

# **Knowledge Visualisation of Wikipedia**

**Jinhui Wang**

A dissertation submitted in partial fulfilment of the requirements of  
Dublin Institute of Technology for the degree of  
M.Sc. in Computing (Knowledge Management)

**September 2008**

I certify that this dissertation which I now submit for examination for the award of MSc in Computing (Knowledge Management), is entirely my own work and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

This dissertation was prepared according to the regulations for postgraduate study of the Dublin Institute of Technology and has not been submitted in whole or part for an award in any other Institute or University.

The work reported on in this dissertation conforms to the principles and requirements of the Institute's guidelines for ethics in research.

**Signed:** \_\_\_\_\_

**Date:**                    *08 09 2008*

# 1 ABSTRACT

Wikipedia is a popular online encyclopaedia, in which articles are created together by anyone who is willing to contribute. There is debate on the quality, reliability and consistency of articles due to its openness. Besides, it is hard to see the growing path of articles because there is no centralised control in Wikipedia. The history of articles is available but in form that is difficult to process to get a picture of article evolution. Knowledge visualisation addresses the problem by providing analysis and comprehension of large amounts of data to gain insights and to facilitate knowledge creation and sharing. However, currently there is no research focusing on visualising the content change for Wikipedia articles. This project is to investigate the usefulness of a visual representation of article history as a tool for Wikipedia users and the usefulness of this representation as a tool to improve knowledge creation and sharing in Wikipedia. By utilising several data retrieving, parsing and visualising tools, the project builds the visual representations of Knowledge Management article on Wikipedia based on the word count and contribution metrics. It is found that the dynamic visualisation is useful in tracking the evolution of article and is helpful for gaining better understanding of the topic after seeing the evolution of its article on Wikipedia.

**Key words:** *knowledge visualisation, knowledge management, Wikipedia*

## **ACKNOWLEDGEMENTS**

I would like to express my sincere thanks to my supervisor Deirdre Lawless for her valuable advice and insights throughout the duration of this project.

I would also like to express my gratitude to Damian Gordon for all of his help to shot the video for this project.

Finally I would like to thank my family and friends for their support during the year.

# TABLE OF CONTENTS

<b>1</b>	<b>ABSTRACT.....</b>	<b>II</b>
	<b>TABLE OF FIGURES .....</b>	<b>VII</b>
	<b>TABLE OF TABLES .....</b>	<b>IX</b>
<b>1.</b>	<b>INTRODUCTION .....</b>	<b>1</b>
1.1	INTRODUCTION TO PROJECT .....	1
1.2	BACKGROUND .....	3
1.3	RESEARCH PROBLEM .....	6
1.4	INTELLECTUAL CHALLENGE .....	6
1.5	RESEARCH OBJECTIVES .....	7
1.6	RESEARCH METHODOLOGY .....	8
1.7	RESOURCES .....	9
1.8	SCOPE AND LIMITATIONS .....	10
1.9	ORGANISATION OF THE DISSERTATION .....	11
<b>2</b>	<b>THE WIKIPEDIA .....</b>	<b>12</b>
2.1	INTRODUCTION.....	12
2.2	WIKIPEDIA OVERVIEW .....	12
2.2.1	<i>Featured Articles</i> .....	15
2.2.2	<i>Five Pillars</i> .....	17
2.3	COOPERATION IN WIKIPEDIA.....	19
2.4	QUALITY OF WIKIPEDIA .....	20
2.5	WIKIPEDIA APPLICATION .....	24
2.6	CONCLUSION .....	26
<b>3</b>	<b>WIKIPEDIA AS A KNOWLEDGE MANAGEMENT RESOURCE .....</b>	<b>27</b>
3.1	INTRODUCTION.....	27
3.2	KNOWLEDGE MANAGEMENT.....	27
3.2.1	<i>Spiral of Knowledge</i> .....	29
3.2.2	<i>Knowledge Management Process</i> .....	30
3.3	WIKIPEDIA AND KNOWLEDGE MANAGEMENT .....	32

3.3.1	<i>Spiral of Knowledge in Wikipedia</i> .....	33
3.3.2	<i>Wikipedia as Knowledge Base</i> .....	33
3.3.3	<i>Wikipedia as Communities of Practice (CoP)</i> .....	34
3.4	CONCLUSION .....	35
<b>4</b>	<b>KNOWLEDGE VISUALISATION .....</b>	<b>36</b>
4.1	INTRODUCTION.....	36
4.2	VISUALISATION .....	36
4.3	STATIC VISUALISATION.....	38
4.4	DYNAMIC VISUALISATION.....	38
4.5	KNOWLEDGE VISUALISATION AND KNOWLEDGE MANAGEMENT.....	38
4.6	VISUALISATION OF WIKIPEDIA .....	38
4.6.1	<i>Trends in Revision History</i> .....	38
4.6.2	<i>Visualisation of Wikipedia Collaboration</i> .....	38
4.6.3	<i>Mosaic Visualisation of Wikipedia</i> .....	38
4.6.4	<i>Visual Side of Wikipedia</i> .....	38
4.7	CONCLUSION .....	38
<b>5</b>	<b>STATIC VISUALISATION OF CONTENT CHANGE IN WIKIPEDIA....</b>	<b>38</b>
5.1	INTRODUCTION.....	38
5.2	REQUIREMENTS FOR VISUALISATION .....	38
5.3	THE TEST BED.....	38
5.4	VISUALISATION PROCESS AND SUPPORTING TECHNICAL ARCHITECTURE.....	38
5.4.1	<i>Retrieving Data from Wikipedia</i> .....	38
5.4.1.1	<i>The Tool</i> .....	38
5.4.1.2	<i>Crawl Data</i> .....	38
5.4.2	<i>Parsing the Data For Analysis</i> .....	38
5.4.2.1	<i>XML Parsing</i> .....	38
5.4.2.2	<i>Wiki Markup</i> .....	38
5.4.2.3	<i>Parsing Wikipedia Notation for Static Visualisation</i> .....	38
5.4.3	<i>Static Visualisation with GraphViz</i> .....	38
5.4.4	<i>The Result</i> .....	38
5.5	CONCLUSION .....	38
<b>6</b>	<b>DYNAMIC VISUALISATION OF CONTENT CHANGE IN WIKIPEDIA</b>	<b>38</b>

6.1	INTRODUCTION.....	38
6.2	REQUIREMENTS FOR DYNAMIC VISUALISATION.....	38
6.3	VISUALISATION PROCESS AND SUPPORTING TECHNICAL ARCHITECTURE.....	38
	6.3.1.1 <i>Parsing Wikipedia Notation for Dynamic Visualisation</i> .....	38
	6.3.1.2 <i>Loading Data Into Database</i> .....	38
	6.3.1.3 <i>Creating Data Summary To Support Required Analysis</i> .....	38
6.4	DYNAMIC VISUALISATION.....	38
	6.4.1 <i>Introduction to Google Motion Visualisation API</i> .....	38
	6.4.2 <i>Wikipedia Visualisation Using Google Motion API</i> .....	38
6.5	THE RESULT.....	38
6.6	SURVEY AND EVALUATION.....	38
6.7	CONCLUSION.....	38
<b>7</b>	<b>CONCLUSIONS AND EVALUATION.....</b>	<b>38</b>
7.1	INTRODUCTION.....	38
7.2	RESEARCH DEFINITION & RESEARCH OVERVIEW.....	38
7.3	CONTRIBUTIONS TO THE BODY OF KNOWLEDGE.....	38
7.4	EXPERIMENTATION, EVALUATION AND LIMITATION.....	38
7.5	FUTURE WORK & RESEARCH.....	38
7.6	CONCLUSION.....	38
	<b>BIBLIOGRAPHY.....</b>	<b>38</b>
	<b>APPENDIX A.....</b>	<b>38</b>

## TABLE OF FIGURES

FIGURE 1 WIKIPEDIA PAGE FOR KNOWLEDGE MANAGEMENT ON AUGUST 19, 2008.....	13
FIGURE 2 EDIT PAGE FOR KNOWLEDGE MANAGEMENT.....	14
FIGURE 3 REVISION HISTORY OF KNOWLEDGE MANAGEMENT UP TO AUGUST 19, 2008	14
FIGURE 4 TALK PAGE FOR KNOWLEDGE MANAGEMENT UNTIL AUGUST 19, 2008 .....	15
FIGURE 5 FEATURED ARTICLE ANTARCTICA ON AUGUST 19, 2008 .....	16
FIGURE 6 SCREEN SHOOT OF WIKED IN ACTION .....	19
FIGURE 7 FOUR PILLARS OF KNOWLEDGE MANAGEMENT .....	28
FIGURE 8 CONVERSION OF KNOWLEDGE BETWEEN TACIT AND EXPLICIT FORMS .....	30
FIGURE 9 A SKETCH OF THE USABILITY LAB .....	37
FIGURE 10 HOW THE INTERNET INFLUENCES INDUSTRY STRUCTURE.....	38
FIGURE 11 THE NEGOTIATION BRIDGE: A VISUAL METAPHOR THAT OUTLINES A NEGOTIATION METHOD.....	38
FIGURE 12 AN INTERACTIVE VISUALIZATION HELPS TO SUPERVISE THE NEW YORK STOCK EXCHANGE .....	38
FIGURE 13 FINDING OF ACCOMMODATION VIA A GOOGLE MAP.....	38
FIGURE 14 A VISUAL LITERATURE REVIEW DIAGRAM ON INFORMATION OVERLOAD... 38	
FIGURE 15 VISUALISATION OF INTERNET WITH OVER 5 MILLION EDGES AND ESTIMATED 50 MILLION HOP COUNT.....	38
FIGURE 16 VISUALISATION OF GOLD MEDALS FOR 2008 BEIJING OLYMPICS UP TO AUGUST 23, 2008.....	38
FIGURE 17 WORD CLOUD FOR KNOWLEDGE MANAGEMENT ARTICLE (AUGUST 23, 2008) ON WIKIPEDIA.....	38
FIGURE 18 WORD MAP FOR HAPPY AND PLEASED GENERATED BY VISUAL THEASURUS .....	38
FIGURE 19 VISUALISED SEARCHING FOR KNOWLEDGE MANAGEMENT BY KARTOO ....	38
FIGURE 20 GAPMINDER WORLD - FERTILITY VERSUS LIFE EXPECTANCY IN YEAR 1950 AND 2010.....	38
FIGURE 21 HISTORY FLOW VISUALISATION OF THE WIKIPEDIA ENTRY ON 'EVOLUTION', 2006.....	38



FIGURE 22 REVERT GRAPH USES FORCE DIRECTED LAYOUT TO SIMULATE SOCIAL STRUCTURES BETWEEN USERS.....	38
FIGURE 23 MOSAIC VISUALISATION FOR WIKIPEDIA.....	38
FIGURE 24 PROCESS OF STATIC VISUALISATION.....	38
FIGURE 25 SAMPLE DIRECTED GRAPH GENERATED BY GRAPHVIZ.....	38
FIGURE 26 STATIC VISUALISATION OF “KNOWLEDGE MANAGEMENT” ON WIKIPEDIA JULY 17, 2008.....	38
FIGURE 27 TABLE OF CONTENTS FOR KNOWLEDGE MANAGEMENT ON WIKIPEDIA JULY 17, 2008.....	38
FIGURE 28 PROCESS OF DYNAMIC VISUALISATION FOR WIKIPEDIA.....	38
FIGURE 29 AMERICAN ECONOMIC FROM YEAR 2000 TO 2006.....	38
FIGURE 30 VISUALISE DATA WITH GOOGLE SPREADSHEET.....	38
FIGURE 31 MOTION WIDGET GENERATED WITH JAVASCRIPT.....	38
FIGURE 32 DYNAMIC VISUALISATION FOR KNOWLEDGE MANAGEMENT ARTICLE IN WIKIPEDIA USING WORD AND CONTRIBUTION COUNT METRICS. (REVISIONS FROM AUGUST 2002 TO AUGUST 2008).....	38
FIGURE 33 INVESTIGATION ON DEFINITION AND EXTERNAL LINKS WITH THE DYNAMIC VISUALISATION TOOL.....	38
FIGURE 34 COMPARISON OF KNOWLEDGE MANAGEMENT DEFINITION BETWEEN 07:53, 3 NOVEMBER 2005 AND 14:45, 3 NOVEMBER 2005 REVISIONS.....	38
FIGURE 35 MAIN PAGE FOR DEMONSTRATING VISUALISATION TOOL.....	38
FIGURE 36 VIDEOS FOR EXPLAINING THE VISUALISATION TOOLS.....	38
FIGURE 37 PERCENTAGE OF VIEWING ARTICLES IN WIKIPEDIA.....	38
FIGURE 38 PERCENTAGE OF RELIABILITY OF ARTICLE CONTENT ON WIKIPEDIA.....	38
FIGURE 39 PERCENTAGE OF QUALITY OF ARTICLES ON WIKIPEDIA.....	38
FIGURE 40 COMPARISON OF USEFULNESS OF HISTORY BETWEEN READER AND CONTRIBUTOR.....	38
FIGURE 41 RESPONSE TO QUALITY OF KNOWLEDGE MANAGEMENT ARTICLE ON WIKIPEDIA.....	38
FIGURE 42 RESPONSES TO USEFULNESS OF VISUALISATION TOOL FOR TRACKING EVOLUTION OF ARTICLE FOR KNOWLEDGE MANAGEMENT.....	38
FIGURE 43 RESPONSE TO THE EXTENSIBILITY OF THE VISUALISATION TOOL.....	38

## TABLE OF TABLES

TABLE 1 SUMMARY FOR WIKIPEDIA.ORG DATABASE DUMP ON 20080312 .....	38
TABLE 2 HTTP PARAMETERS FOR EXPORTING ARTICLES FROM WIKIPEDIA.ORG .....	38
TABLE 3 EXAMPLE OF HTML FORM USING POST REQUEST METHOD.....	38
TABLE 4 EXAMPLE OF RETURNED XML DUMP FOR KNOWLEDGE MANAGEMENT .....	38
TABLE 5 CAPABILITY AND LIMITATION OF SAX PROCESSOR .....	38
TABLE 6 XSD OUTPUT BY MEDIAWIKI'S SPECIAL:EXPORT SYSTEM.....	38
TABLE 7 TYPICAL WIKI MARKUP .....	38
TABLE 8 DOT DESCRIPTION FOR GENERATING A SIMPLE DIRECTED GRAPHS .....	38
TABLE 9 EXAMPLE OF HIERARCHY STRUCTURE OF WIKIPEDIA ARTICLE (SOURCE: “KNOWLEDGE MANAGEMENT” FROM WIKIPEDIA.ORG ON AUGUST 6, 2008).....	38
TABLE 10 DATABASE SCHEMA <i>BASIC</i> FOR IMPORTING XML DUMP .....	38
TABLE 11 DATABASE SCHEMA <i>SECTION</i> FOR SUMMARISING SECTION DETAILS .....	38
TABLE 12 EMBEDDED JAVASCRIPT CODE FOR GENERATING MOTION VISUALISATION WIDGET .....	38
TABLE 13 COLUMNS FOR WIKIPEDIA DYNAMIC VISUALISATION.....	38



# 1. INTRODUCTION

## *1.1 Introduction to Project*

Wikipedia is a free, open content encyclopaedia project operated by the non profit Wikimedia Foundation (Wikimedia Foundation Financial Statements, 2007), in which articles are collaboratively written and maintained by anyone who can access to the Internet. As of April 26th, 2007, a total of 1,755,932 articles are already available in the English language edition. Wikipedia is a world wide cooperation platform containing articles in more than 200 languages (Ortega and Barahona 2007). In July 2008, when work on this dissertation started, it had over 2,450,000 articles in English. Wikipedia has now become the top ten most-visited web sites worldwide (Alexa 2008).

Articles in Wikipedia cover many areas including arts, biography, geography, history, mathematics, science, society and technology. Wikipedia can be seen as a repository of knowledge. It is accessible anywhere in the world with Internet connection. Knowledge can be captured by anyone who is willing to contribute as long as he or she follows the policies and guidelines. Over 1,500 articles have been designated by the Wikipedia community as featured articles (Blumenstock 2008).

Wikipedia has portals, which organise content around topic areas in a loose hierarchy structure. In addition, it provides searching facilities as well as simple hyperlinks for locating articles. Wikipedia can be used in many different ways. Visitors can acquire knowledge from Wikipedia by simply exploring articles. They can also browse articles around topics in a hierarchy structure, read random articles and search articles.

In May 2006, The University of Washington Libraries Digital Initiatives unit began a project to integrate the UW Libraries Digital Collections into the information workflow of their students by inserting links into Wikipedia (Ann and Carolyn 2007). As a result, analysis of server statistics indicates that Wikipedia is indeed driving more traffic to their library web site. Yu et al. (2007) seek to evaluate ontologies based on categories found in Wikipedia. It is found that tangledness may be desirable in

ontologies and category structures for browsing in general knowledge application areas like Wikipedia. Ponzetto and Strube (2007) present experiments on using Wikipedia for computing semantic relatedness. It is found that existing relatedness measures perform better using Wikipedia than a baseline given by Google counts.

As Wikipedia is open to a large contributor base and anyone can edit it, it is important to have mechanisms and tools such as policies and guidelines to assist the development of articles. Constant revisions leave the Wikipedia in an incomplete state. It is common that articles grow with multiple authors and revision cycles in Wikipedia. Wikipedia is supported by MediaWiki.org (2003), which has robust version and reversion controls to prevent poor quality edits from doing permanent harm to articles. However, the moderation in Wikipedia is poor. As anyone can contribute, factual errors or overtly vandalise articles can be introduced. If there is conflict among contributors, resolution may only result after months-long disagreements.

It is interesting to track history of Wikipedia articles. Articles may contain false or debatable information at start and gradually has a consensus view after a long process of discussion, debate and argument. Other articles may be caught in a heavily unbalanced viewpoint and take months or years to achieve better balance point. Each article has its own growing path. In addition, articles evolve in different ways: contributors may focus on one aspect of the concept and move on to another or they may simultaneously work on all the aspects of the concept. By tracking the history of Wikipedia articles, it is notable to observe the unique growing path of each article and gain insights how knowledge is created.

The project described in this dissertation explored the usefulness of visualisation as a technique for tracking content change in Wikipedia and for the purposes of knowledge exploration and creation within topics in Wikipedia. Knowledge visualisation, in particular the dynamic visualisation, allows the manipulation of different types of information in a way that improves knowledge creation and transfer. The dynamic visualisation developed was useful to both Wikipedia users and experts in different domains. The visual representation allowed them to evaluate user perspective on key concepts in their domains of interest. By combing, aggregating and summarising the

Wikipedia history, the animation visualised the content change in articles from different perspectives, showing the process of change that goes beyond the stored data.

The concept of Knowledge Management was used as a test bed in this project. Knowledge Management is the systemic and organisationally specified process for acquiring, organising, and communicating knowledge of employees so that other employees may make use of it to be more effective and productive in their work (Alavi & Leidner 1999). However, Knowledge Management is a concept which crosses many disciplines and therefore there are differences in the views of the various user communities. This should be reflected in the Knowledge Management Wikipedia article which makes it very suitable to support experimentation in this project.

## ***1.2 Background***

The World-Wide Web has achieved a large scale of cooperation for knowledge creation and sharing all over the world. Berners-Lee (1996) writes that the Web's major goal was to be a shared information space through which people and machines could communicate. The World Wide Web has now become for many the primary resource for creating and sharing knowledge. O'Reilly (2005) states that Web 2.0 is the network as platform, spanning all connected devices; Web 2.0 applications are those that make the most of the intrinsic advantages of that platform: delivering software as a continually-updated service that gets better the more people use it, consuming and remixing data from multiple sources, including individual users, while providing their own data and services in a form that allows remixing by others, creating network effects through an "architecture of participation," and going beyond the page metaphor of Web 1.0 to deliver rich user experiences. With the popularity of Web 2.0, people not only acquire information published by organisations but also start to gain knowledge from individual contributions through Wikipedia, blogs and social networks. There are a number of popular Web 2.0 applications:

- Gmail. Gmail is Google's Web-based e-mail service introduced in 2004. Gmail interface resembles desktop applications, generating the so-called 'Rich User Experience'. Users are meant to find mails by either conversations replies to and forwards of messages, starring adding a 'star' to a message or searching

full text on all messages, thus relying on Google's roots as being a search engine originally (Best 2006).

- Flickr. Flickr is a photo publishing Web site. Flickr is the first photo service that introduced sorting photos by tags describing what is being depicted. Moving images into and within sets can be done by drag-and-drop. Processes like uploading or renaming of images are supported by JavaScript actions showing progress and displaying changes immediately (Best 2006).
- Facebook. Facebook has become hugely one of the popular social networking applications in the last few years. It provides users with a profile space, facilities for uploading content (e.g. photos, music), messaging in various forms and the ability to make connections to other people (Joinson 2008).

With strong power of modern desktops as well as mobile devices, people are able to retrieve knowledge in multimedia forms such as pictures, audios and videos from the web with no space time limitation.

Wikipedia attempts to provide a reliable knowledge creation and sharing platform. It is a widespread project in that it has more than 200 languages and over 1,750,000 articles in English language and top 10 language editions (English, German, French, Japanese, Polish, Dutch, Italian, Portuguese, Spanish and Swedish) accumulate a total sum over 4,800,000 articles (Ortega and Barahona). Every minute, there are new contributions made to various articles. Wikipedia has five pillars that define the character of the project. Built upon the five pillars, policies and guidelines ensure the quality of articles (Butler et al. 2008).

Wikipedia has robust version and reversion controls to prevent poor quality edits from doing permanent harm to articles. Contribution is not limited to editing the pages. For example, one can cleanup the article by changing spelling, grammar, tone, and sourcing. Verifiability ensures readers are able to check that material added to Wikipedia has already been published by a reliable source. Talk pages offer the ability to discuss articles and other issues with other contributors when conflict arises.

While Wikipedia owes the incredible growth to open-source editing, it also suffers from its openness. Dedicated and knowledgeable editors can and do effectively reverse the process of entropy by making entries better over time. Other editors, through ignorance, sloppy research, or, on occasion, malice or zeal, can and do introduce or perpetuate errors in fact or interpretation. The reader never knows whether the last editor was one of this latter group; most editors leave no trace save a whimsical cyber-handle. It is frowned upon as an academic reference in that not only Wikipedia may appear in the future to escape the consequences of errors but also it states in its guidelines that its contents are not suitable for academic citation, because Wikipedia is, like a print encyclopaedia, a tertiary source (Waters 2007).

