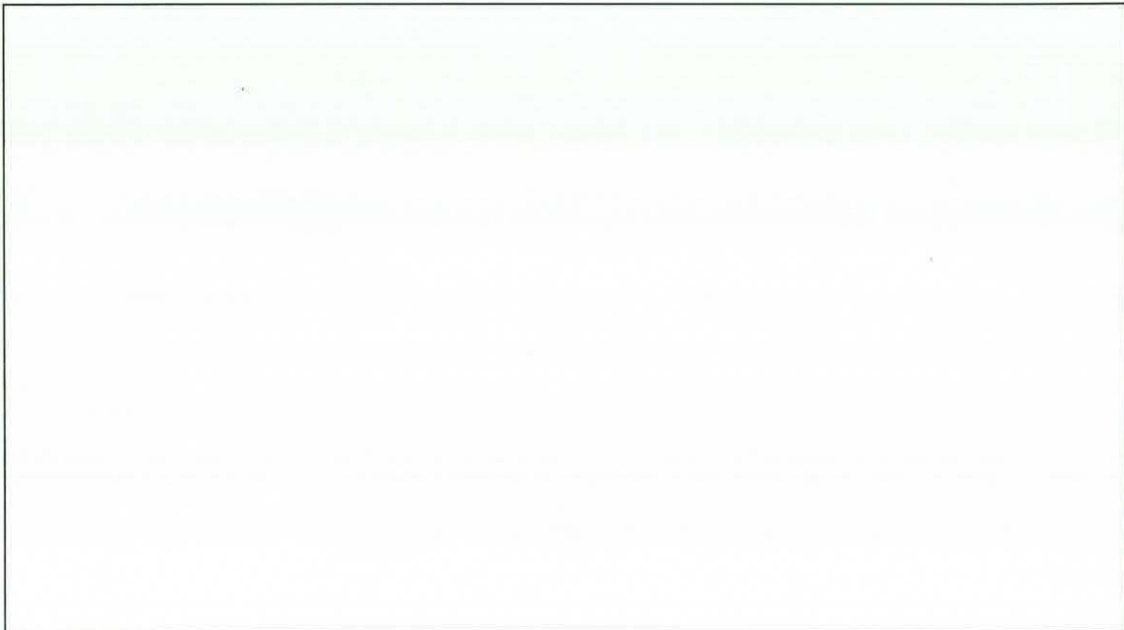
## Pre Questionnaire

Unless the question states otherwise, please circle one answer per question.

1. How would you rank your awareness of ontologies
  - a. Strong Awareness, Fair Awareness, Low Awareness, No Awareness
2. Do you think ontologies can help you retrieve information more effectively?
  - a. Definitely, Maybe, Sometimes, Rarely, Never
3. What are the main barriers you perceive to creating an ontology? Please circle one more of the following:  
Time, Expertise, Lack of Tools, Lack of Knowledge, Its not my job to create, other \_\_\_\_\_
4. Do you consider only one person creating and maintaining an ontology a problem?
  - a. Definitely, Maybe, Sometimes, Rarely, Never
5. Do you consider the time that it takes to create an ontology worth the benefit?
  - a. Definitely, Maybe, Sometimes, Rarely, Never
6. Do you consider the time it takes to develop an ontology a problem when the ontology is first created?
  - a. Definitely, Maybe, Sometimes, Rarely, Never
7. Do you consider the time it takes to discover concepts to add to an ontology a problem when the ontology is first created?
  - a. Definitely, Maybe, Sometimes, Rarely, Never
8. Do you think users need training to create an ontology?
  - a. Definitely, Maybe, Sometimes, Rarely, Never

## Issues with ontology development

Please note difficulties you see with traditional ontology development:

A large, empty rectangular box with a thin black border, intended for the user to write their response to the question about difficulties in traditional ontology development.

## Post Questionnaire

### Discovering Information

1. The system allows users to add concepts (classes) to the system in a faster manor than before (e.g. Protégé):
  - a. Strongly Agree, Agree, Neutral, Disagree, Strongly Disagree
2. The harvesting tool provides a good list of harvested, potential, concepts to add to the system:
  - a. Strongly Agree, Agree, Neutral, Disagree, Strongly Disagree
3. The harvesting view allows you to find adequate information regarding a concept that you may be trying to add to the system:
  - a. Strongly Agree, Agree, Neutral, Disagree, Strongly Disagree
4. The harvesting tool's concept cloud gives a good overview of the page you are currently viewing:
  - a. Strongly Agree, Agree, Neutral, Disagree, Strongly Disagree
5. The tool makes it easy to add concepts from the list provided to the ontology
  - a. Strongly Agree, Agree, Neutral, Disagree, Strongly Disagree