Knowledge Visualisation, which designates all graphic means that can be used to construct and convey complex insights, plays an important role in Knowledge Management (Eppler & Burkhard 2005). There are a number of knowledge visualisation formats: heuristic sketches, conceptual diagrams, visual metaphors, knowledge animations, knowledge maps and scientific charts (Eppler & Burkhard 2005). The conceptual diagrams are for structuring information and illustrating relationships while knowledge animations are for dynamic and interactive visualisations (Su et al. 2007). With the dramatic decreasing storage cost, information overload has become a major problem for organisations, especially for those who are knowledge-intensive and even the entire society. Organisations are drowning in data but starving for knowledge. Knowledge visualisation helps to compress large amounts of information with the help of analytical frameworks, theories, and models that absorb complexity and render it accessible (Eppler & Burkhard 2005). Knowledge visualisation also helps organisations find insights and gain actionable knowledge in a quick and direct manner. The use of knowledge visualisation can therefore be a mechanism which can improve the creation and transfer of knowledge between two or more people.

A static visualisation is a snapshot or an image while a dynamic visualisation is an animation, both serving as intermedium for knowledge creation, transferring and cognition. A static visualisation allows exploring data by offering different methods such as overview, zooming in and filtering and then showing details on demand to achieve the cognition. On the other hand, dynamic visualisation helps to explore large



time-varying datasets with reoccurring data objects that alter in time. Static visualisation fits casual users as it shows a simple image for the underlying data while dynamic visualisation is novel for advanced users, providing more interactions ability for viewing complex datasets in multiple angles.

By visualising and delivering the content change of Wikipedia articles to the interested groups and domain experts, new knowledge and insights could be discovered. Currently, edit trail is the only resource that records the content change for Wikipedia articles. However, due to the active contribution, the history of articles dramatically grows, making it difficult to find insights in a plain text format. A visualisation of article history is helpful to show the change in a more concrete and understandable form.

### ***1.3 Research problem***

The primary aim of the project described in this dissertation was to investigate the usefulness of a visual representation of article history as a tool for Wikipedia users and to investigate the usefulness of this representation as a tool to improve knowledge creation and sharing in Wikipedia. The research involved in this project investigated how a knowledge visualisation could be created for tracking content change in Wikipedia articles and how the visualisation tool could be used by Wikipedia users and domain experts to improve knowledge creation and sharing. Both static and dynamic visualisations were considered in this research.

The focus of the research was to highlight the usefulness of knowledge visualisation as a knowledge management tool and moreover, how useful such a tool could be in a user controlled resource in Wikipedia.

### ***1.4 Intellectual challenge***

The intellectual challenges for this dissertation span many areas. These are:

- Explore the role of knowledge visualisation in knowledge management, the formats of knowledge visualisation and its current application.

- Investigate the working mechanism, usage and popularity of Wikipedia.
- Research types of data and changes related to Wikipedia revisions for the purpose of visualisation.
- Research existing file transferring tools to download history revisions of articles in Wikipedia.
- Research the existing visualisation tools to represent the article in both static and dynamic format for tracking the content change.
- Investigate the appropriateness and usefulness of knowledge visualisation for Wikipedia.
- Visualise content change of Knowledge Management article in Wikipedia in both static and dynamic format.
- Get communities from both Wikipedia users as well as domain experts to evaluate the output of the tool and to criticise the usefulness of a visualisation tool for Wikipedia.

### ***1.5 Research objectives***

The following objectives have been achieved throughout the dissertation and contributed to the overall outcome to highlight the appropriateness and usefulness of knowledge visualisation for Wikipedia:

To achieve this end, the project is divided up into six objectives. These are,

1. Perform a literature review on Wikipedia in particular the knowledge management issues related to the creation and content change management.

2. Perform a literature review on knowledge visualisation reviewing types of knowledge visualisation and typical usage of knowledge visualisation and the role of knowledge visualisation in knowledge management.
3. Identify an appropriate toolkit to create the visualisation, investigating file transferring utilities, parsers and visualisation tools for developing appropriate static and dynamic visualisation.
4. Identify the key aspects of content change for Wikipedia articles to develop the visualisation such as what data is of interest and what content should be visualised.
5. Create a static and dynamic series of visualisations for the Knowledge Management article in Wikipedia with the toolkit identified.
6. Evaluate the resulting knowledge visualisation of Wikipedia with a number of evaluation techniques.

### ***1.6 Research methodology***

Both primary and secondary research was performed throughout the duration of this project. The secondary research comprised of a literature review of material pertaining to three topics:

- The Wikipedia: its history, quality, syntax and its usage as a tool for collaboration in community of practice.
- Knowledge Management: The definition of knowledge management, spiral of knowledge and knowledge management process.
- Knowledge Visualisation: The definition, formats and application of knowledge visualisation. The role of knowledge visualisation in knowledge management.

The varying sources were used to complete the literature review topics: ACM Digital Library, IEEE Electronic Library, books and journals from DIT library. Other sources such as websites and dictionary were also used.

The primary research of this project involved determining what type of data and changes are of interest for tracking content change. A toolkit to facilitate the creation of the visualisation was researched and implemented. The Knowledge Management concept was used to as a test bed for the visualisation. In addition, secondary research on data of interest for visualisation will be of help to affect the shape of visualisation.

The results of the Wikipedia visualisation was published to interested groups and domain experts for examining whether new knowledge could be gained and created through visualisation and for gap analysis. Two short videos describing the project itself and illustrating the usage of tools were published on YouTube (YouTube 2005) for public evaluation.

Finally, a secondary survey regarding the quality and appropriateness of the visualisation was conducted in order to exam the usefulness of visualisation. Domain experts including lectures in Ireland colleges and MSc students in knowledge management course were invited to give their views and comments on the visualisation.

## ***1.7 Resources***

The following resources were essential components to the completion of this project:

- **File Transferring and Visualisation Tools**

File Transferring tools were vital to retrieve essential revisions from Wikipedia website for analysis. The visualisation tools were essential to present the content change of Wikipedia articles in a multimedia form.

- **Library Facilities**

Access to Dublin Institute of Technology Library facilities was one of the key resources for finishing this project. It was a great source for literature review.

- Computer

Access to a computer and word processing package was necessary to complete this dissertation and to store the relevant work done during the project.

- Internet and Email Access

Access to Internet was one of the main methods to keep contact with supervisor on VoIP. Access to Email allows discussing the dissertation with supervisor.

- Access to Supervisor

Access to supervisor on meetings was of great help to guide and give advice to this project.

- Publishing Website and Survey Software

Access to publishing website was critical to publish the visualisation result to interested audience.

- Survey Software

Access to survey software was essential to complete the organisational survey. The survey tool was available via <http://www.surveymonkey.com> for creating and distributing the survey for this project.

## ***1.8 Scope and limitations***

This project's sole focus was on part of the Wikipedia content. In order for the project to be achievable in the timescale the project focused on the Knowledge Management article on Wikipedia, in which experts are accessible to facilitate the required evaluation. Only word and contribution count metrics are selected for the dynamic visualisation. The limited number of metrics makes it difficult to give a complete view on the content change. The two metrics work on syntax level, which may not effectively reflect the semantic change. For example, while holding the same number of words, the meaning of the content could be totally different.

The article and its meta data are all from Wikipedia website. In this respect, all the experiment and conclusion are based on the data collected.

## ***1.9 Organisation of the dissertation***

This dissertation is divided into seven chapters. Chapter two will introduce the reader to Wikipedia. The chapter will examine cooperation, quality and application of Wikipedia and address the necessity of tracking evolution of Wikipedia articles to gain more understanding on articles.

Chapter three will introduce the definition of Knowledge Management, spiral of knowledge and knowledge management process. It will assess Wikipedia from knowledge management perspective. The chapter will illustrate that Wikipedia is a good sample as an online community of practice (CoP).

Chapter four will introduce concept of knowledge visualisation and its forms. It will explain static and dynamic visualisation and show the differences between them. The chapter will offer an assessment on how knowledge visualisation could help with knowledge management. The chapter will also present various existing examples of visualisation for Wikipedia.

Chapter five will outline the requirements for visualisation. It will discuss why the Knowledge Management article on Wikipedia is a suitable test bed for the visualisation. The chapter will describe the process of static visualisation. It will critically assess the result of static visualisation and illustrate why it is not particular useful for tracking content change.

Chapter six will assess the weakness of static visualisation and shows how the dynamic visualisation can help to achieve the goal of visualisation. It will describe the process and tools for dynamic visualisation. The chapter will present the result of dynamic visualisation for the Knowledge Management Wikipedia article as well as analyse the survey results.

Finally chapter seven contains results, conclusions and future areas of work identified as a result of the research conducted for this project.

## **2 THE WIKIPEDIA**

### ***2.1 Introduction***

Wikipedia puts control into the hands of users who decide what topics are covered and at what depth. As the project is to visualise the content change of Wikipedia articles, it is important to investigate how the community build the articles and how the quality of these articles can be assessed. In addition, while Wikipedia can be integrated as an invaluable knowledge base to build various applications, it has become a central resource for research and reference in a variety of areas.

This chapter presents the result of literature survey investigating Wikipedia. It starts by giving an overview of Wikipedia - what an article looks like, how article grows and the five pillars that define the characteristic of Wikipedia. It then describes in detail how contributors cooperate with each other and points out some of the problems with the way Wikipedia articles are created. The chapter also examines the quality of Wikipedia and the dangers of referring to it as an encyclopaedia even though it is not comprehensive and independently controlled. The chapter describes various applications that have been built on top of Wikipedia. The chapter concludes the necessity of tracking evolution of Wikipedia articles to gain more understanding on articles.

### ***2.2 Wikipedia Overview***

Wikipedia is a free, open content encyclopaedia project operated by the non profit Wikimedia Foundation (Wikimedia Foundation Financial Statements, 2007). Articles are collaboratively written and maintained by any volunteer who can access to the Internet. Wikipedia attempts to provide a reliable knowledge creation and sharing platform. As of April 26th, 2007, a total of 1,755,932 articles are already available in the English language edition and it is a world wide project with more than 200. Despite concerns about the quality of openly editable information, Wikipedia has now become the top ten most-visited web sites worldwide (Alexa 2008).

Figure 1 shows the article page for Knowledge Management on August 19, 2008. It starts with an introduction to Knowledge Management. It has a table of contents showing the article structure. It then shows different aspects of Knowledge Management section by section. By clicking the hyperlinks in the table of contents, one can easily navigate to the interested sections within the same page.



**Figure 1 Wikipedia Page for Knowledge Management on August 19, 2008**

As Wikipedia is an open-content knowledge sharing platform, anyone can edit the article by simply clicking the “edit this page” hyperlink at the top of the page shown in Figure 2. One can change the content of the page as well as format and organise the page by manipulating text with wiki syntax. Wikipedia supports a number of markups including links and URLs, images, headings, character formatting and tables. This makes it easy for both casual and professional users to create well formed, novel articles.



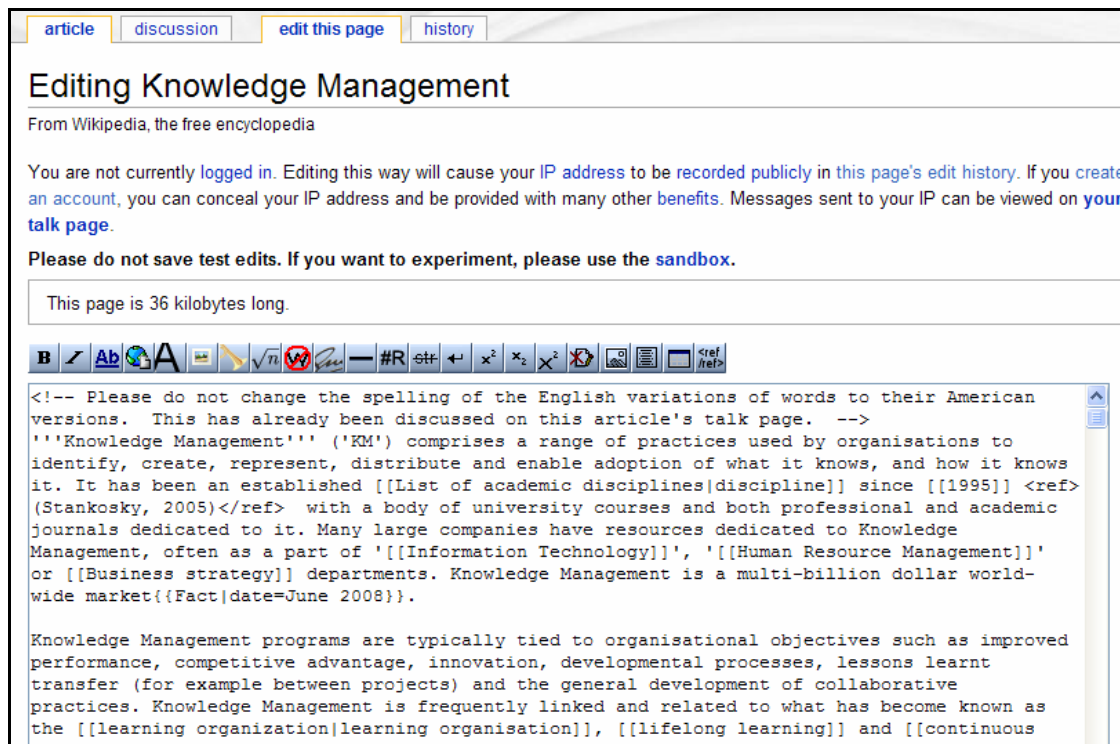


Figure 2 Edit Page for Knowledge Management

Figure 3 shows the revision history of Knowledge Management article dated to August 19, 2008. As a collaboration platform, Wikipedia maintains all the revisions of articles along with the date and time for each edit, the username or IP address of contributors as well as the edit summaries. Several actions can be taken on revisions such as viewing history versions, comparing different versions among pages or reverting to previous version.

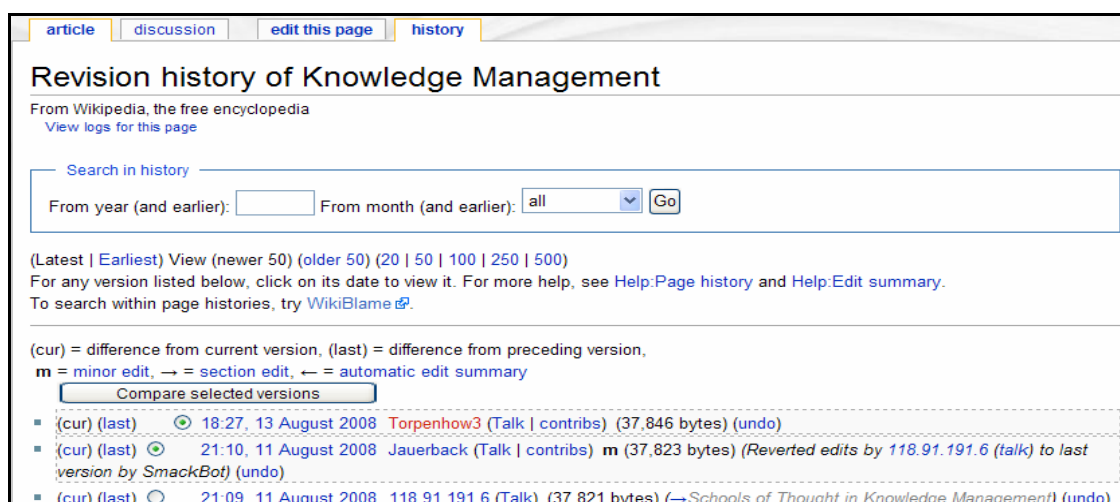


Figure 3 Revision history of Knowledge Management Up to August 19, 2008

When a conflict arises among contributors, Wikipedia provides an online communication mechanism called “Talk Page” for discussing the issues. Figure 4 shows the “Talk Page” for the Knowledge Management article dated to August 19, 2008. In the Talk Page, users discuss the content of articles as well as debate the appropriate content to be included. Talk pages are useful in that they may contain information that is not on the article and often unverified but useful.



**Figure 4 Talk Page for Knowledge Management until August 19, 2008**

### 2.2.1 Featured Articles

Featured articles are considered to be the best articles in Wikipedia, determined by Wikipedia’s editors. Figure 5 shows the article page for featured article Antarctica on August 19, 2008. As can be seen, an article is marked as featured with a star symbol at the right top of the page. The page on Antarctica contains an appropriate picture showing the location of Antarctica as well as a number of figures describing the overall status such as area and population. A featured article is considered outstanding due to its characteristics of well-written, comprehensive, accurate, neutral and stable. However, those characteristics are mainly qualitative judgement, which can not be

measured with any simple metric. For example, the comprehension of the article could largely vary from person to person. An article could be easily understood by domain experts while it is quite difficult for casual readers.

**Antarctica**

From Wikipedia, the free encyclopedia

*For other uses, see [Antarctica \(disambiguation\)](#).*

**Antarctica** is Earth's southernmost continent, overlying the South Pole. It is situated in the southern hemisphere, almost entirely south of the Antarctic Circle, and is surrounded by the Southern Ocean. At 14.4 million km<sup>2</sup> (5.4 million sq mi), it is the fifth-largest continent in area after Asia, Africa, North America, and South America. About 98% of Antarctica is covered by ice, which averages at least 1.6 kilometres (1.0 mi) in thickness. On average, Antarctica is the coldest, driest and windiest continent, and has the highest average elevation of all the continents.<sup>[1]</sup> Since there is little precipitation, except at the coasts, the interior of the continent is technically the largest desert in the world. There are no permanent human residents and there is no evidence of any existing or pre-historic indigenous population. Only cold-adapted plants and animals survive there, including penguins, fur seals, mosses, lichen, and many types of algae.

The name *Antarctica* is a romanized version of the Greek compound word *Ανταρκτική* (*Antarktiké*), meaning "Opposite of the Arctic".<sup>[2]</sup> Although myths and speculation about a *Terra Australis* ("Southern Land") date back to antiquity, the first confirmed sighting of

Antarctica	
<b>Area (Overall)</b>	14,000,000 km <sup>2</sup> (5,405,430.2 sq mi)
<b>Area (ice-free) (ice-covered)</b>	280,000 km <sup>2</sup> (108,108.6 sq mi)
<b>Population (permanent)</b>	7th ≈0

**Figure 5 Featured Article Antarctica on August 19, 2008**

Before becoming featured, articles are reviewed for accuracy, neutrality, completeness, and style. Reviewing for featured article is done by any user who is willing to contribute to the featured article review process, although featured article director and his delegates is responsible for terminating of the review process. According to the community in Wikipedia, the criteria for reviewing featured articles consist of:

- A number of great attributes such as well-written, comprehensive, factually accurate, neutral and stable.
- Good style which includes a concise lead section that summarises the topic and prepares the reader for the detail in the subsequent sections; appropriate structure of hierarchical headings and a substantial but not overwhelming table of contents; and consistent citations.
- It has images and other media where appropriate, with succinct captions and acceptable copyright status.

- It has proper length and stays focused on the main topic without going into unnecessary detail.

Featured article reviewing process can improve the candidates in various ways: articles may need updating, formatting, and general copyediting. Other issues such as a failure to meet current standards of prose, comprehensiveness, factual accuracy, and neutrality, may also be addressed. Featured articles drive the quality of other Wikipedia articles as they are fabulous examples in quality. From featured articles, contributors are able to see what makes a feature article and what effort is necessary to achieve a quality article.

### 2.2.2 Five Pillars

Wikipedia has official policies and guidelines to further the goal of creating a free encyclopaedia. According to Wikipedia itself, those policies and guidelines can be summarised as five pillars that define the characteristics of Wikipedia:

- Wikipedia is an encyclopaedia incorporating elements of general encyclopaedias, specialised encyclopaedias, and almanacs. All articles must follow no original research policy, and strive for verifiable accuracy: unreferenced material may be removed.
- Wikipedia has a neutral point of view. It strives for articles that advocate no single point of view. Sometimes this requires representing multiple points of view, accurately presenting each point of view, providing context for any given point of view, and presenting no one point of view as “the truth” or “the best view.”
- Wikipedia is free content that anyone may edit. Articles can be changed by anyone and no individual controls any specific article. Any writing contribution can be edited and redistributed at will by the community.
- Wikipedia has a code of conduct. One should be civil and avoid conflicts of interest, personal attacks or sweeping generalisations.

- Wikipedia does not have firm rules besides the five general principles. One should be bold in editing, moving, and modifying articles. All prior versions of articles are kept, so there is no way to accidentally damage Wikipedia or irretrievably destroy content.

Although the five pillars are the key success to Wikipedia, they are not enforced by the community. Users do not sign up to adhere to five pillars when they create accounts. Instead, the community work with the five pillars. For instance, the pillar “Wikipedia does not have firm rules besides the five general principles” makes it possible for the community to develop the rules and policies in great flexibility. Rules can be seen as self-propagating entities that are the result of an evolving, competitive process. This perspective rejects the idea of intention, design, and agency as the primary drivers of policy development, largely because of the bounded rationality of individuals and high levels of complexity in the organisational system. Instead it is argued that rules are the result of competition for shifting attention (Butler et al. 2008).

There are a number of tools available for improving the content of articles on Wikipedia. As shown in Figure 6, wikEd (2006) is a full-featured text editor that adds enhanced text processing functions to Wikipedia. It has a number of exciting features such as wiki syntax highlighting, pasting formatted content from external word processor and regular expression searching and replacement. As wikEd is a complete rich-text pseudo-WYSIWYG editor, it encourages contributors, especially prospective contributors who are kept ashore due to the inconvenience of wiki syntax to contribute their knowledge to Wikipedia.

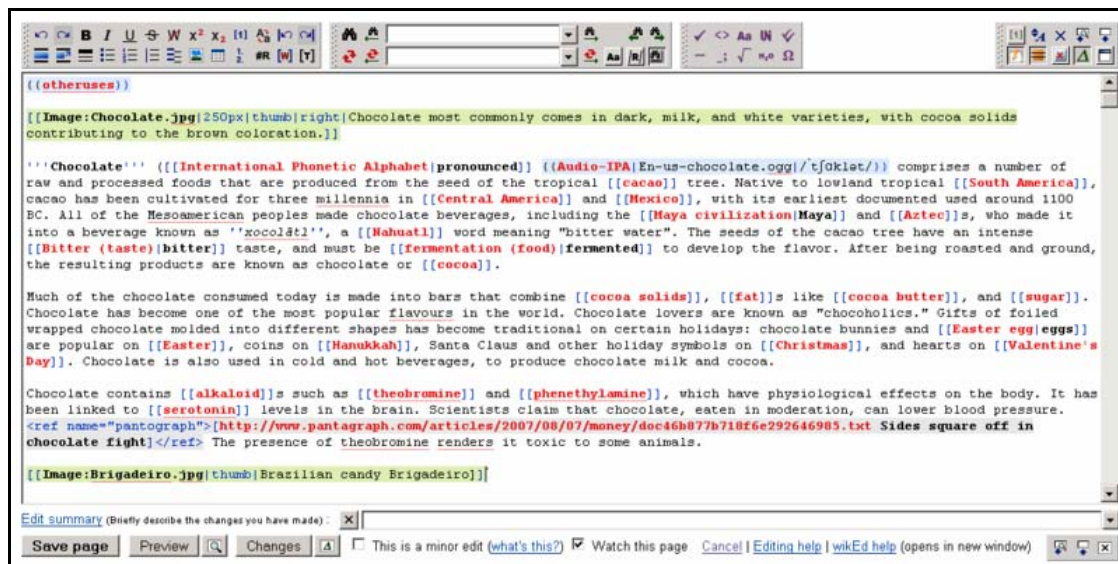


Figure 6 Screen Shoot of wikEd in Action

While Wikipedia owe its incredible growth to its openness, it also suffers from vandalism. As a result, a number of vandalism tools have been developed to protect the content from being corrupted. VandalProof (2006) allows users to peruse recent changes, watch lists, and user contributions. If vandalism is found, in one click administrators can revert it, post the appropriate warning template on the vandal’s talk page, add that person to their blacklist and add the revert to their automated vandalism log. While high volume of edits occurs each second, the vandalism and monitoring tools are essential to keep Wikipedia in a healthy status.

### 2.3 Cooperation in Wikipedia

Wikipedia is a mirror of society. Articles are created by online community with no centralised control unlike traditional encyclopaedias. There is a variety of research investigating Wikipedia from a variety of perspectives which contribute to considerations of Wikipedia cooperation.

Butler et al. (2008) study the nature and roles of policies and rules in Wikipedia. It is found that the policies in Wikipedia and the systems and mechanisms that operate around them are multi-faceted. Wikipedia itself shows that wikis are capable of supporting a broader range of structures and activities than other collaborative platforms. Wikipedia provides a valuable opportunity for using the “sidewalk design

strategy” (Evans 1990) of providing a field of grass and watching where and how the users walk, or so-called desire paths. However, this path is only recorded in a plain text, which is hard to process. It leads to the requirements of visualising the history of articles to provide a more direct way to explore and investigate Wikipedia.

Some interesting findings are found when study the correlation between contributors and articles. Ortega et al. (2008) study the inequality in the contributions to several language editions of the Wikipedia. It is found there is a high level of inequality in the total number of contributions to each Wikipedia language edition, with less than 10% of the total number of authors being responsible for more than 90% of the total number of contributions. It is also discovered that this level of inequality has remained somewhat constant in the history of every language edition. The result strongly supports the idea to observe knowledge creation and sharing process among those ten percent contributors, which can effectively reflect the cooperation phenomenon in Wikipedia.

Several insights are gained from an empirical analysis of Wikipedia (Viegas 2007). First, the community maintains a strong resilience to malicious editing, despite tremendous growth and high traffic. Second, the fastest growing areas of Wikipedia are devoted to coordination and organisation. Finally, by manually coding the content of a subset of “Talk Page”, the pages serve many purposes, notably supporting strategic planning of edits and enforcement of standard guidelines and conventions. Despite the potential for anarchy, the Wikipedia community places a strong emphasis on group coordination, policy, and process. For the purpose of this research, it is interesting to visualise this cooperation process from the article evolving perspective.

#### ***2.4 Quality of Wikipedia***

Wikipedia is considered as one of the biggest free encyclopaedias in the world. However, due to its openness, fast dynamic changing of content and lack of central coordination government, the quality of articles greatly differs from one to another. Distinguishing between good and bad quality articles is not a simple task to human users, let alone computer programs. The difficulties can be attributed to several reasons (Lim et al. 2006):

- Large number of articles for quality judgement: The larger the wiki site, the harder it is to determine the quality of each article by comparing with other articles from the same site.
- Diverse content among articles: Wide range of topics can be covered by the articles. It is extremely difficult to perform content analysis on the article to determine their qualities without human judgements and high quality benchmark collection for each topic.
- Unknown contributors: The expertise and experience of contributors are usually not explicitly captured by the collaborative software. Without knowing this, it is difficult to determine the quality of articles created by users.
- Abuse: Wiki sites with open access can easily be targets of abuse in that contributors can intentionally create articles of specific patterns to circumvent quality checking. In this case, a human expert may be able to detect such instances but designing software to detect them will be a challenge.

There is a variety of research investigating on the quality of Wikipedia. Wilkinson & Huberman (2007) study the correlation between number of edits, number of distinct editors and articles quality. Based on the mutual reinforcement principle, Lim et al. (2006) developed two models for measuring the quality of latest articles and the authority of their contributors. Stein & Hess (2007) study the featured articles on Wikipedia – articles marked by a community’s vote as being of outstanding quality. Zeng et al. (2006) develop an article fragment trust model to assess the trustworthiness of articles. Blumenstock (2008) measures article quality with a simple word count metric.

It is dangerous to refer to Wikipedia as an encyclopaedia as it lacks a central point of control. Due to the openness of Wikipedia, while dedicated and knowledgeable editors can improve pages better over time, other editors can introduce errors in fact or interpretation on purpose or in accident. The reader never knows whether the last editor is the latter group. Even if the metrics discussed in this section can be used to



gain a measurement on the quality of Wikipedia articles, they can only serve this purpose to a certain degree and none of them can be used as a method to identify a true measure of the total quality of articles.

Wilkinson & Huberman (2007) study the quality of Wikipedia articles up to 1.5 million in the English language. They demonstrate a strong overall correlation between number of edits, number of distinct editors and quality of articles. It is found the high-quality articles in Wikipedia are distinguished from the rest by a larger number of edits and distinct editors. It shows more cooperation in the development of the high-quality articles than other articles, including a strong correlation between discussion activity and article quality, more edits per editor to high-quality articles, and a markedly different pattern of editors' responses to other edits on these pages.

Based on the mutual reinforcement principle, Lim et al. (2006) developed two models for measuring the quality of latest articles and the authority of contributors. The mutual reinforcement principle is:

- Quality: An article has high quality if it is contributed by high authority authors.
- Authority: A contributor has high authority if he or she contributes high quality articles.

Two models are used to measure the quality of articles. The basic model measures the quality of an article using both the authority of contributors and the amount of contribution from each contributor. The peer review model extends the former by considering the review aspect of article content. It is shown the basic model and peer review model are able to derive article qualities (and contributor authority) from the collaboration information and edit histories through a mutual reinforcement approach.

Stein & Hess (2007) study the featured articles in Wikipedia – articles marked by a community's vote as being of outstanding quality. The research investigates the XML-dump of German Wikipedia metadata containing 976,016 regular articles. They find a relation between the quality of articles and authors: featured articles have higher rating than other articles. It matters that users with a reputation for high quality writing

contribute. Pages edited in the very beginning by authors with high reputation have a higher chance to get featured in the future.

Zeng et al. (2006) develop an article fragment trust model to assess the trustworthiness of articles by utilising Wikipedia revision history. By applying the model to articles in Geography category, it is found featured articles have the highest trustworthiness value compared to other articles. The fragment trust model has 91% classification accuracy of featured articles.

In contrast to complex methods, Blumenstock (2008) measures article quality with a simple metric - the length of the article counted in words. They test the performance of article length as a discriminant between high and low quality articles based on the assumption that featured articles are of much higher quality than random articles. As a result, by classifying articles with greater than 2,000 words as featured and those with fewer than 2,000 words as random, 96.31% accuracy in the binary classification is achieved. It is believed article length is a very good predictor of whether an article will be featured on Wikipedia.

The metrics for measuring the quality of articles on Wikipedia can be summarised as:

- Word count. The more words an article contains, the better chance the article is of high quality.
- Contribution count. The more revisions for an article, the better the article is. This is due to the fact that more number of edits reflects high focus of community contribution to the article, which improves the chance for an article being in high quality.
- Number of distinct contributors. More distinct contributors give more opinions and views when an article is created. The article has more chance to have a neutral point of view.
- High authority authors. If an author has contributed numerous high quality articles, it is likely that he or she will stay on contributing high quality content.

While all those metrics can measure the quality in some degree, they cannot give a qualitative judgement as humans do. For example, it is difficult for a machine to judge whether an article is comprehensive. Nor can a machine determine how accurate the article is. The quality of article is largely dependent on readers, who have complex criteria to judge the article in their knowledge context. Thus, one has to be careful with the quantity measurement for the quality of articles.

## ***2.5 Wikipedia Application***

Wikipedia articles not only can be viewed by web browsers but also can be integrated to build various applications.

Several applications have been developed based on Wikipedia. Milne et al. (2007) use extracted thesauri from Wikipedia to facilitate query expansion. Banerjee et al. (2007) propose a method of improving the accuracy of clustering short texts by enriching their representation with additional features from Wikipedia. Schonhofen (2006) exploits the titles and categories of Wikipedia articles to determine the characteristic of a document. Pei et al. (2008) construct a global ontology by using Wikipedia thesaurus to provide an intermediary for ontology mapping. Sinclair et al. (2007) develops a system that extracts information from the free text descriptions and try to identify the respective Wikipedia article describing each entity extracted from the text.

Milne et al. (2007) use extracted thesauri from Wikipedia to automatically and interactively facilitate query expansion. Wikipedia is particularly attractive for thesauri extraction because it represents a vast domain-independent pool of manually defined terms, concepts and relations. They develop a search interface Koru, allowing a thesaurus to be intuitively and unobtrusively used. By comparing Koru with another traditional search interface, it is found that the knowledge base provided by the thesaurus is relevant and accurate enough to make a perceptible difference to the retrieval process. However, thesaurus-based query expansion is highly dependent on the quality and relevance of the thesaurus. Nearly half of the terms are ambiguous according to Wikipedia. Although the disambiguation techniques in their research reduce the number of multiple matches, the final thesaurus still has 17% ambiguous.

As documents in the collection to derive the thesaurus are not restricted to any particular domain, irrelevant terms could be returned when users search in a particular domain.

Banerjee et al. (2007) proposes a method for improving the accuracy of clustering short texts by enriching representation with additional features from Wikipedia. It shows Wikipedia can substantially help improve clustering accuracy and in different information retrieval tasks. While popular news or blog feeds often face the problem of information overload as feed sources periodically deliver large number of items, the method could be applied to clustering similar items in the feed reader to make the information more manageable for users. However, not all the clustering algorithms achieved higher accuracy and the method does not apply to the incremental clustering problem, which is a more realistic scenario for a feed reader.

Schonhofen (2006) presents a simple method that exploits only the titles and categories of Wikipedia articles. The method can effectively characterise documents by Wikipedia categories. It is observed that the Wikipedia categories, especially when augmented by words represent documents equal or better than their full text. However, this method heavily relies on the quality of titles and categories of Wikipedia articles. The categorisation of Wikipedia articles is not always consistent. The density of the Wikipedia category net is very uneven, some topics are discussed in more detail than others. Many Wikipedia categories cover semantically unrelated articles.

Pei et al. (2008) propose a new approach of constructing a global ontology by using Wikipedia thesaurus to provide an intermediary for ontology mapping. They attempt to deduce relations among the concepts in the thesaurus by a two-step method: name mapping and logic-based mapping. From the experiments, it is confirmed that high accuracy can be achieved by giving a proper threshold for each factor. Thus it is possible to use Wikipedia knowledge to construct a global ontology. However, the name mapping does not work when they have many common related concepts. For example, “Pacific War” is incorrectly mapped to “pacific”. On the other hand, concepts relating to same terms that presenting totally different meanings result in an incorrect relation inferring for the logic-based mapping. For example, Doll is mapped as a Girl. However, Wikipedia concepts Doll did not represent the meaning of Girl.

Sinclair et al. (2007) develop a system that extracts information from the free text descriptions and try to identify the respective Wikipedia article describing each entity extracted from the text. They have focused on extracting peoples' names from the text, and aims to retrieve structured information from the Wikipedia article to augment the knowledge base. By using the whole of Wikipedia as its linkbase, the system is able to dynamically add links to any person described on Wikipedia. It is found due to its incredibly wide coverage of subjects Wikipedia is a fantastic resource for such a system. However, the system may link the person name to unrelated article on Wikipedia due to name ambiguous. Besides, their work only limits to extracting people's names. Whether the method will work for places or organisation names remains further investigation.

In summary, Wikipedia is an attractive resource for different research areas as well as for organisations and individual users. Wikipedia is not just a brunch of articles. In the knowledge management area, ontology can be built upon the collection of articles with relationship connected via hyperlinks. In the information retrieval area, thesauri can be extracted to improve the accuracy of clustering short texts as well as to facilitate query expansion for search engine. Hyperlinks to the Wikipedia articles can be injected to enrich web pages. Similar items in the feed reader can be clustered with extracted thesauri from Wikipedia to make the information more manageable for a user. Wikipedia will attract more researchers and organisations to put investigation efforts on it.

## ***2.6 Conclusion***

This chapter presented an overview of Wikipedia. It described how Wikipedia can be integrated to build various novel applications and demonstrated how Wikipedia's usefulness extends beyond its best known function as an encyclopaedia. Issues related to the quality of Wikipedia articles have been discussed and assessed including details of the guiding principles of Wikipedia, mechanisms of cooperation, featured article creation, cooperation and, in particular, existing measures of quality for Wikipedia articles.

## **3 WIKIPEDIA AS A KNOWLEDGE MANAGEMENT RESOURCE**

### ***3.1 Introduction***

The project described in this dissertation is concerned with investigating the usefulness of a visual representation of article history as a tool for Wikipedia users and to investigate the usefulness of this representation as a tool to improve knowledge creation and sharing in Wikipedia. Knowledge creation and sharing in Wikipedia must be seen in the wider context of knowledge management of Wikipedia. This chapter therefore introduces key concepts of knowledge management and offers an assessment of Wikipedia as a Knowledge Management resource.

The chapter begins by introducing the definition of knowledge, spiral of knowledge and knowledge management process. It then discusses knowledge management issues which impact Wikipedia describing the spiral of knowledge in Wikipedia and explaining why Wikipedia is a knowledge base. The chapter also illustrates Wikipedia is a good sample as an online community of practice (CoP).

### ***3.2 Knowledge Management***

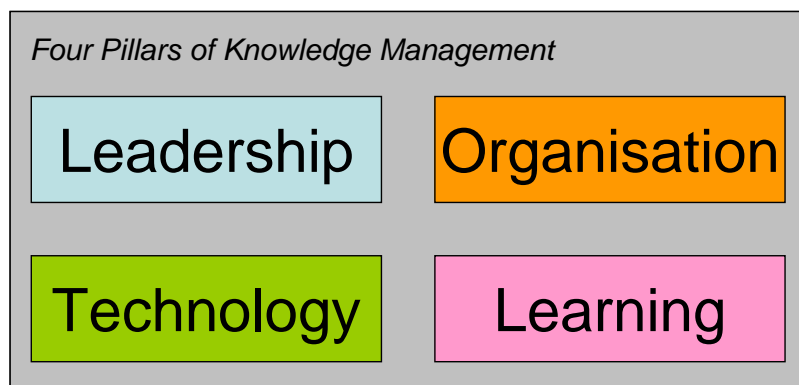
Knowledge Management is a scattered subject area. There is no single point of view on what Knowledge Management is. A variety of definitions exist for the Knowledge Management concept. While none definition considered definitive, they all offer useful insights to gaining understanding on this complex area.

Alavi and Leidner (1999) define Knowledge Management as a systemic and organisationally specified process for acquiring, organising, and communicating knowledge of employees so that other employees may make use of it to be more effective and productive in their work. Schreiber et al. (2000) defines Knowledge Management as a framework and a tool set for improving the organisations knowledge infrastructure, aimed at getting the right knowledge to the right people in the right time. Liss (1999) states that Knowledge Management is a directed process of figuring out

what knowledge individuals (have) within an organisation and then devising ways of making it available to others. These definitions try to offer a concise description on what knowledge management is, focusing on particular aspects but neglecting others. However, knowledge management is a complex and scattered area. It is hard to agree upon what knowledge management is.

Instead of defining Knowledge Management, Rao (2003) proposes the eight keys to successful knowledge management practices: connectivity, content, community, culture, cooperation, capacity, commerce and capital. This “8 Cs” framework achieves the goal of explaining Knowledge Management better than simply defining the concept. Knowledge Management is not only about technology infrastructure but also about organisation context. Different organisations expect various benefits from knowledge management such as better decisions, new business opportunities, improved motivation and retention of employees. The “8 Cs” framework gives organisations opportunity balancing among the eight practices and thus implements the effective ones to achieve maximum benefits from the knowledge management activities.

On the other hand, Bixler (2002) points out that the four pillars of knowledge management are leadership, organisation, technology and learning as shown in Figure 7. All the four pillars must be addressed to achieve successful knowledge management implementation.



**Figure 7 Four Pillars of Knowledge Management**

Knowledge Management must be implemented from top to bottom, which requires a leader at or near the top of an organisation who can provide the strong and dedicated leadership needed for cultural change. The knowledge management activities must happen within the whole organisation, from chief executive officer to decision makers, from senior managers to employees. Organisation must be tailored with a knowledge management framework and strategy, including all performance metrics and objectives, for a successful knowledge management programme. While culture change is vital to a knowledge management programme, it is the technology that provides reliable tools and infrastructure to enable knowledge creation, sharing and transferring. Learning is an integral part of knowledge management and is an ongoing activity. People create knowledge in the process of social interaction and learning. They collaborate, share knowledge and build on each other's ideas through the organisation learning. New organisation behaviour must be created to facilitate long time learning as part of the knowledge management programme.

The four pillars of Knowledge Management are adopted in this dissertation. Although there are various attempts trying to precisely define Knowledge Management, none of them covered all aspects of Knowledge Management. While the "8 Cs" framework defines the characteristics of Knowledge Management, it is too complex for use in this project. The four pillars clearly address the dimensions of Knowledge Management – people, process and technology – and can be easily applied to assess Wikipedia as will see in section 3.3.

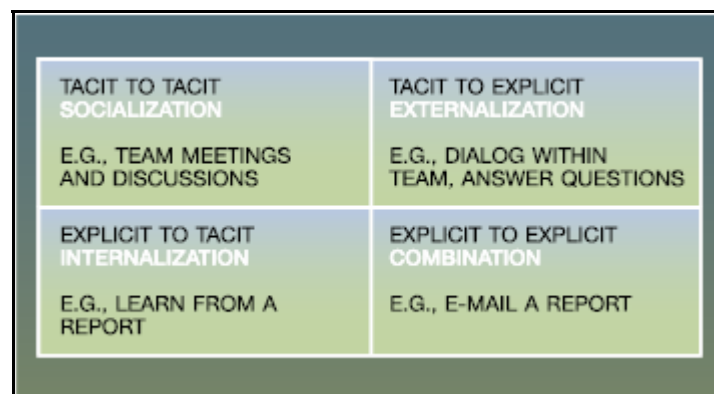
### 3.2.1 Spiral of Knowledge

According to Nonaka and Takeuchi (1995), there are two types of knowledge: tacit knowledge and explicit knowledge. Tacit knowledge is what the knower knows, which is derived from experience and embodies beliefs and values. Tacit knowledge is actionable knowledge, and therefore the most valuable. Explicit knowledge is represented by some artefact, such as a document or a video, which has typically been created with the goal of communicating with another person.

Marwick (2001) proposes the transformation of knowledge between its tacit and explicit forms. As shown in Figure 8, there are four types of knowledge transformation.



- Socialization (tacit to tacit). In the socialisation process knowledge is acquired and shared without being made explicit. Socialisation usually occurs between people or within groups of workers with a common interest.
- Externalisation (tacit to explicit). In the externalisation process tacit knowledge is transformed into explicit knowledge. Externalisation is a very hard process.
- Combination (explicit to explicit). In the combination process various sorts of explicit knowledge are brought together to form more complex or more useful knowledge.
- Internalisation (explicit to tacit). In the internalisation process tacit knowledge is acquired by examining explicit knowledge from many sources.



**Figure 8 Conversion of knowledge between tacit and explicit forms**

The spiral of knowledge processes is helpful to understand not only how knowledge is acquired and shared but also how knowledge may be created. As will see in section 3.3.1, Wikipedia is an excellent example for illustrating the spiral of knowledge.

### 3.2.2 Knowledge Management Process

According to Murray and Jones (2005), Knowledge Management embodies organisational processes that seek synergistic combination of data and information-processing capacity of information technologies, and the creative and innovative

capacity of human beings. In this definition, the process consists of capturing, organising, targeting, transferring and maintaining knowledge.

- Capture knowledge. The capture knowledge process involves finding out where the knowledge is and capture tacit and explicit knowledge. For example, an organisation can build a knowledge map points to the people that have knowledge and the places that contains knowledge.
- Organise knowledge. The organise knowledge process involves devising a taxonomy, generating a thesaurus, designing metadata and supporting data structures, generating metadata, devising and generating indexes.
- Target knowledge. The target knowledge process is about setting up data structures in the knowledge repository that hold details of users so that knowledge can be targeted to interested users, interactions with the knowledge repository can be personalised, operations on the content and the knowledge repository itself can be controlled, communities of practice (CoP) can be supported.
- Transfer knowledge. The transfer knowledge process is essentially about making knowledge content visible and available to users for sharing and collaboration purposes so that knowledge can be absorbed and put into action and new knowledge can be created.
- Maintain knowledge. The maintain knowledge is an ongoing process. It involves a number of discrete steps such as maintaining knowledge yellow pages, taxonomy, thesaurus, indexes, content, user profiles and publication services.

The process gives a guideline for organisations who want to initiate knowledge management programme. It helps the organisation to:

- Makes visible organisational knowledge no matter where it is.
- Provides access to an organisation's collective expertise anywhere in the organisation.

- Retains the organisation's knowledge in times of change.
- Exploits knowledge as an organisational asset.
- Helps to ensure that knowledge is up to date and relevant.
- Helps the organisation to do the right thing.
- Embeds knowledge in the organisation's processes
- Enables the survival of the organisation.

As will see in section 3.3.3, as a none-commercial website, Wikipedia embodies the process to facilitate knowledge creation and sharing all over the world.