### Quality of Information

1. The system highlights concepts that may already exist in the system:
  - a. Strongly Agree, Agree, Neutral, Disagree, Strongly Disagree
2. The system prevents duplication of concepts that are already entered into the system:
  - a. Strongly Agree, Agree, Neutral, Disagree, Strongly Disagree
3. The system gives a clear separation of namespaces and allows modularity of the ontology (for example separating products from technologies boprod:SAP\_ERP and botech:Ontologies):
  - a. Strongly Agree, Agree, Neutral, Disagree, Strongly Disagree
4. OntoFarm allows different lexical representations of the same word to be entered. This will allow the user more freedom when entering information into a system and prevent them being restricted to certain terms. Do you agree with this functionality?
  - a. Strongly Agree, Agree, Neutral, Disagree, Strongly Disagree

5. The search functionality makes it easy to discover concepts in order to edit them
  - a. Strongly Agree, Agree, Neutral, Disagree, Strongly Disagree
6. Do you prefer the tree structure of Protégé or the search everything based approach of OntoFarm

### Collaboration

1. It is important that the system allows more than one person to work with the ontology at **different** times:
  - a. Strongly Agree, Agree, Neutral, Disagree, Strongly Disagree
2. It is important that the system enables more than one person to work with the ontology at the **same** time:
  - a. Strongly Agree, Agree, Neutral, Disagree, Strongly Disagree
3. Restriction of the predicates that may be entered into the system helps make the system easier to understand and use:
  - a. Strongly Agree, Agree, Neutral, Disagree, Strongly Disagree
4. Simplification of concept entry (removal of design decisions such as classes, instances and the restriction on predicates that can be entered) allows users who are not so familiar with the ontology to enter concepts that are important to them.
  - a. Strongly Agree, Agree, Neutral, Disagree, Strongly Disagree
5. The System allows users who are not experts working with ontologies to enter items into the ontology.
  - a. Strongly Agree, Agree, Neutral, Disagree, Strongly Disagree
6. The description given to a concept helps prevent duplication of concepts and misunderstanding
  - a. Strongly Agree, Agree, Neutral, Disagree, Strongly Disagree

### Restrictions

1. Restricting the predicates that people may use when creating a triple in the ontology makes the process of creating an ontology simpler?
  - a. Strongly Agree, Agree, Neutral, Disagree, Strongly Disagree
2. Restricting the predicates that people may use when creating a triple in the ontology removes value from the ontology?
  - a. Strongly Agree, Agree, Neutral, Disagree, Strongly Disagree

## Overall

1. The system takes a long time to understand
  - a. Strongly Agree, Agree, Neutral, Disagree, Strongly Disagree
2. Which do you feel would take the least time to create an initial ontology?
  - a. OntoFarm, a Text Editor, Protégé
3. Which do you feel would take longer to add to and maintain an ontology?
  - a. OntoFarm, a Text Editor, Protégé
4. Which do you feel would provide a better structured ontology?
  - a. OntoFarm, a Text Editor, Protégé

## Repeated Questions

1. How would you rank your awareness of ontologies
  - a. Strong Awareness, Fair Awareness, Low Awareness, No Awareness
2. Do you think ontologies can help you retrieve information more effectively?
  - a. Definitely, Maybe, Sometimes, Rarely, Never
3. What are the main barriers you perceive to creating an ontology? Please circle one more of the following:  
Time, Expertise, Lack of Tools, Lack of Knowledge, Its not my job to create, other \_\_\_\_\_
4. Do you consider only one person creating and maintaining an ontology a problem?
  - a. Definitely, Maybe, Sometimes, Rarely, Never
5. Do you consider the time that it takes to create an ontology worth the benefit?
  - a. Definitely, Maybe, Sometimes, Rarely, Never
6. Do you consider the time it takes to develop an ontology a problem when the ontology is first created?
  - a. Definitely, Maybe, Sometimes, Rarely, Never
7. Do you consider the time it takes to discover concepts to add to an ontology a problem when the ontology is first created?
  - a. Definitely, Maybe, Sometimes, Rarely, Never

8. Do you think users need training to create an ontology?

a. Definitely, Maybe, Sometimes, Rarely, Never



### Strengths, Weaknesses, Opportunities and Threats of the Tool

Please analyse the tool and place your comments into the relevant boxes:

Strengths: The advantages of the tool (i.e. what does the tool do well?)

Weaknesses: The disadvantages of the tool (i.e. what does the tool do badly?)

Opportunities: The advantages of the tool to the organisation (i.e. how will the listed strengths benefit you and/or the organisation)

Threats: The disadvantages of the tool to the organisation (i.e. how will the listed strengths or weaknesses threaten you and/or the organisation)

|               |            |
|---------------|------------|
| Strengths     | Weaknesses |
| Opportunities | Threats    |