### ***3.3 Wikipedia And Knowledge Management***

As discussed in section 3.2, the four pillars of knowledge management – leadership, organisation, technology and learning – clearly and fully depicts important aspects of knowledge management. Wikipedia can be assessed from the knowledge management perspective using the four pillars. There is no centralised control in Wikipedia, which means there is no authority user. This leads to powerless leadership in Wikipedia. However, leadership does implicitly exist. For example, when reviewing a featured article, it is the featured article director and his delegates responsible for terminating the review process. Wikipedia is a non profit knowledge sharing project rather than a formal organisation. No users work for Wikipedia - they contribute. As people are loosely cooperated, the concept of organisation is obscure in Wikipedia. Wikipedia employs various technologies and tools to improve the content of articles. There are tools for enhancing text processing ability, importing external resources to increase interoperability, and vandalism tools to protect the content from being corrupted. Learning in Wikipedia happens everywhere. Knowledge seekers can quickly get an overview on the concept by reading articles on Wikipedia. Knowledge contributors gain further insight on the concept through discussion and debate with each other with on “Talk Page”. As will see in next section, the spiral of knowledge also exists in Wikipedia.

### 3.3.1 Spiral of Knowledge in Wikipedia

As discussed in section 3.2.1, the two types of knowledge – tacit and explicit – can be transformed. As a knowledge creation and sharing platform, Wikipedia is a good example to show the spiral of knowledge.

- Socialisation (tacit to tacit). When conflicts arise on the content of article, contributors use “Talk Page” to discuss changes to its associated article or project page as shown in Figure 4. Knowledge flows from one contributor to another through reading or putting comments on the talk page.
- Externalisation (tacit to explicit). This is one of the significant transformations in Wikipedia. Groups of people with shared interest contribute their tacit knowledge and article is the artefact that makes the knowledge explicit and widely available to others.
- Combination (explicit to explicit). Typically, each article contains links pointing to other knowledge including internal articles on Wikipedia website and external hyperlinks to web pages or documents. Knowledge is combined in a cohesive manner so that readers can gain deeper and wider insight on their interested topics.
- Internalisation (explicit to tacit). While a plenty of people contributing to Wikipedia, there are even more huge audience reading Wikipedia articles. In this case, people gain new knowledge and insights through reading articles from Wikipedia. Knowledge is transformed from explicit to tacit.

In summary, Wikipedia supports impressive technologies to foster knowledge creation and sharing for the four categorises of knowledge transformation.

### 3.3.2 Wikipedia as Knowledge Base

A knowledge base is a collection of data, information and knowledge with an implied organisation and links to provide navigation among items within the organisation (Knowledge Base 2002). According to this definition, Wikipedia can be seen as an invaluable knowledge base. It is a collection of articles with both internal links to other

articles within Wikipedia website and external hyperlinks to other web pages and documents. Articles are encoded in wiki notation in both machine readable and human readable format.

The articles are categorised according to their natures and can be easily accessed by a static HTTP hyperlink injected with the title of articles. The hyperlinks themselves encodes valuable knowledge as discussed in section 2.5, where Schonhofen (2006) presents a method of exploiting the titles and categories of Wikipedia articles to determine the Wikipedia categories and characteristic of a document.

The advantage of Wikipedia is its availability for free and constant updating. Unlike proprietary knowledge base within organisations, Wikipedia is free of charge for all the researchers and interested users. While it contains all the revision history of articles, it always reflects the most up-to-date view on topics. That leads to zero maintaining cost for the knowledge base compared to proprietary ones within organisations. Wikipedia is an invaluable knowledge base that keeps evolving and updating all the time.

### 3.3.3 Wikipedia as Communities of Practice (CoP)

Communities of practice are groups of people who share a concern, a set of problems, or a passion about a topic, and who deepen their knowledge and expertise in this area by interacting on an ongoing basis (Wenger et al. 2002). Meanwhile, Wikipedia is an example of what can be accomplished by a disparate group of individuals, with a shared interest in a topic, working on such a foundation. It can be shown that Wikipedia by itself is a widely open community of practice. Three characteristics are crucial for communities of practice (Wenger et al. 2006):

- The domain. A community of practice is not merely a club of friends or a network of connections between people. It has an identity defined by a shared domain of interest. In Wikipedia, each article has an association group of people, which is a community, who are interested in contributing their knowledge to the content of article. Each community seeks to reach a neutral view on the concept being described on Wikipedia.

- The community. In pursuing their interest in their domain, members engage in joint activities and discussions, help each other, and share information. In Wikipedia, when conflicts arise on the content of article, contributors debate and discuss the issues through “Talk Page” as shown in Figure 4. By adding and reading comments, contributors learn from each other through creating and enhancing Wikipedia articles.
- The practice. Members of a community of practice are practitioners. They develop a shared repertoire of resources: experiences, stories, tools and ways of addressing recurring problems - in short a shared practice. For Wikipedia, communities have developed several tools to ensure the quality of Wikipedia article. As discussed in section 2.2.2, there are a variety of tools to ensure the quality of Wikipedia such as editing tools to add enhanced text processing functions to Wikipedia, importing and converting tools to reuse external knowledge and tools to monitor and detect vandalism.

Communities in Wikipedia are more likely to be active. Wikipedia serves as a good example for a successful community of practice.

### ***3.4 Conclusion***

This chapter gave several definitions for Knowledge Management and showed there is no consensus view on what Knowledge Management is. The four pillars of Knowledge Management were adopted to assess Wikipedia from Knowledge Management perspective. The chapter described the spiral of knowledge and illustrated how knowledge is transformed in Wikipedia. It also demonstrated that Wikipedia is not only a valuable knowledge base but also a good example of virtual online community of practice (CoP).

## **4 KNOWLEDGE VISUALISATION**

### ***4.1 Introduction***

The project described in this dissertation is concerned with investigating the usefulness of a visual representation of article history as a tool for Wikipedia users. This chapter therefore introduces the concept of visualisation and explores the expected benefits of employing knowledge visualisation in general.

The chapter starts by introducing key concepts of visualisation and its more common forms. It moves on to explain static and dynamic visualisation exploring the differences in uses and benefits and providing examples. It then offers an assessment on how knowledge visualisation could help with knowledge management. Finally, the chapter presents various existing examples of visualisation for Wikipedia.

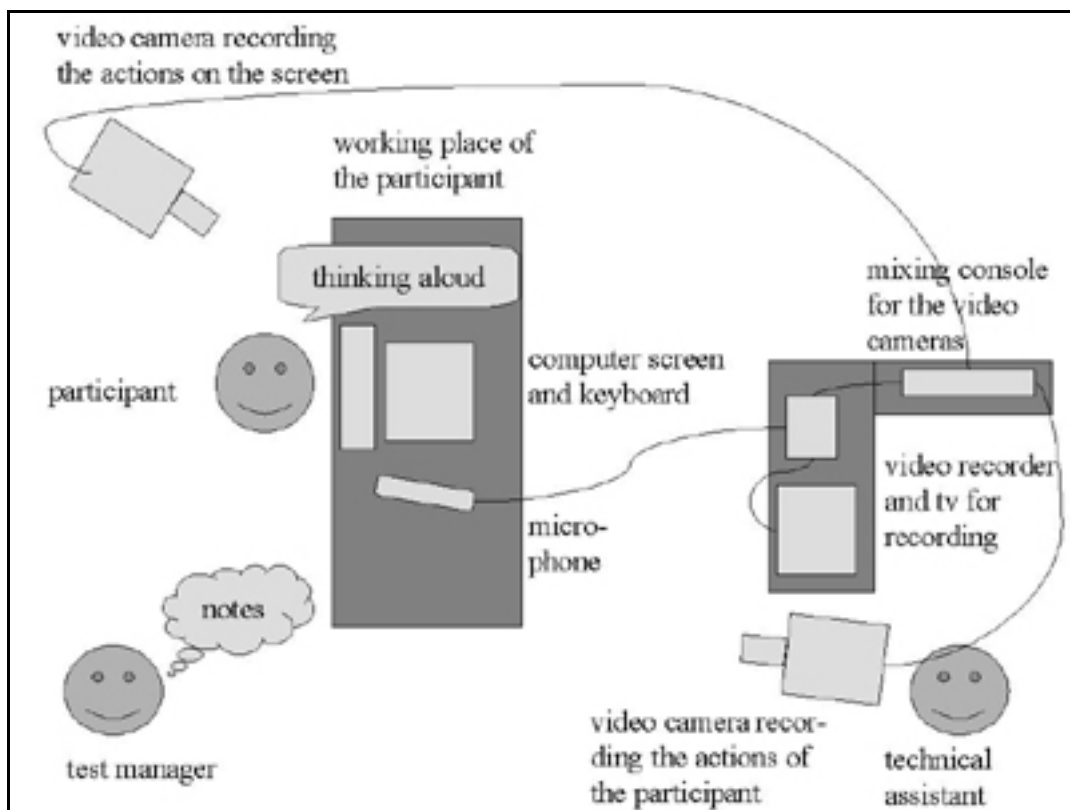
### ***4.2 Visualisation***

Visualisation is meant to support the analysis and comprehension of (often large) datasets through techniques intended to show/enhance features, patterns, clusters and trends, not always visible even when using a graphical representation (Valiati et al 2008).

Visual representations invite the user to explore his or her data. This exploration requires that the user be able to interact with the data to understand trends and anomalies, isolate and reorganise information as appropriate, and engage the analytical reasoning process. It is through interactions that the analyst achieves insight. (Hanrahan et al. 2005)

From the format perspective, Eppler & Burkhard (2005) structure the visualisation methods to six main groups: heuristic sketches, conceptual diagrams, visual metaphors, knowledge animations, knowledge maps and scientific formats.

- Heuristic Sketches. They are drawings that are used to assist the group reflection and communication process by making unstable knowledge explicit and debatable. Figure 9 shows a sketch of the user usability test for the laboratory. The sketch shows how the thinking aloud method is used for user testing in the usability laboratory. Heuristic sketches help to quickly visualise an idea. The use of a pen on a flipchart attracts the attention towards the communicator. However, the sketch might not be precise enough for wide sharing.



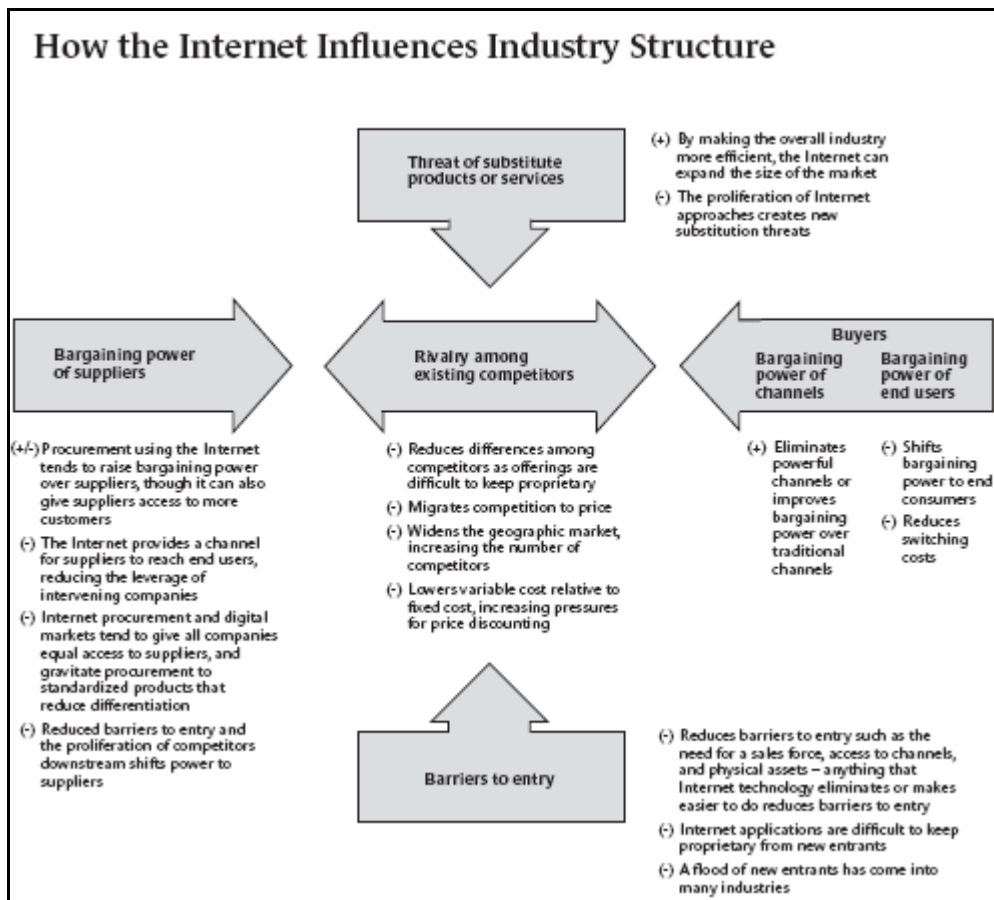
**Figure 9 A sketch of the usability lab**

*Source: Harms & Schweibenz 2001*

- Conceptual Diagrams. They are schematic depictions of abstract ideas with the help of standardised shapes (such as arrows, circles, pyramids or matrices) used to structure information and illustrate relationships. Figure 10 shows how the internet enable the reconfiguration of existing industries that had been constrained by high costs for communicating, gathering information, or accomplishing transactions. The conceptual diagram is helpful to make abstract concepts accessible and to



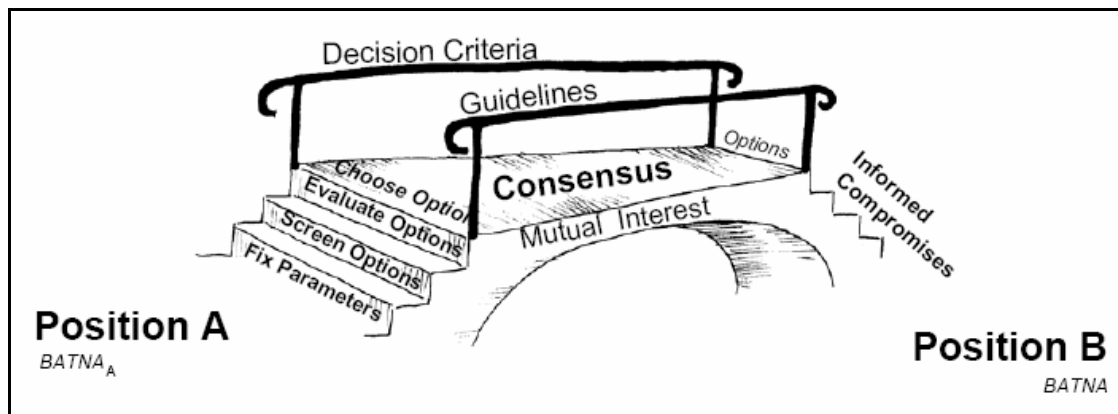
reduce the complexity to the key issues. However, it is only suitable to illustrate formula concepts and ideas due to its usage of standardised shapes.



**Figure 10 How the Internet Influences Industry Structure**

*Source: Porter 2001*

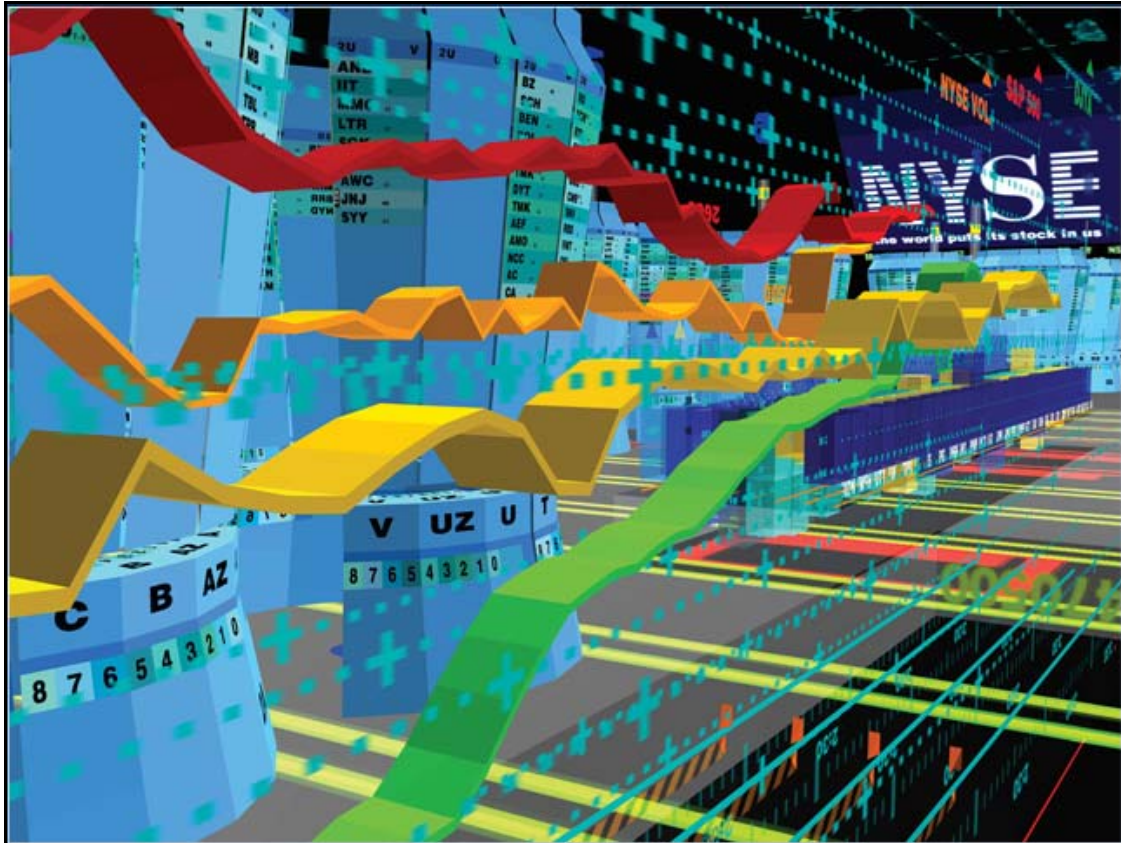
- **Visual Metaphors.** A metaphor provides the path from the understanding of something familiar to something new by carrying elements of understanding from the mastered subject to a new domain. Figure 11 uses the image of a bridge to convey how to lead successful negotiations and the picture of stairs leading to a fortress in order to illustrate the necessary steps that lead to market innovations. The visual metaphors can effectively link unfamiliar concepts to familiar ones to reduce the barrier for understanding. However, it might be difficult to find proper mappings between concepts.



**Figure 11 The Negotiation Bridge: A visual metaphor that outlines a negotiation method**

*Source: Lewicki et al 1997*

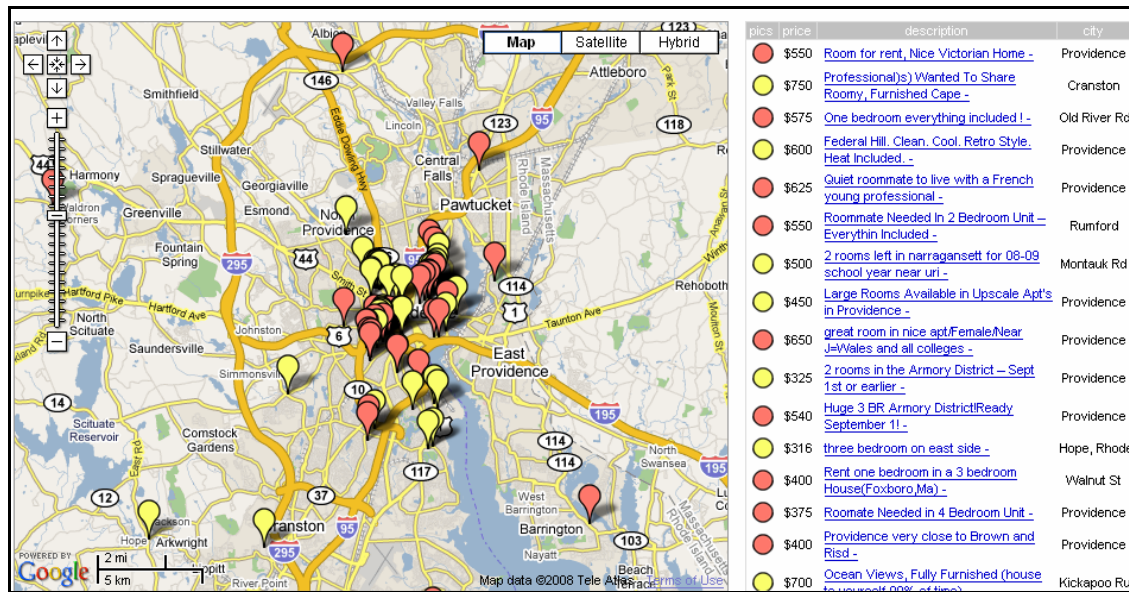
- Knowledge Animations. Knowledge animations are computer-supported interactive visualisations that allow users to control, interact and manipulate different types of information in a way that fosters knowledge creation and transfer. Figure 12 illustrates an interactive, three dimensional interface that visualises the data of the New York Stock Exchange. It is a dynamic visualisation for managers who are used to supervise and control the New York Stock Exchange. Knowledge animation is handy for combining, reducing, aggregating and assembling information. However, it is not easy to find a proper dynamic representation for the underlying large amounts of data.



**Figure 12 An Interactive Visualization helps to supervise the New York Stock Exchange**

*Source: ASYMPOTOTE 1998*

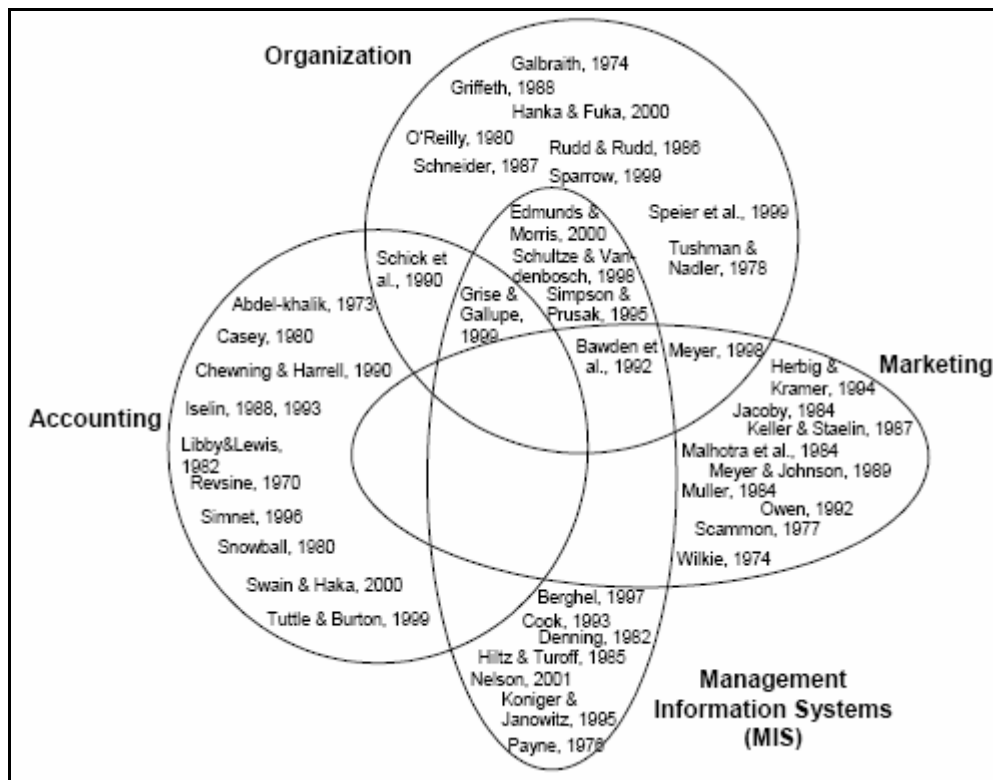
- Knowledge Maps. A knowledge map navigates and structures expertise. In general, it consists of two parts: a ground layer which represents the context for the mapping, and the individual elements that are mapped within this context. Figure 13 shows how data about houses from another web site can be interwoven with Google Map (Best 2006) to assistant accommodation finding. Knowledge map is particular good for illustrating geography based knowledge.



**Figure 13 Finding of Accommodation via a Google Map**

*Source: HousingMaps 2008*

- Scientific Charts. A scientific chart visualises domain knowledge and intellectual structures. Figure 14 shows the visual literature review diagrams for information overload. It illustrates the low degree of interdisciplinary research regarding information overload research topic. As it is manually designed by a reviewer, it could cost a lot time to construct such a diagram.



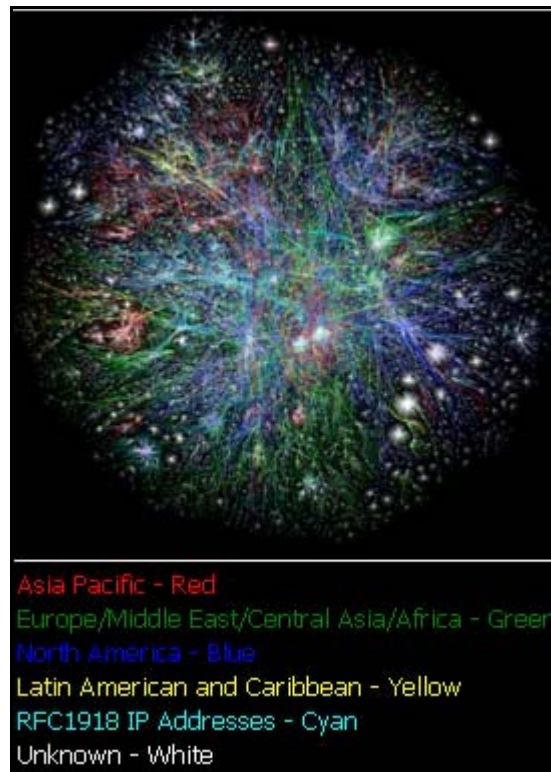
**Figure 14 A Visual Literature Review Diagram on Information Overload**

For the purpose of this research, visualisation is categorised into two types: static and dynamic visualisation. In short, a static visualisation is a snapshot or an image while a dynamic visualisation is an animation. A static visualisation allows exploring data by offering different methods such as overview, zooming in and filtering and then showing details on demand to achieve the cognition. On the other hand, dynamic visualisation helps to explore large time-varying datasets with reoccurring data objects that alter in time. Static visualisation fits casual users well as it shows a simple image for the underlying data while dynamic visualisation is novel for advanced users, providing more interactions and options to view the complex datasets in multiple angles.

### **4.3 Static Visualisation**

Static visualisation is commonly used in different areas for various purposes. The Opte project (2003) makes visual representation of the extent of the Internet. Figure 15 shows an example of Internet visualisation with over 5 million edges and estimated 50 million hop count. As shown at the bottom of the figure, different colours represent

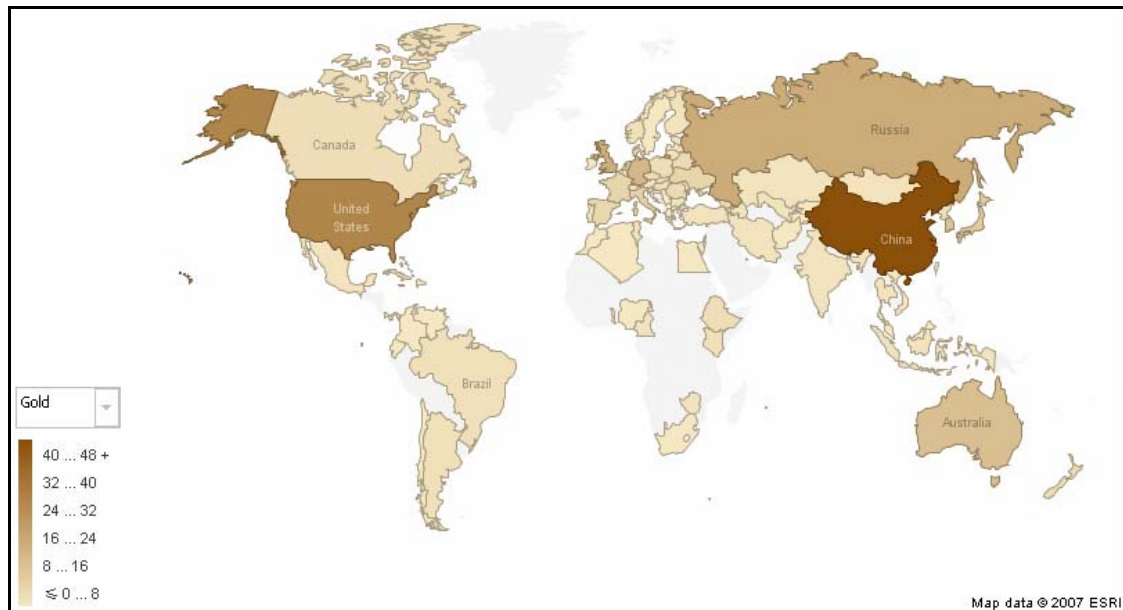
different IP addresses from the world. For example, the node in red indicates Asia Pacific while the node in blue indicates North America. It is believed that the network mapping can help teach students more about the Internet. The data represented and collected serves a multitude of purposes: modelling the Internet, analyzing wasted IP space, IP space distribution, detecting the result of natural disasters, weather, war, and art.



**Figure 15 Visualisation of Internet with over 5 Million Edges and Estimated 50 Million Hop Count.**

A tendency to integrate data with the world map has become one of the popular forms in static visualisation. Figure 16 shows the distribution of gold medals for Beijing 2008 Olympics from Many Eyes Visualisation (Viégas 2007). Each country in the map is colored according to number of gold medals won. The deeper the color is the more gold medals a country has won. It can be seen from the map that China is at first place of wining gold medals followed by United States up to August 23, 2008. The visualisation gives a quick view on what is happening in the world. In that case, knowledge is made in a form that can be easily absorbed and transmitted.





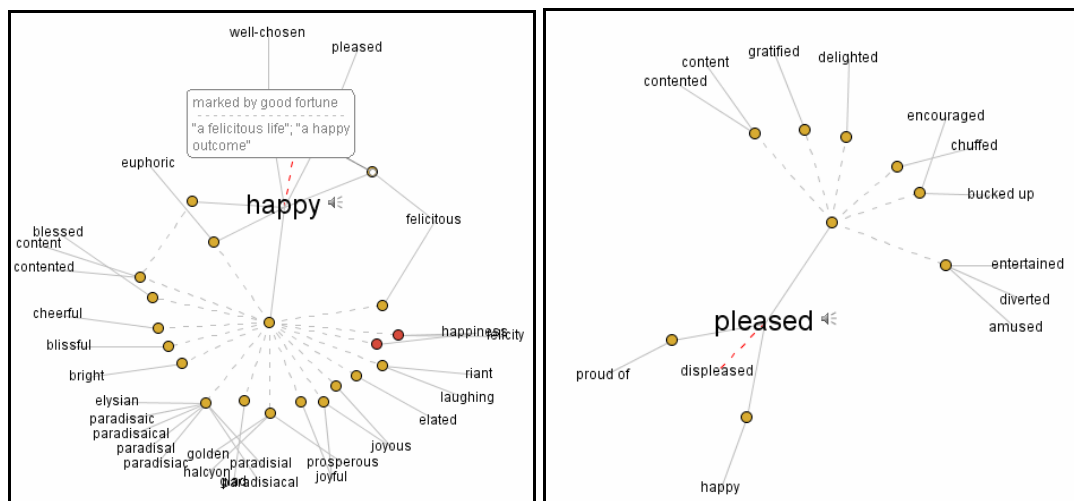
**Figure 16 Visualisation of Gold Medals for 2008 Beijing Olympics Up to August 23, 2008**

Wordle (2008) is a toy for generating “word clouds” from provided text. The clouds give greater prominence to words that appear more frequently in the source text. Clouds can be tweaked with different fonts, layouts, and color schemes. Figure 17 shows the word cloud for Knowledge Management on Wikipedia August 23, 2008. As can be seen, “knowledge” is the most occurred word followed by the word “management”. This makes sense as the article talks about Knowledge Management. It is interesting to see that “information” also occurs a lot. This could lead to more investigation on the relationship between “knowledge” and “information”. The created images can be easily shared so that knowledge is transformed and new insight could be gained. Many websites especially for portals include tag clouds/word clouds on their home pages.





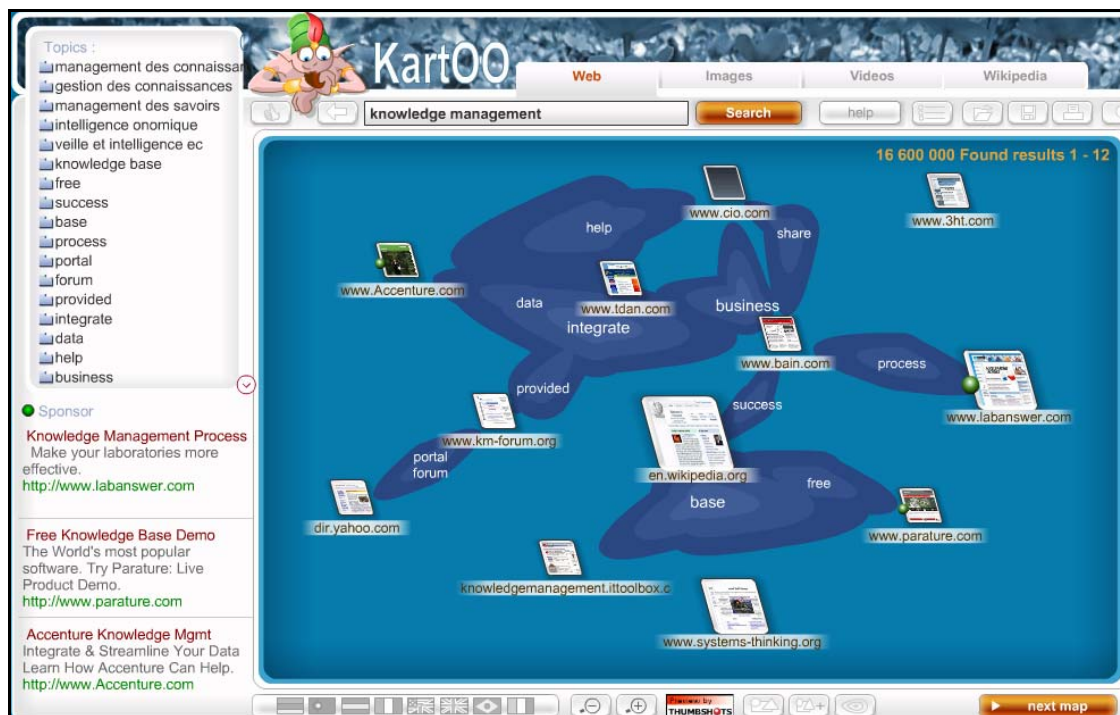
moving over the nodes, the corresponding explanation is popped up for further investigation. By clicking the related word such as “pleased” in Figure 18 (left), a new word map is generated as shown in Figure 18 (right), in which the word “pleased” is put in the centre and the word “happy” can be found at the bottom of the map. As it works like a brain by connecting related words altogether, one will discover and naturally learn. Knowledge of English language is gained by finding the right word and writing more descriptively. This dynamic visualisation differs from static visualisation in that it emphasises on interaction and learning by allowing users to explore relationships among words and to apply the word to their daily English reading and writing tasks.



**Figure 18 Word Map for HAPPY and PLEASSED Generated by Visual Theasurus**

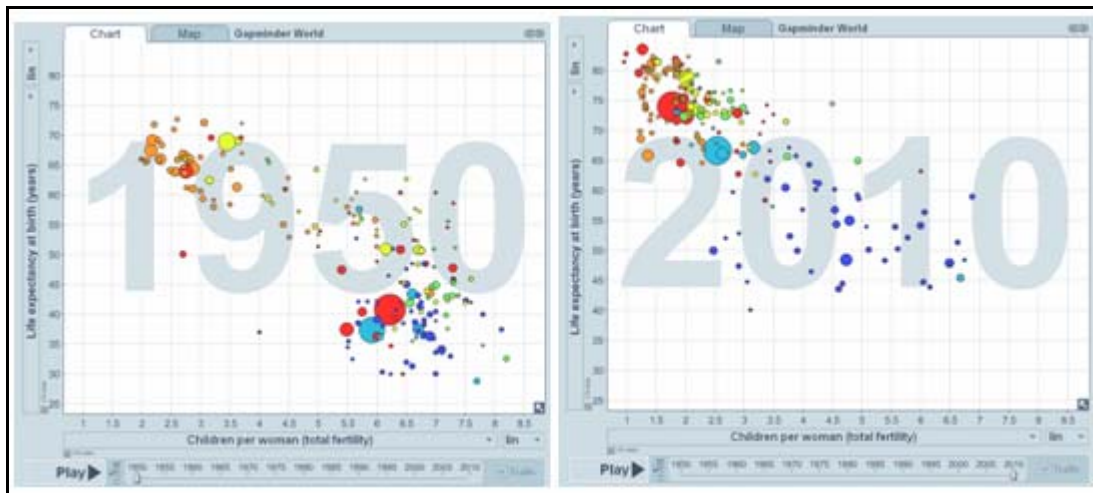
*Source: Visual Theasurus 1998*

KartOO (2008) is a visual search engine that employs several different visualisation methods. Figure 19 shows the search for “Knowledge Management”. The left side lists additional related topics to Knowledge Management, while the right gives the possible links to follow shown in clouds. When moving over one of the links, the left side is replaced with a preview of the target page in snapshot, while the right side shows the relationship to other links. By providing the search result in a visualised format and showing the relationship between results, the search engine enables users to find the most interested and related links within a seconds. Compared with traditional searching, the cost of knowledge discovery is reduced with the visual presentation of searching results.



**Figure 19 Visualised Searching for Knowledge Management by KartOO**

Gapminder (2007) is a piece of software for animation of statistics. It unveils the beauty of statistical time series by converting boring numbers into enjoyable, animated and interactive graphics. Gapminder is able to visualise a number categorisation of data including environment, economy, education, health, geograph and population, births and deaths energy, technology and infrastructure, poverty and inequality and trade, aid and investment. As shown in Figure 20, the software shows the relationship between children per woman and life expectancy on X-axis and Y-axis. Each bubble represents a country, which are plotted in the two dimensional coordinate system. The size of bubble depends on the population of the country: the larger the bubble the more population it is. The color of the bubble indicates the region of the country: yellow for America, orange for Europe & Central Asia, red for East Asia & Pacific, light blue for South Asia, dark blue for Sub-Saharan Africa and green for Middle East & North Africa.



**Figure 20 Gapminder World - Fertility versus Life expectancy in year 1950 and 2010**

The Gapminder allows gaining insights from time series data in the form of animation. As shown in Figure 20 (left), in 1950 industrialised countries, especially European countries rendered in orange bubble, have lower fertility and longer life expectancy. As shown in Figure 20 (right), in 2010 the world has completely changed. Countries all over the world have been moving towards the left top corner of the rendering area, with lower fertility and longer life expectancy. It shows that as time goes by, people are living a better life while families breed fewer children. As will see in section 6.4.2, similar components from Gapminder will be used to visualise the evolution of articles on Wikipedia.

The dynamic visualisation provides more interaction capability of exploring large volume of complex data. It is particular useful for dealing with time series data to track the evolution of objects. Chapter 6 shows how this type of powerful visualisation helps to track the content change for Wikipedia articles.

#### ***4.5 Knowledge Visualisation and Knowledge Management***

Knowledge visualisation examines the use of visual representations to improve the creation and transfer of knowledge between at least two people. Knowledge visualisation aims to transfer insights, experiences, attitudes, values, expectations, perspectives, opinions and predictions, and this in a way that enables someone else to re-construct, remember and apply these insights correctly (Eppler & Burkhard 2005).

The use of visual representations and interactions can accelerate rapid insight into complex data. Visual representations translate data into a visible form that highlights important features, including commonalities and anomalies. Visual representations make it easy for users to perceive salient aspects of their data quickly. Augmenting the cognitive reasoning process with perceptual reasoning through visual representations permits the analytical reasoning process to become faster and more focused. (Hanrahan et al. 2005)

Knowledge visualisation helps to solve several predominant, knowledge-related problems in organisations (Eppler & Burkhard 2005):

- **Knowledge transfer.** Knowledge visualisation offers a systematic approach on how visual representations can be used for the transfer of knowledge in order to increase its speed and its quality. Knowledge visualisation can serve as a conceptual bridge, linking not only minds, but also departments and professional groups. Knowledge visualisation can also facilitate inter-functional knowledge communication by making differing basic assumptions visible and communicable and by providing common contexts that help to bridge differing backgrounds.
- **Knowledge creation.** Knowledge visualisation offers great potential for the creation of new knowledge, thus enabling innovation. Knowledge visualisation offers methods to use the creative power of imagery and the possibility of fluid rearrangements and changes. It enables groups to create new knowledge, for instance by use of heuristic sketches or rich graphic metaphors. Unlike text, these graphic formats can be quickly and collectively changed and thus propagate the rapid and joint improvement of ideas.
- **Information overload.** Knowledge visualisation can be used as an effective strategy against information overload. Knowledge visualisations help to compress large amounts of information with the help of analytical frameworks, theories, and models that absorb complexity and render it accessible.

Knowledge visualisation plays an important role in the spiral of knowledge. Heuristic sketches as discussed in section 4.2 make the tacit knowledge explicit and debatable. Knowledge can be visualised in a number of angles and then combined to form more complex insights. The visual representation of knowledge can be widely shared and knowledge is internalised to individuals by manipulating the visualisation.

Knowledge visualisation also assists the knowledge management processes, especially for capturing knowledge and transferring knowledge process. In the capturing knowledge process, knowledge is encoded in the visual representation rather than the plain boring data. Visualisation increases the interruption ability of knowledge and thus improves the value of knowledge. In the transferring knowledge process, as the visual representation of knowledge is more intuitive and impressive, it can accelerate the speed of knowledge transferring and improve the quality of knowledge sharing.

#### ***4.6 Visualisation of Wikipedia***

The idea of using visualisation to monitor Wikipedia activity is an area currently being investigated. In this section, a number of visualisation approaches on Wikipedia will be discussed and the usefulness of those approaches will be assessed.

##### **4.6.1 Trends in Revision History**

Viégas et al. (2004) introduce an exploratory data analysis tool, the history flow visualisation, which makes broad trends in revision histories immediately visible and is effective in revealing patterns within the wiki context. As shown in Figure 21, each version of the document is represented by a vertical “revision line” with length proportional to the length of its text. The contributors are each assigned a different colour in the visualisation, and sections of each revision line are coloured according to who originally authored them. In order to visually link sections of text that have been kept the same between consecutive versions, the tool draws shaded connections between corresponding segments on adjacent revision lines. The tool lets the space between successive revision lines be proportional to the time between the revision dates.



**Figure 21 History Flow Visualisation of the Wikipedia Entry on 'Evolution', 2006**

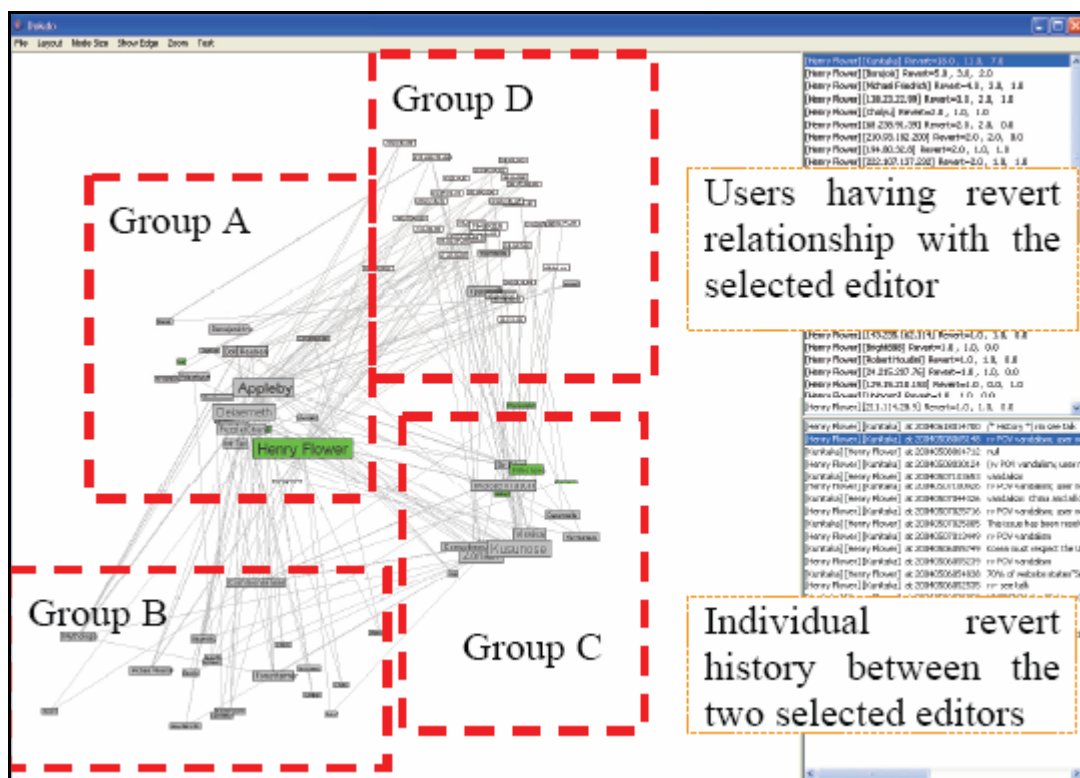
The history flow visualisation tool can be used to reveal several common patterns of collaboration and negotiation including vandalism and repair; anonymity versus named authorship; negotiation; and content stability. For example, mass deletions - one common form of vandalism in Wikipedia - are easily spotted in the visualisations because they appear as breaks in the continuous horizontal flow of changes.

However, the tool does not track the content change of articles. Although it keeps tracking on the length of text contributed by each member, it does not look at the content text itself. This limits the users only to see which portion of article is being edited without telling what is being edited. It will be more straightforward and helpful to observe the evolution of article from the content perspective view in that user is able to not only get a quick overview on the organisation of the article but also tell the community focus on the article from time to time.



#### 4.6.2 Visualisation of Wikipedia Collaboration

Suh et al. (2007) develops a user conflict model based on users' editing histories, specifically revisions that void previous edits, known as "reverts". Based on the model, a tool called Revert Graph is developed to visualise the revert relationships between opinion groups as shown in Figure 22. Node size is proportional to the log of the number of reverts or revisions. The thickness of the edges represent the degree of revert relationships between users. Nodes are color-coded based on users' registration status: an administrator in green, a normal registered user in grey and an unregistered anonymous in white. The tool provides users to drill down the graph allowing investigation to the level of an individual revert.



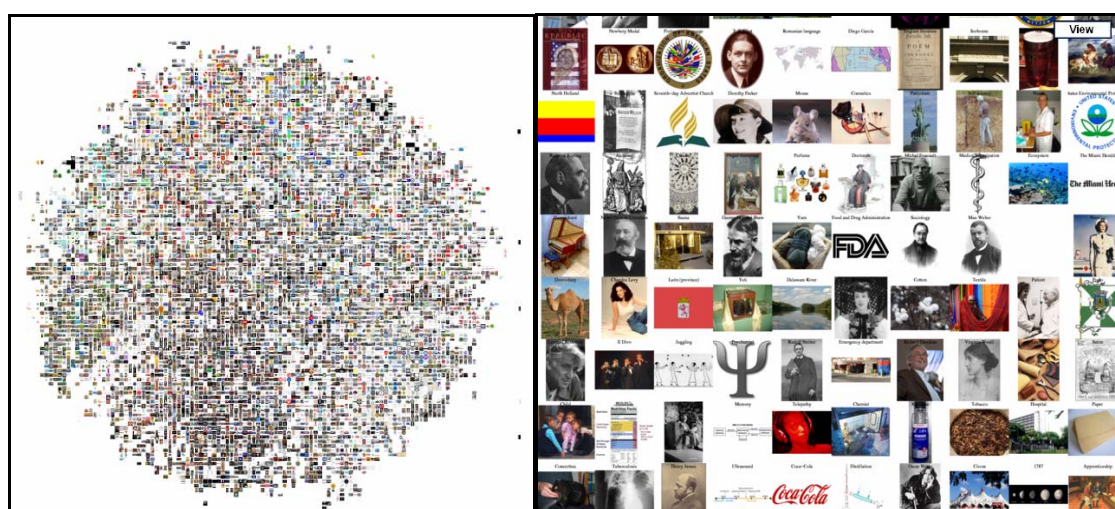
**Figure 22** Revert Graph uses force directed layout to simulate social structures between users.

The tool is capable of dividing users into groups with different opinions. It can effectively show edit wars and debating among groups, especially for controversial articles. It is useful for figuring out distinct roles in contributors. For example, the tool tells there are a group of users attempting to mediate between user groups with divergent points of view. These users are not active in expressing a particular view.

Instead, they usually revert edits from many other user groups. The limitation of the tool is that it only works with the article with a lot of reverts. Little insights can be gained for articles without many revert while articles contain a plenty of content revisions. Nor can the tool explain what the groups are arguing about and why. As content of articles attracts more focus than contributors of articles, it will be useful to visualise reverts of content. This will give more knowledge on which part of article is controversial so that further investigation can be performed on the conflicts.

#### 4.6.3 Mosaic Visualisation of Wikipedia

Mosaic Wikipedia Visualisation (2008) attempts to show which topics are contained in the online Wikipedia, and those most hotly contested. As shown in Figure 23, the visualisation is a chaotic-looking mosaic, which is created clusters of 300 or so articles that touch on a related topic, such as a religion or a famous person. For each cluster one picture is taken from the most popular article and laid out in a circular grid. Atop the grid are coloured dots showing how often and how recently each article has been edited. The larger, darker dots mean more intense activity. However, it is a type of static visualisation, which cannot instantly reflect the hot activities. It will be great if the image is updated in real time so that Wikipedia administrators can spot where arguments are taking place.



**Figure 23 Mosaic Visualisation for Wikipedia**



#### 4.6.4 Visual Side of Wikipedia

Instead of focusing on the pure text collaboration of Wikipedia, Viégas (2007) studies the visual side of the online encyclopaedia such as images, maps, diagrams, illustrations. It tries to find the difference in collaborating around images as opposed to text.

A survey is conducted, in which participants are selected from the list of users who had contributed images to the “Featured Pictures” page in the English Wikipedia. It is found collaboration around images presents a series of challenges for wiki adopters. The technical infrastructure needed to support image editing is completely external to wiki platforms, which means several key aspects of wiki collaboration features are not available to image creators at present. For example, the lack of public versioning history is a key difference from how text gets edited on wikis and it carries critical consequences to users’ ability to engage in collaborative image editing. By not being able to easily revert back to earlier, public versions of pictures, image contributors do not experience the same level of flexibility that text editors encounter in a wiki site.

One of the most difficult problems for Wikipedia sysops and editors is to quickly take a picture of the current structure and evolution over time of a certain article. However, current researches on Wikipedia visualisation are mainly focusing on the revision and editing patterns of the article rather than the article content itself. This brings up the requirements of visualising the content change of articles to facilitate knowledge creation, sharing and transferring for Wikipedia users and to allow academic researchers to evaluate the usefulness of such visualisation.

### ***4.7 Conclusion***

This chapter explained the concept of visualisation. The chapter detailed the static and dynamic visualisation and gave some examples. It then described knowledge visualisation and how it can help to facilitate knowledge creation and sharing in knowledge management. In addition, the chapter discussed various form of Wikipedia visualisation and concluded that tracking content change of Wikipedia is an untouched research area worth investigating.

## **5 STATIC VISUALISATION OF CONTENT CHANGE IN WIKIPEDIA**

### ***5.1 Introduction***

This chapter outlines the requirements for the Wikipedia visualisation. It discusses why the Knowledge Management article on Wikipedia is a suitable test bed for visualisation. It then describes the static visualisation process in three stages: extracting text from Wikipedia, parsing it for input and creating a static visualisation. The chapter critically assesses the usefulness of the static visualisation and conclude why it is not particularly useful.

### ***5.2 Requirements for Visualisation***

Current research for Wikipedia visualisation mainly focuses on the revision and editing patterns rather than on the content of article itself. This research proposes a method to visualise the content change of articles on Wikipedia. By using the visualisation tool developed, the user should be able to tell what is actually happening to the article from the content perspective view. For example, the tool should show the structures of article at a certain period of time as well as the evolution of structure as time goes by. With the tool, the user should be able to observe community focus on articles from time to time as well as the correlation between sections when the content shifts.

Wikipedia has no centralised control mechanism unlike traditional encyclopaedias. Each article has many revisions and it is difficult to get a clear picture of how an article has reach its current state being. Because it is difficult to see the path by which different people have edited, amended, deleted and corrected points in an article.

Articles are created by a section of a community which may not reflect the true views of the discipline to which article refers. There are potential bias as articles are contributed by a particular section of people who have particular bias about the topic.

As articles are always evolving, which means there is no final version of articles. To understand how the knowledge is changing within an article requires a visualisation process.

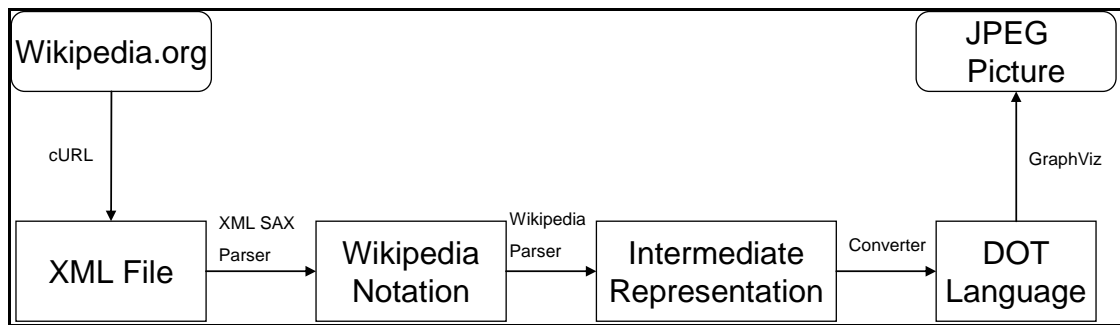
### ***5.3 The Test Bed***

The test bed for this research is the Knowledge Management article on Wikipedia. As shown in Figure 1, the article starts with an introduction section presenting the community view on definition of Knowledge Management and continues to present the contents section by section. Figure 3 shows the revision page for the article, which lists all the edits done to the article. In detail, the revision history contains the date and time of the editing, contributor as well as the editing comments.

The Knowledge Management article is particularly interesting as a visualisation topic because Knowledge Management is a cross discipline research area. The article itself presents a number of perspectives, the technical perspective, the organisational perspective, and the ecological perspective. If one views the Knowledge Management article, it could be expected that at particular moment in time, members holding particular perspective will be contributing and others will be contributing less so. It will be interesting to see how that would reflect the changing content.

### ***5.4 Visualisation Process and Supporting Technical Architecture***

The static visualisation focuses on the structure of a particular article, especially the sections and hyperlinks. The three stages of static visualisation are shown in Figure 24. The first stage is to retrieve data from Wikipedia website, which is mainly achieved by the network file transferring tool cURL (1996) as will be discussed in section 5.4.1.1. The second stage is to parse XML and Wikipedia notation for further analysis. As a result, an intermediate representation of sections and hyperlinks in the article is generated. The last stage is to convert intermediate representation to DOT language, which can be recognised by a visualisation tool GraphViz (Ellson et al. 2002) as will be discussed in section 5.4.3.



**Figure 24 Process of Static Visualisation**

In the following sub sections, each part of the process as well as the technical architecture will be described in detail to create the static visualisation.

#### 5.4.1 Retrieving Data from Wikipedia

Wikipedia offers free copies of all available content to interested users. All text content is licensed under the GNU Free Documentation License (2002). Table 1 details the type of dump available from Wikipedia. It can be seen that the complete page edit history (pages-meta-history.xml.bz2) is huge - up to more than 147.7 Gigabytes. The archive (pages-meta-current.xml.bz2) of current pages including discussion and user pages is also big – up to 6.4 Gigabytes. Even the meta-data about editing history (stub-meta-history.xml.gz) reaches 6.7 Gigabytes. There are also some useful dumps such as the abstract pages (abstract.xml) of each article, list of page titles (all-titles-in-ns0.gz), page to page linking records (pagelinks.sql.gz).

<b>File Name</b>	<b>Size</b>	<b>Description</b>
pages-meta-history.xml.bz2	More than 147.7 GB	All pages with complete page edit history
pages-meta-current.xml.bz2	6.4 GB	All the current pages only, including discussion and user pages.
pages-articles.xml.bz2	3.5 GB	Articles, templates, image descriptions, and primary meta-pages. This contains current versions of article content
stub-meta-history.xml.gz	6.7 GB	These files contain no page text, only revision metadata.
stub-meta-current.xml.gz	710.6 MB	
stub-articles.xml.gz	383.9 MB	
abstract.xml	1.8 GB	Extracted page abstracts for Yahoo
all-titles-in-ns0.gz	26.2 MB	List of page titles
redirect.sql.gz	15.7 MB	Redirect list
page_restrictions.sql.gz	197 KB	Newer per-page restrictions table
page.sql.gz	384.8 MB	Base per-page data
user_groups.sql.gz	12 KB	User group assignments
logging.sql.gz	386.4 MB	Data for various events (deletions, uploads, etc)
interwiki.sql.gz	7 KB	Set of defined interwiki prefixes and links for this wiki
langlinks.sql.gz	51.7 MB	Wiki interlanguage link records
Externallinks.sql.gz	593.1 MB	Wiki external URL link records
Templatelinks.sql.gz	177.7 MB	Wiki template inclusion link records
Imagelinks.sql.gz	99.7 MB	Wiki image usage records
Categorylinks.sql.gz	295.0 MB	Wiki category membership link records
pagelinks.sql.gz	1.5 GB	Wiki page-to-page link records
oldimage.sql.gz	12.0 MB	Metadata on prior versions of uploaded images
image.sql.gz	81.5 MB	Metadata on current versions of uploaded images
site_stats.sql.gz	456 bytes	A few statistics such as the page count

**Table 1 Summary for Wikipedia.org Database Dump on 20080312**

In past years, Wikipedia used to provide SQL dump for pages, revision history and text. However, in the middle of 2005, they upgraded the Wikimedia sites, which use a very different database layout than earlier versions. Changes to the backend storage are aggressive. As a result, Wikipedia provides XML dump of article histories for forward and backward compatibility without requiring authors of third-party dump processing or statistics tools to reproduce every internal hack.

Although the archived dump is impressive, it is not ideal for use in this research for the following reasons:

- Failure of data dump. The archive dump is easy to fail due to the large volume of data. That means it is hard to get the required dump for research.
- Large volume of data. The data is huge, with up to more than 147.7 Gigabytes in compressed data. According to Wikipedia, the size of uncompressing archive could become up to 100 times. The limitation of the data storage makes it quite hard to hold such large volume of data. Meanwhile, as the research is only interested in the Knowledge Management article, there is no need to download the whole archive.
- Spread of interested data. As the research is interested in not only the text of pages but also the editing history of pages, using the database dump requires extracting data from different archive such as pages-meta-history.xml.bz2 and stub-meta-history.xml.gz. This makes it even harder to work with several archives regarding their volume size.

On the other hand, Wikipedia offers alternative choice to export text and editing history of a particular article wrapped in XML format. As the research only interested in a small portion of articles along with their editing histories, it would be handy to download the dump for the interested articles. Another benefit is that the latest version of data is returned due to instant data retrieving.

One can download a single article dump from Wikipedia by issuing a HTTP POST request. For example, one can export the oldest 100 history revisions for Knowledge Management by issuing the following HTTP POST request:

```
http://en.wikipedia.org/w/index.php?title=Special:Export&pages=Knowledge
Management&offset=&limit=100&action=submit
```

The key parameters for the exporting are shown in Table 2.

Parameters	Description
Pages	A list of page titles, separated by linefeed characters.
Offset	The timestamp at which to start, non-inclusive.
Limit	The maximum number of revisions to return up to 100.

**Table 2 HTTP Parameters for Exporting Articles from Wikipedia.org**

The problem with this method is that only one hundred revision histories at maximum are returned per HTTP request. To retrieve the next one hundred records, one needs to specify the timestamp to start. For example, by setting the offset parameter to 2002-01-27T20:25:56Z, the following request will return the next 100 revisions newer than 2002-01-27 20:25:56:

```
http://en.wikipedia.org/w/index.php?title=Special:Export&pages=Knowledge
Management&offset=2002-01-27T20:25:56Z &limit=100&action=submit
```

The work around is to keep downloading the dump data while copy the timestamp from the last revision of the previous query. For example, if there are five hundred revisions, they will be downloaded in separation for five times, one hundred revisions each.

#### 5.4.1.1 The Tool

In order to automatically retrieve dump from Wikipedia website, several tools have been considered for appropriateness.

Wget (1996) is a free software package for retrieving files using HTTP, HTTPS and FTP, the most widely-used Internet protocols. It is a non-interactive command line tool, which can easily be called from scripts, cron jobs and terminals.

cURL (1996) is a command line tool for transferring files with URL syntax. It supports over 10 transfer protocols and a wide range of platforms. cURL has a number of great features such as SSL certificates, HTTP POST, HTTP PUT, FTP uploading, proxies, cookies, authentication, file transfer resume.

cURL is selected for crawling revisions from Wikipedia for the following reasons:

- cURL supports multiple programming language. It supports almost the modern programming languages such as C/C++, Java, PHP and Smalltalk. It is a cross-platform library with a stable API that can be used by each and everyone.
- Pipes. cURL is more in the traditional unix-style, it sends more stuff to stdout, and reads more from stdin in a “everything is a pipe” manner. This feature makes the tool callable by other process.
- Single shot. cURL is made to do single-shot transfers of data. It transfers just the URLs that the user specifies, and does not contain any recursive downloading logic nor any sort of HTML parser.
- More protocols. cURL supports more protocols including FTP, FTPS, HTTP, HTTPS, SCP, SFTP, TFTP, TELNET, DICT, LDAP, LDAPS and FILE.
- More portable. Ironically cURL builds and runs on lots of more platforms such as DOS, Linux, OS/2, Solaris and Windows.
- More SSL libraries and SSL support. cURL can be built with one out of four different SSL/TLS libraries, and it offers more control and wider support for protocol details.



- cURL supports more HTTP authentication methods, and especially when try over HTTP proxies.
- cURL can emulate browsers and do HTTP automation to a wider extent.

One of the features of cURL is that it can emulate the behaviour of web browser by automating HTTP jobs. For example, the HTML form shown in Table 3 can be submitted with cURL command:

```
curl -d "birthyear=1905&press=%20OK%20" www.hotmail.com/when/junk.cgi
```

---

```
<form method="POST" action="junk.cgi">
  <input type="text" name="birthyear">
  <input type="submit" name="press" value=" OK ">
</form>
```

---

**Table 3 Example of HTML Form using POST Request Method**

#### 5.4.1.2 Crawl Data

cURL gives great simplicity to grab data from Wikipedia. For example, one can download the dump for Knowledge Management by issuing the cURL command:

```
curl -d "title=Special:Export&pages=Knowledge Management
&offset=&limit=100&action=submit" http://en.wikipedia.org/w/index.php
```

In response, the data returned in XML format is shown in Table 4. Section 5.4.2.1 and 5.4.2.3 show how the XML and wiki text notation can be parsed for further analysis.

---

```

<page>
<title>Knowledge Management</title><id>72896</id>
<revision>
  <id>160446</id>
  <timestamp>2002-08-17T22:14:20Z</timestamp>
  <contributor>
    <username>Pichai Asokan</username>
    <id>3509</id>
  </contributor>
  <comment>Created the page</comment>
  <text xml:space="preserve">&lt;b&gt;Knowledge
Management&lt;/b&gt; is a term associated with the processes for the creation,
dissemination and utilization of knowledge.
  </text>
</revision>
<!-- Other Revisions -->
<revision>...</revision>
</page>

```

---

**Table 4 Example of Returned XML Dump for Knowledge Management**

#### 5.4.2 Parsing the Data For Analysis

The data parsing stage consists of two steps: XML parsing and wiki notation parsing. The XML dump contains multiple revisions of an article. The first step is to retrieve various data from the XML dump. This includes the timestamp, the contributor, the comments as well as the Wikipedia notation for revisions. The second step is to parse the wiki notation and retrieve sections and hyperlinks from each revision. In the following sections, the two steps will be discussed in detail.

##### 5.4.2.1 XML Parsing

Once the XML dump is crawled, it can be processed by any XML parser. There are two typical ways to process XML data, SAX and DOM.

SAX stands for Simple API for XML (Harold 2002). It is an event based parser that invokes methods when mark-up, such as a start tag or an end tag, is encountered. No tree structure is created - data is passed to the application from the XML document as it is found. SAX parsers are typically used for reading XML documents that will not be modified. SAX-based parsers are available for a variety of programming languages.

The XML DOM (XML Document Object Model) defines a standard way for accessing and manipulating XML documents (Harold 2002). The DOM views XML documents as a tree-structure. All elements can be accessed through the DOM tree. Their content (text and attributes) can be modified or deleted, and new elements can be created. The elements, their text, and their attributes are all known as nodes. DOM parsers are typically used for manipulating and transforming XML documents.

Table 5 shows the capabilities and limitations of SAX parsers.

<b>Capability</b>	<b>Limitation</b>
Search a document for an element containing a keyword	Re-order the elements in a document
Print out formatted content	Resolve cross-references between elements
Modify an XML document by making small changes, such as fixing spelling and renaming elements	Verify ID-IDREF links
Read data to build a complex data structure	Validate an XML document

**Table 5 Capability and Limitation of SAX Processor**

In this research, the SAX parser is selected to process XML data as it has two advantages over DOM parser:

- XML as data source. The data in XML format will be loaded into local database for further processing. It is mainly for read only purpose rather than manipulation. SAX is a perfect candidate to do this job.

- Less resource consumption. The DOM reads all the data and builds an internal tree representation in memory; it consumes much more resource than SAX parser. The returned XML dump can potentially be very large. While SAX only reads a small portion of data at one time, it consumes less resource.

Table 4 shows the fragment of revision for Knowledge Management. As can be seen, each revision consist of a revision id, the date time when the revision is created, the contributor, the revision comment and the revised text. In addition, the revision type is encoded, which can be seen from the XML Schema Description by MediaWiki export system shown in Table 6.

---

```

<?xml version="1.0" encoding="UTF-8" ?>
<!-- Other Complex Type Definition -->
<complexType name="RevisionType">
  <sequence>
    <element name="id" type="positiveInteger" minOccurs="0"/>
    <element name="timestamp" type="dateTime"/>
    <element name="contributor" type="mw:ContributorType"/>
    <element name="minor" minOccurs="0" />
    <element name="comment" type="string" minOccurs="0"/>
    <element name="text" type="mw:TextType" />
  </sequence>
</complexType>
<complexType name="ContributorType">
  <sequence>
    <element name="username" type="string" minOccurs="0"/>
    <element name="id" type="positiveInteger" minOccurs="0" />
    <element name="ip" type="string" minOccurs="0"/>
  </sequence>
</complexType>
<!-- Other Complex Type Definition -->

```

---

**Table 6 XSD Output by MediaWiki's Special:Export System**

### 5.4.2.2 Wiki Markup

The wiki markup is the syntax system that one can use to format a Wikipedia page. Table 7 summarise the typical wiki format syntax. One can do basic text formatting such as bolding and italicize text. Section headings are used to organise an article in hierarchy structure by dividing content into sections, subsections and so on. The top level section heading is encoded by wrapping two equal signs (==) around the section title. More “equals” (=) signs creates a subsection. There can be two types of links within the text. An internal link points to other articles within the Wikipedia system and an external links points to the world wide resource in URL.

Type of Format	Appearance of Text	Corresponding Wiki Notation
Basic text formatting	3 apostrophes will <b>bold the text.</b>	3 apostrophes will "'bold the text'"
	You can <i>italicize text</i> by putting 2 apostrophes on each side.	You can "italicize text" by putting 2 apostrophes on each side.
Section headings	<b>Section headings</b>	== Section headings ==
	<b>Subsection</b>	=== Subsection ===
Links	Here's a link to a page named <a href="#">Official position</a> .	Here's a link to a page named [[Official position]].
	<a href="#">Intentionally permanent red link</a> is a page that doesn't exist yet.	[[Intentionally permanent red link]] is a page that doesn't exist yet.
	You can make an external link just by typing a URL: <a href="http://www.nupedia.com">http://www.nupedia.com</a>	You can make an external link just by typing a URL: http://www.nupedia.com

**Table 7 Typical Wiki Markup**

#### 5.4.2.3 Parsing Wikipedia Notation for Static Visualisation

As discussed in section 5.4.2.2, an article is typically organised in a hierarchy structure with sections and subsections. Each section contains text description and hyperlinks to other resources. The aim is to extract sections and hyperlinks and their relationships from articles by parsing the wiki notation. The result is temporarily stored in an intermediate data format, which will be converted to the DOT language as will be discussed in section 5.4.3 for generating article map.

java-wikipedia-parser (Steven 2007) is an open source Wikipedia parser. The parser is fast, consumes little memory and well tested. It supports most of the typical Wikipedia notation including tables, text decoration, ordered and unordered list, headings, regular links and smart links, literals and indents.

java-wikipedia-parser is an event based parser. Whenever it encounters an element of, it will pass the control of processing to programmers. For example, when a heading is recognised by the parser, it will call the corresponding function where the programmer instructs how to process the heading. As the parser goes through the headings and links in Wikipedia notation, an intermediate representation of sections and hyperlinks is generated.

#### 5.4.3 Static Visualisation with GraphViz

GraphViz (Ellson et al. 2002) is a static visualisation tool that can represent structural information as diagrams of abstract graphs and networks. Its layout programs take descriptions of graphs in a simple text language and make diagrams in images.

DOT language is part of the GraphViz package for taking graph description as input and renders the graph in several formats. DOT describes graphs in human readable form and supports both directed graphs and undirected graphs. By running the DOT script shown in Table 8, a directed graph with four nodes and three edges is generated as shown in Figure 25. The intermediate representation of article is converted to the DOT description by a tool developed in house for this research.

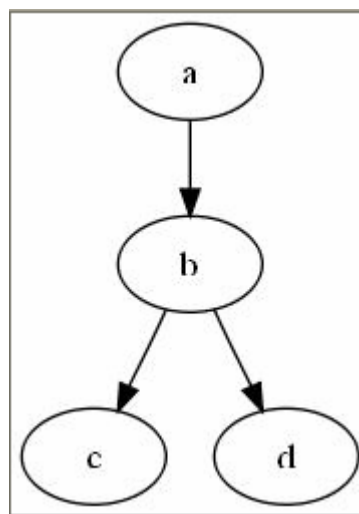
---

Digraph graphname

```
{  
  a -> b -> c;  
  b -> d;  
}
```

---

**Table 8 DOT Description for Generating A Simple Directed Graphs**



**Figure 25 Sample Directed Graph Generated by GraphViz**

#### 5.4.4 The Result

The result of static visualisation is shown in Figure 26. The visualisation is a network graph containing four types of node linked by directed edges. The node in pink is the title of article Knowledge Management that being visualised. The node in blue is the first level section such as Knowledge Management Technologies and Knowledge Management roles and organisational structure. The node in yellow is the second level section while the node in plain text is the hyperlink. For example, the first level section Key concepts in Knowledge Management contains three subsections: Dimensions of knowledge, Adhoc knowledge access and Knowledge access stages. The directed edge is interpreted as the “has” semantic. For example, the Knowledge Management article in pink has several sub sections such as Knowledge Management Reasons of Failure or Success and Schools of Thought in Knowledge Management. Each first level section

and second level section has several hyperlinks. For example, the second level section Dimensions of knowledge contains the hyperlinks to database.

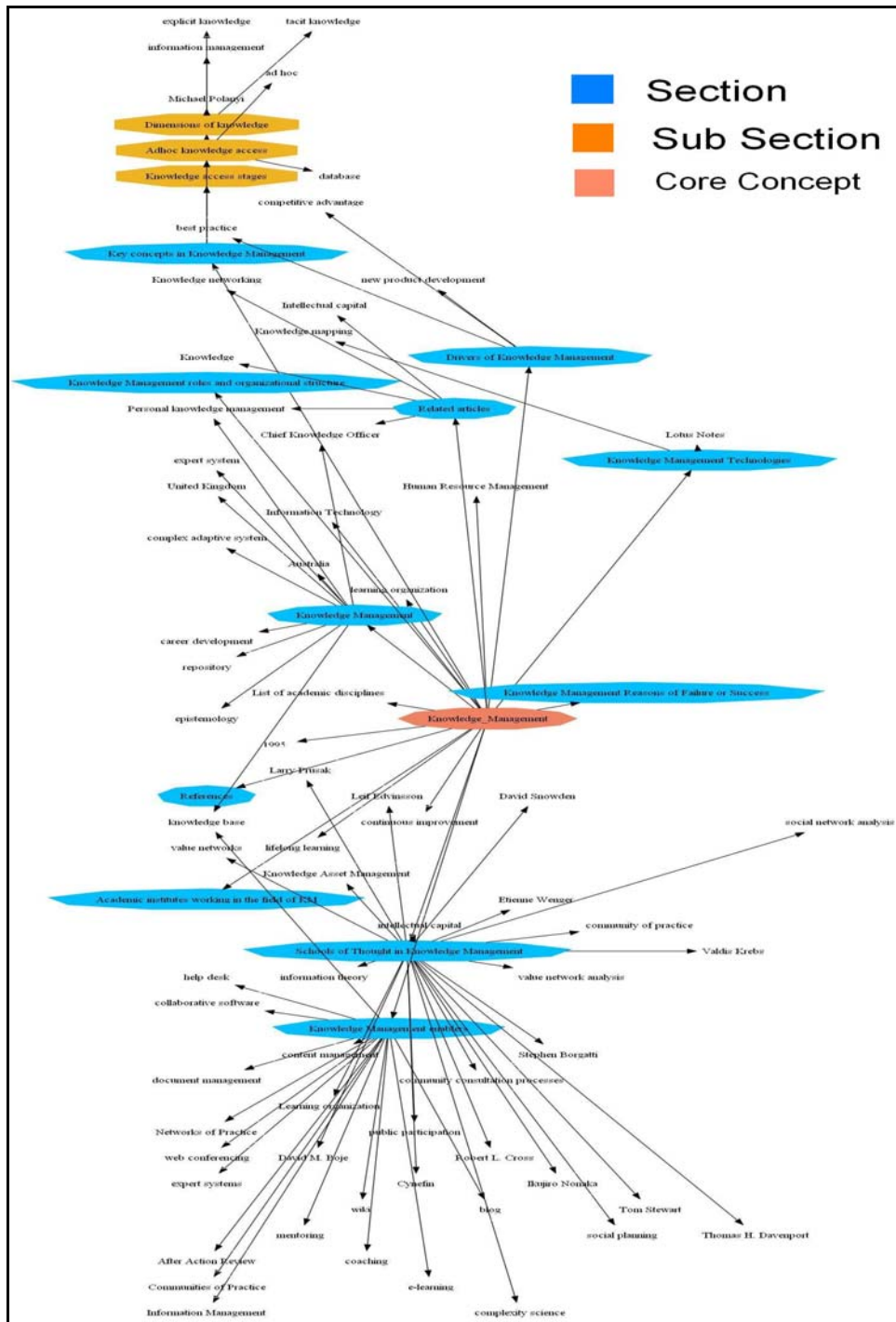


Figure 26 Static Visualisation of “Knowledge Management” on Wikipedia July 17, 2008



<b>Contents</b> [hide]	
1	Knowledge Management
2	Schools of Thought in Knowledge Management
3	Key concepts in Knowledge Management
3.1	Dimensions of knowledge
3.2	Knowledge access stages
3.3	Adhoc knowledge access
4	Drivers of Knowledge Management
5	Knowledge Management Technologies
6	Knowledge Management enablers
7	Knowledge Management roles and organizational structure
8	Knowledge Management Reasons of Failure or Success
9	Academic institutes working in the field of KM
10	Related articles
11	See also
12	References
13	Further reading
13.1	Articles
14	External links

**Figure 27 Table of Contents for Knowledge Management on Wikipedia July 17, 2008**

The static visualisation gives a quick overview on the article by showing major sections and related hyperlinks in some extent. However, it is not particular useful. One of the problems is that it exposes too much detail from the article. For example, the large number of hyperlinks pollutes the diagram, giving little valuable information on the article. Although the diagram shows the major sections within the article, that piece of information can be clearly gained by browsing the table of contents shown in Figure 27. Due to the limitation of static visualisation, the diagram does not have the capability of showing the difference between revisions from the history perspective. Nor can it track the focus of community during a certain period of time. The reason is that static visualisation is merely a snapshot of the object. On the other hand, modern organisation attempts to collect every kind of information available. However, the organisational capacity for producing information far exceeds the human capacity for processing it (Shenk 1997). Organisations are drowning in data, but starving for knowledge. The visualisation is a perfect candidate for reducing information overload and mining valuable knowledge from large volume of data archived by organisations. As will see in chapter 6, dynamic visualisation is a good candidate to present time series complex data.

## ***5.5 Conclusion***

This chapter addressed the problems with Wikipedia and introduced the requirements of visualisation for Wikipedia. It demonstrated that the Knowledge Management article on Wikipedia is a good test bed for the visualisation task. The chapter then described the process for static visualisation from article retrieving, XML and wiki notation parsing to content visualising. Finally, the chapter showed the result of static visualisation and concluded that it is not particular useful for the purpose of tracking content change for Wikipedia articles.

## **6 DYNAMIC VISUALISATION OF CONTENT CHANGE IN WIKIPEDIA**

### ***6.1 Introduction***

This chapter begins by reviewing the requirements of dynamic visualisation for Wikipedia. It assesses the weakness of static visualisation and shows how the dynamic visualisation can help to achieve the goal of visualisation. It then discusses the tool developed to create the dynamic visualisation as well as the process of visualisation. Finally the chapter presents output of the dynamic visualisation for Knowledge Management article on Wikipedia. The chapter discusses the evaluation result of the dynamic visualisation collected through an online survey tool.

### ***6.2 Requirements for Dynamic Visualisation***

Wikipedia articles are created by online community, which may not reflect the true views of the discipline to which article refers. Each article has many revisions and it is difficult to get a clear picture of how an article has reach its current state being. As articles are always evolving, which means there is no final version of articles. Thus, to understand how the knowledge is changing within an article requires a dynamic visualisation process.

The problem with the static visualisation is that it exposes too much detail for the article. Static visualisation is not capable of showing the difference between revisions from history perspective view. As static visualisation is merely a snapshot of the object, it cannot track community focus for a certain period of time.

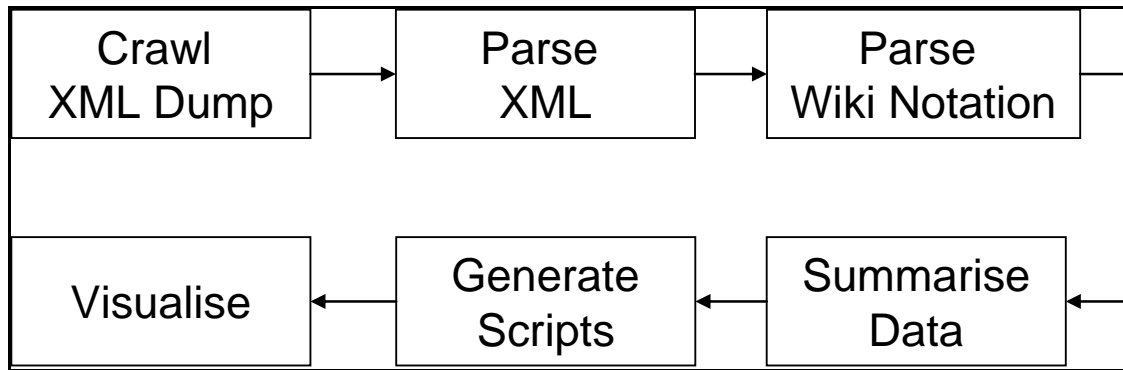
One the other hand, the dynamic visualisation is suitable for showing time series complex data in an interactive manner. The dynamic visualisation developed in this research is able to show the growing path of articles on Wikipedia as well as the community focus from time to time. It helps understand how knowledge creation and sharing can be improved in Wikipedia.

Various metrics could be used for the purpose of visualisation such as word count and contribution count for each section, number of distinct contributors, and authority of authors. As this project is to track content change of articles, metrics related to the article itself is preferred such as word count and contribution count for each section. The data for those two metrics are easy to collect and to explain. In addition, it is novel to track the content change based on sections, which are essential elements to comprise an article. Two metrics – word count and contribution count – together with time series are used to historically and visually evaluate and observe the phenomena of Wikipedia contribution. For example, by observing number of words for sections, one is able to tell appearing and disappearing of sections. This will give a clear picture of the community focus on particular sections during a period of time.

### ***6.3 Visualisation Process and Supporting Technical Architecture***

The dynamic visualisation of content change for Wikipedia is based on the data related to the evolution of article, particularly the evolution of the sections within article.

Figure 28 shows the process for creating the dynamic visualisation. The first stage is to extract data from XML dump. It uses the same tool as the static visualisation process for retrieving XML file and parsing XML and wiki notation. However, the dynamic visualisation stores all the data into MySQL (1998) database including the meta-data on revisions. It employs java-wikipedia-parser (Steven 2007) for counting number of words within each section. The second stage is to summarise the data stored in MySQL database by an analysis tool developed in house in this research project. The last step is to take the output of analysis tool to generate the Motion Visualisation (2007) script embedded in a HTML web page. By sending the visualisation script to Google visualisation server, a flash based interactive visualisation widget is returned.



**Figure 28 Process of Dynamic Visualisation for Wikipedia**

In the following sections, it will be described in detail for each part of the process to create the dynamic visualisation and the technical architecture developed.

#### 6.3.1.1 Parsing Wikipedia Notation for Dynamic Visualisation

As discussed in section 5.4.2.3, static visualisation uses java-wikipedia-parser (Steven 2007) for retrieving relationship between sections and hyperlinks. In dynamic visualisation, the parser is mainly for counting number of words for each section. As the user can organise the article by applying different level of section heading syntax, the structure of the article is naturally formed in a hierarchy structure. When counting number of words for different level of sections, the hierarchy structure is reflected. That is, the number of words in higher level section is the sum of lower level sections. For example, Table 9 shows the fragment of hierarchy structure for Knowledge Management on August 6, 2008. The three sub sections Dimensions of Knowledge, Knowledge access stages and Adhoc knowledge access contain 391, 251, 177 words respectively. As a result, the number of words for higher level section Key concepts in Knowledge Management will be the sum of word count for lower sections 819. When counting the words, the section is simply split into words by a single space delimiter. Although this counting method is not as accurate as other word processor package, the technique tolerable for this research project.

---

/\* other sections\*/

### 3 Key concepts in Knowledge Management

3.1 Dimensions of knowledge

3.2 Knowledge access stages

3.3 Adhoc knowledge access

/\* other sections \*/

---

**Table 9 Example of Hierarchy Structure of Wikipedia Article (source: “Knowledge Management” from Wikipedia.org on August 6, 2008)**

#### 6.3.1.2 Loading Data Into Database

In order to effectively summarise the data for dynamic visualisation, the XML dump is loaded into MySQL (1998) database with the schema shown in Table 10. The Title field is title of the interested article. The Timestamp field is when the revision is created in date time format. The Comment field is high of interest as it contains important information about section editing. When a revision is type of section editing, the heading of the section is encoded into the comment tag. The Notation field is the text of the article in wiki syntax, which will be parsed by Wikipedia parser as discussed in section 5.4.2.3.

---

#### **Schema Basic**

Fields	Mapping to XML Dump	Example
Title	article title	Knowledge Management
Timestamp	tag <timestamp>	2002-08-17T22:14:20Z
Contributor	tag <contributor>	Pichai Asokan (3509)
Comment	tag <comment>	Created the page
Notation	tag <text>	Knowledge Management is a term associated with the processes for the creation, dissemination and utilization of knowledge.

---

**Table 10 Database Schema Basic for Importing XML Dump**

As long as the raw data is loaded into database, one can enjoy the simplicity of data manipulation with SQL statements and high performance of data processing provided by database management system.

When loading XML dump into database with SAX, small modifications are applied to <comment> and <text> tags. The text within some of the <comment> tags is wrapped with C-Style comment (/\*\*/) such as:

```
<comment>/** Schools of Thought in Knowledge Management */</comment>
```

This type of comment indicates that the revision applies to a particular section, such as “Schools of Thought in Knowledge Management” for the above example. The section being revised is extracted from the C-Style comment. For example, the output of the comment tag for the above example is

```
<comment>Schools of Thought in Knowledge Management</comment>
```

### 6.3.1.3 Creating Data Summary To Support Required Analysis

The research is interested in the evolution of articles, especially the section evolution such as number of words within each section and the number of revisions during a period of time. The detail revisions need summarising for the purpose of visualisation. Table 11 shows the database schema for section summarising.

<b>Schema Section</b>		
<b>Fields</b>	<b>Description</b>	<b>Example</b>
Title	The article title	Knowledge Management
Period	Year month	2008-07
Section	1 <sup>st</sup> and 2 <sup>nd</sup> level heading of an article	Drivers of Knowledge Management
NumContribution	Contribution count for section	10
NumWord	Word count for section	200

**Table 11 Database Schema Section for Summarising Section Details**

The Period field of the schema records the year and month information for a particular section. It is decided that the summarisation be on a monthly base because it makes little sense to summarise the revision on a daily base, which only contains few observations. On the contrary, a yearly base summarisation will lead to the loss of information, making it difficult to see the section evolution. The data source of Section field is mapped to the Comment field shown in Table 6. The NumContribution field is the number of times that a section is revised within a month. The NumWord field is the number of words that a particular section contains. As there are multiple revisions of articles during a period, the latest revision in that period is chosen for word counting.

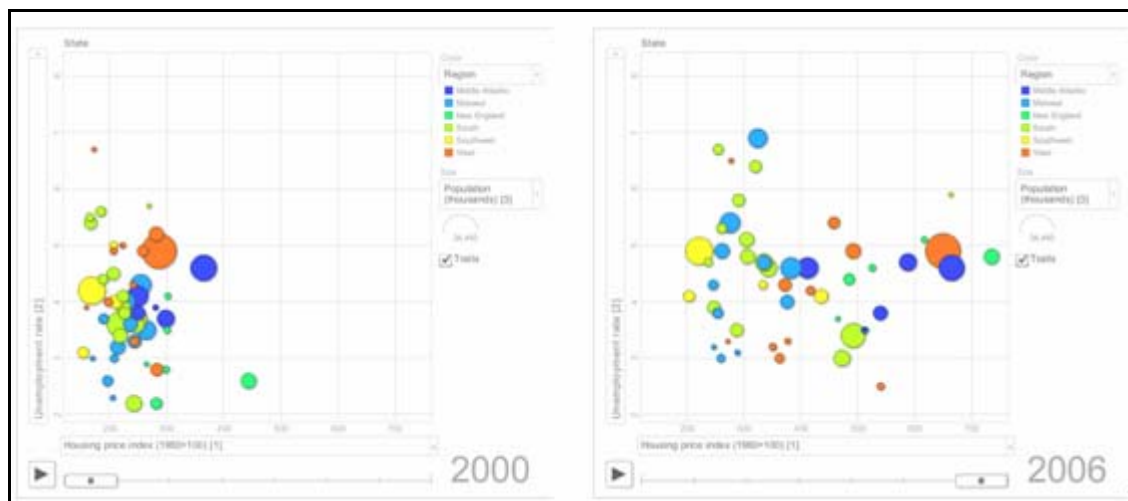
A number of problems were encountered in creating the data summary. When summarising the section editing activities, the diversity of comments brings a number of difficulties. One of the difficulties is that not all the comments describe which section is on revision. For example, the comment for Knowledge Management at 2002-08-17T22:14:20Z is “Created the page”. The solution is that the section retrieved from the comment is matched with the sections in the article and only the matched sections will be picked for visualisation. Another problem is that while a section is semantic identical, it has a variety of form such as capitals, spaces. For example, the “Criticisms of KM - control versus creativity” and “Criticisms of KM - Control versus creativity” are semantic equal. However, they are not treated as the same section as the word “Control” within the section heading differs from one to another. The work around is to trail all the spaces and convert all the letters to lower case.

## ***6.4 Dynamic Visualisation***

### **6.4.1 Introduction to Google Motion Visualisation API**

As discussed in section 4.4, Gapminder is a piece of software for animation of statistics. In March 2006 Google acquired Gapminder (2007) and the team of developers that worked for Gapminder has joined Google since April 2007. As a result, Google publish the Motion Visualisation API (2007). Motion is a dynamic flash based chart to explore several indicators over time.



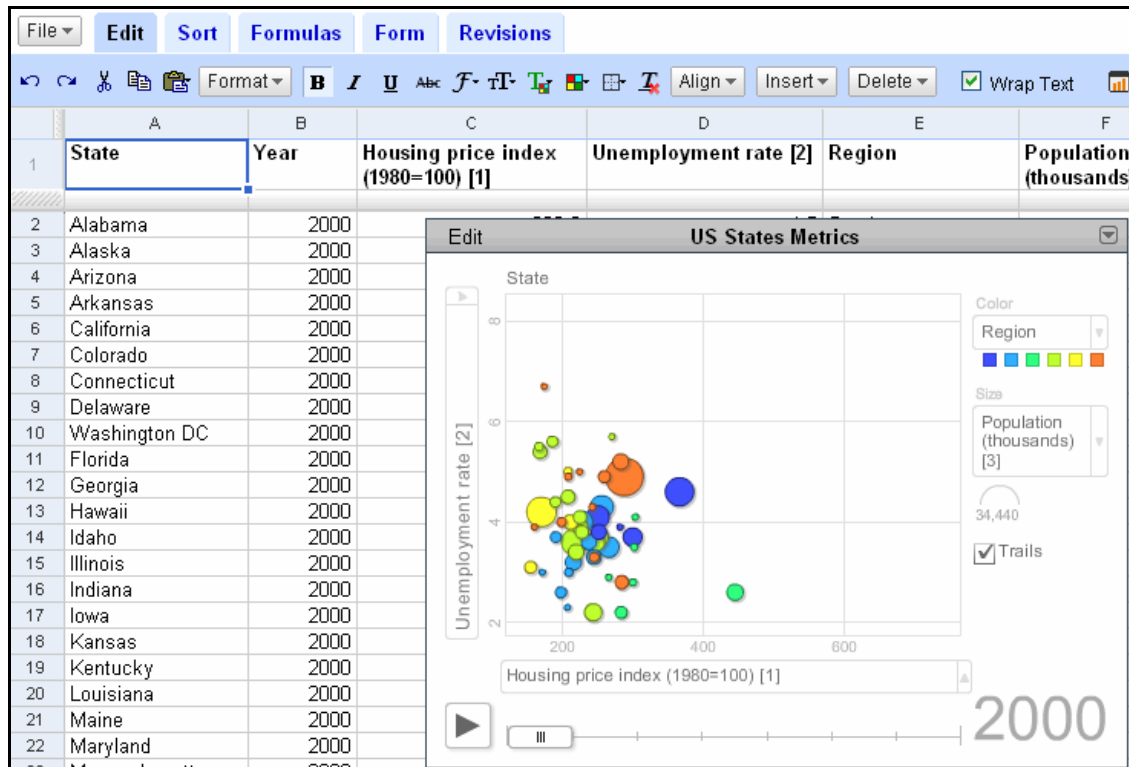


**Figure 29 American Economic From Year 2000 to 2006**

Derived from excellent characteristics of Gapminder, Motion is able to simultaneously display historical data in four dimensional. As shown in Figure 29, each bubble represents a state in US. The X-axis is the housing price index while the Y-axis is the unemployment rate. Bubbles are painted with different colors according to the region of the states. The size reflects the population of the state: larger bubble indicates bigger population. The right bottom shows the corresponding date time (year in this case) for the states. When clicks the “Play” button shown in the left bottom, bubbles start to move around as time goes by. On the right side of Figure 29, it shows the position of bubbles when the animation stops in 2006.

#### 6.4.2 Wikipedia Visualisation Using Google Motion API

There are two alternatives to visualise data with Motion API. One of the choices is to enter data in Google Spreadsheet (Siegle 2007) and insert visualisation widget as shown in Figure 30.



**Figure 30 Visualise Data with Google Spreadsheet**

Another choice is to take advantage of the visualisation service in a web based manner by embedding Javascript in a static HTML web page. Table 12 shows the content of an ordinary HTML page with embedded Javascript code. In the script, one of the functionality is to create data for visualising. This consists of defining columns and corresponding column and filling rows of data. Another functionality of the code is to call the Motion widget API to render the data. As a result, a Flash based visualisation chart is returned and rendered in the browser as shown in Figure 31.

---

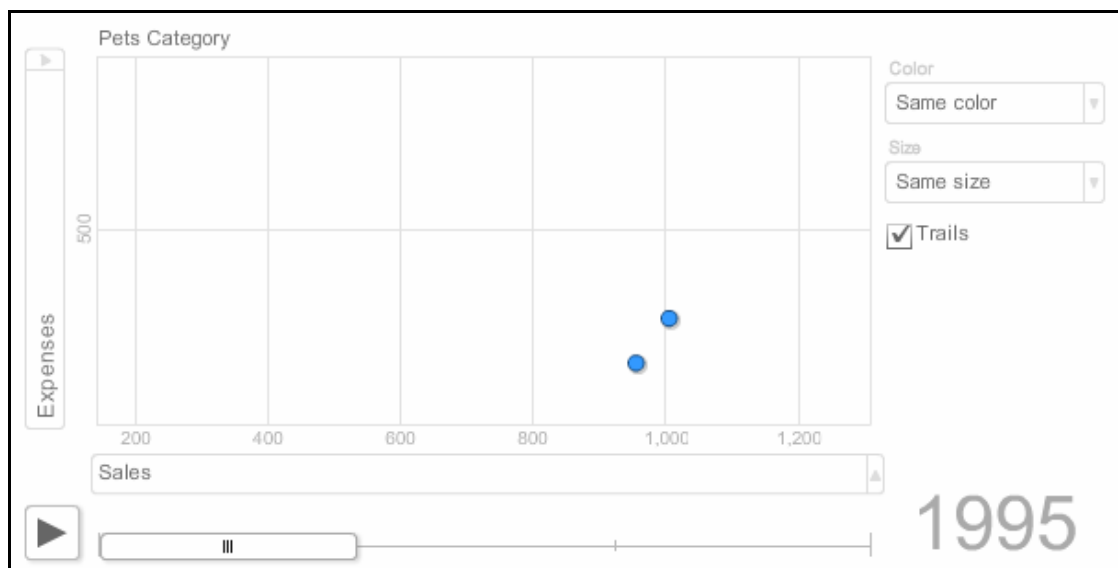
```

<html>
<head>
  <script type="text/javascript" src="http://www.google.com/jsapi"></script>
  <script type="text/javascript">
    google.load("visualisation", "1", {packages:["motionchart"]});
    google.setOnLoadCallback(drawChart);
    function drawChart() {
      var data = new google.visualization.DataTable();
      data.addRows(6);
      data.addColumn('string', 'Department');
      data.addColumn('number', 'Year');
      data.addColumn('number', 'Sales');
      data.addColumn('number', 'Expenses');
      data.setValue(0, 0, 'Dogs');
      data.setValue(0, 1, 1995);
      data.setValue(0, 2, 1000);
      data.setValue(0, 3, 300);
      data.setValue(1, 0, 'Cats');
      data.setValue(1, 1, 1995);
      data.setValue(1, 2, 950);
      data.setValue(1, 3, 200);
      <!-- Other data setting statements -->
      var chart =
        new google.visualization.MotionChart(document.getElementById('chart_div'));
      chart.draw(data, { width: 600, height:300});
    }
  </script>
</head>
<body><div id="chart_div" style="width: 600px; height: 300px;"></div></body>
</html>

```

---

**Table 12 Embedded Javascript Code for Generating Motion Visualisation Widget**



**Figure 31 Motion Widget Generated with Javascript**

As the web based visualisation is easy for publishing and has little couple with external application such as Google Spreadsheet, it is used for the dynamic visualisation in this research project.

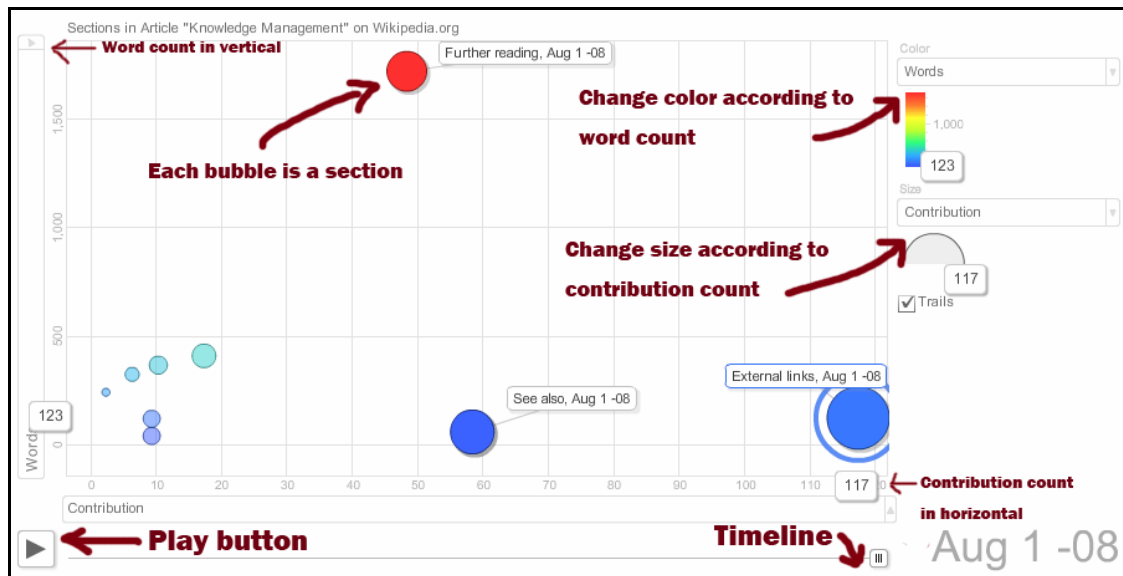
A tool is developed to dynamically generate the HTML page, which in turn generates the visualisation widget. The tool reads the data defined in Table 11 and output four columns shown in Table 13. The Section column is the article section being observed. The YearMonth column is the time period in month. The Contribution column is the total number of revision up to that particular period. For example, if there are 3 revisions up to June 2004 for the section “The development of KM” and there are 4 more revisions in July 2004, the Contribution value for July 2004 is the accumulated number of revisions 7.

Column	Type	Description
Section	string	The interested section
YearMonth	Date	Period of the data in year and month
Contribution	numeric	Number of revision for the section
Words	numeric	Number of words within the section

**Table 13 Columns for Wikipedia Dynamic Visualisation**

## 6.5 The Result

Figure 32 shows the snapshot for the dynamic visualisation, which uses word and contribution count metrics for visualising. Each bubble on the grid is a section within the article. The X-axis is accumulated the number of contributions up to date and the Y-axis is the number of words at a particular time. One can change the colour for the bubble based on the number of words appearing and the size based on the contributions to that particular section. When playing the visualisation, one can get a picture of how the contributions are increasing decreasing along the horizontal axis and number of words increasing decreasing along the vertical axis. One can investigate particular bubbles by simply click on them. When mouse over particular bubble, the title of the section is shown nearby and one can see the number of contributions shown on the X-axis horizontal and number of words shown on the Y-axis vertical. For example, the section “External links” has 123 words and 117 contributions in August 2008.

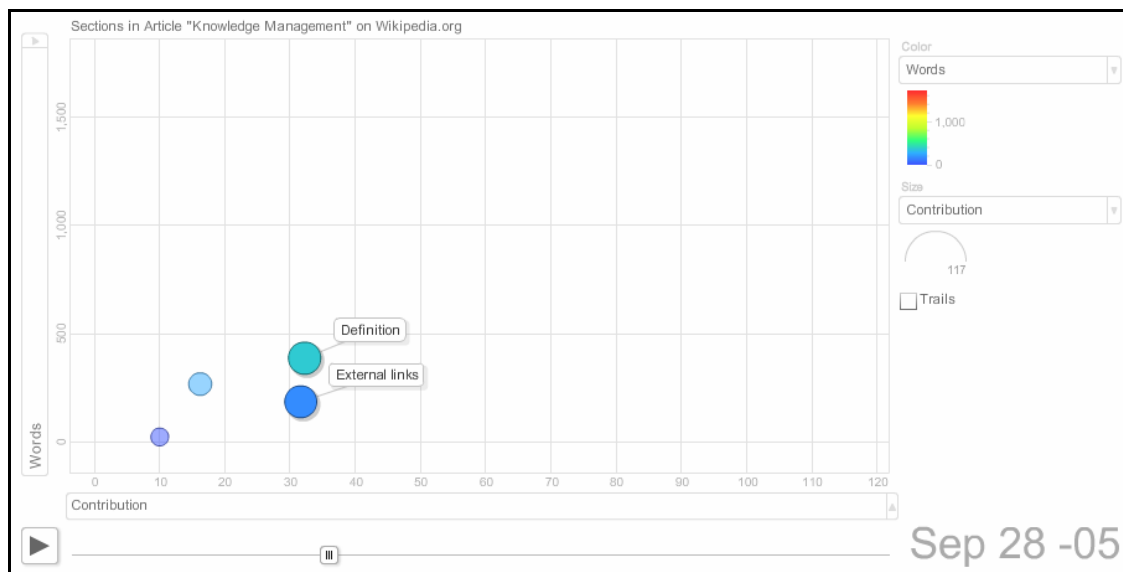


**Figure 32 Dynamic Visualisation for Knowledge Management Article in Wikipedia Using Word and Contribution Count Metrics. (Revisions from August 2002 to August 2008)**

It can be seen that at the end of the visualisation August 2008, contributions to “External links” and “See also” are moving quite consistently along the horizontal axis but are not particular moving along the vertical axis. In contrast, “Further reading” starts off as a quite insignificant bubble. However, by the end it contains more than ten

times words than “External links” but has less than half of the contributions. This means many people contributed little content to “External links” while few people contributed many words to “Further reading”.


The dynamic visualisation helps as a tool for exploration and knowledge creation and sharing. By replaying the timeline, one can investigate the relationships between bubbles. Start on July 2004, one can choose the “Definition” bubble and track what happens to that particular section as well as the “External links”. By moving the slide on, it can be seen that while the “Definition” is increasing, the “External links” is also moving towards the same direction by the end of September 2005 shown in Figure 33. But all the sudden, “Definition” disappears. It could be concluded that while people were contributing to the definition of Knowledge Management, they were also contributing to external links to support their definition. The fact that “Definition” disappeared in 2005 triggers to see whether the community changed content by moving the definition into other sections of the article.



**Figure 33 Investigation on Definition and External links with the Dynamic Visualisation Tool**

The movement of content can be verified by comparing the two revisions between 07:53 and 14:45, 3 November 2005, shown in Figure 34. The shared text between two revisions is highlighted in red rectangle. As can be seen, some content of the definition in the earlier revision was moved to the later revision.

**Definition**

A widely accepted 'working definition' of knowledge management applied in worldwide organizations is available from the [WWW Virtual Library Knowledge Management](#) .

"Knowledge management caters to the critical issues of organizational **adaptation**, **survival**, and **competence** in face of increasingly discontinuous environmental change.... Essentially, it embodies organizational processes that seek synergistic combination of data and **information processing** capacity of information technologies, and the creative and innovative capacity of **human beings**."

In simpler terms, knowledge management seeks to make the best use of the knowledge that is available to an organization, creating new knowledge, increasing awareness and understanding in the process.

Knowledge Management can also be defined as:

Capturing, organizing, and storing knowledge and experiences of individual workers and groups within an organization and making this information available to others in the organization.

Knowledge management is most frequently associated with two types of activities:

1. To document and appropriate individuals' knowledge and then disseminate it through such venues as a companywide database.
2. Activities that facilitate human exchanges using such tools as groupware, email, and the Internet.

**(a) Fragment of Knowledge Management Definition at 07:53, 3 November 2005**

**Definitions of 'Knowledge Management'**

- **Organizational Knowledge management (KM)** is the creation, organization, sharing and flow of **knowledge** in organizations.
- Knowledge Management seeks to make the best use of the knowledge that is available to an organization, creating new knowledge, increasing awareness and understanding in the process.
- Knowledge Management can also be defined as the capturing, organizing, and storing of knowledge and experiences of individual workers and groups within an organization and making this information available to others in the organization.

**(b) Knowledge Management Definition at 14:45, 3 November 2005**

**Figure 34 Comparison of Knowledge Management Definition between 07:53, 3 November 2005 and 14:45, 3 November 2005 Revisions**

**6.6 Survey and Evaluation**

The survey conducted as part of this project assesses the usefulness of a visual representation of article history as a tool for Wikipedia users as well as the usefulness of this representation as a tool to improve knowledge creation and sharing in Wikipedia. The survey is quite short and qualitative in nature, it attempts to collect users views rather than focusing on gathering statistical data.

The survey addressed a number of issues. Firstly it attempted to elicit the respondents' level of experience with Wikipedia so that their views on the tool could be analysed more deeply. Questions covered respondent usage of Wikipedia e.g. some users may only read articles on Wikipedia without any contribution while others contribute a lot. Further respondent views on the usefulness, reliability and quality of articles on Wikipedia were sought.

The survey then addressed respondents' views on the areas of specific interest to this dissertation: article history and visualisation. Respondents were polled on their views on the importance and usefulness of article histories on Wikipedia and the potential usefulness of visually tracking history of Knowledge Management article on Wikipedia. The survey finished by querying the extensibility and applicability of the visualisation tool to other articles on Wikipedia.

The visualisation tool was made available on a website for public evaluation as shown in Figure 35. The page gives a brief introduction to the visualisation tool and a link to Knowledge Management article on Wikipedia. The visualisation of word count and contribution count metrics for the article on Knowledge Management are available for evaluation. Finally, a link to the online survey created with SurveyMonkey (<http://www.surveymonkey.com>) is provided to allow the user to give feedback.

## Knowledge Visualisation of Wikipedia with Google Motion Visualisation API

This page relates to an MSc project to visualise the evolution of articles in Wikipedia.

By using Google [Motion](#) Visualisation API, the visualisation works with two simple metrics - word count and contributions count.

The sample visualisation was constructed using the [Knowledge Management](#) on Wikipedia. Before viewing the visualisations you might find it useful to read this article online.

A presentation about the dissertation is available in two parts as videos on YouTube.com. [Part 1](#) describes the premise of this dissertation, [Part 2](#) explains the dynamic visualisations of the Knowledge Management article which available on this page.

Once you have finished viewing the visualisations, it would be appreciated if you could take a few moments of your time to complete a survey which will contribute to the evaluation of the work in this MSc Project.

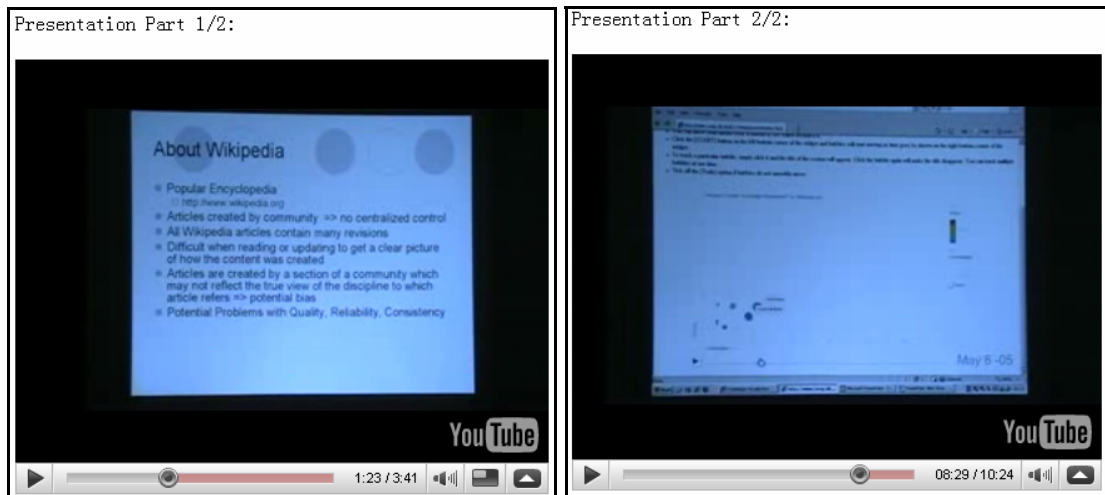
- [Start visualise with the word count metrics for the article Knowledge Management.](#)
- [Start visualise with the contribution count and word count metrics metrics for the article Knowledge Management.](#)
- [Complete the Survey on Knowledge visualisation of Wikipedia.](#)

**Figure 35 Main Page for Demonstrating Visualisation Tool**

In order to explain the purpose and usage of the tool, two videos are shot and published on YouTube (2005) for evaluation shown in Figure 36. The first part of the video Figure 36 (left) describes the premise of this dissertation while the second part of



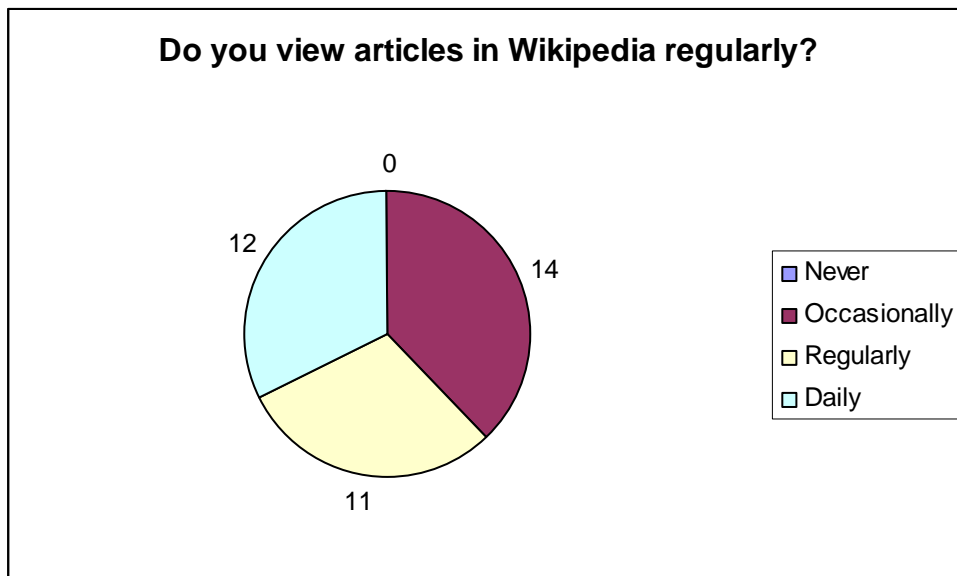
the video Figure 36 (right) explains the dynamic visualisations of the Knowledge Management article on Wikipedia. The first part of the videos reached No.9 in the top views in the education section on YouTube Ireland on August 23, 2008 after the videos were published for one day. By the end of this research, the first part of the videos has attracted 225 views after two weeks publishing. The result is promising that people showed great interest in the visualisation tool by viewing the videos.



**Figure 36 Videos for Explaining the Visualisation Tools**

The survey was broadcasted to a variety of audience for feedback collection. This includes computer science lecturers and students in Irish universities, researchers interested in Wikipedia as well as the postgraduate students in Knowledge Management course in Dublin Institute of Technology, Ireland. The survey was also posted on a variety of Knowledge Management boards for attractiveness. In the end, 37 responses were collected from the online survey.

Figure 37 below illustrates the survey responses to the question of how regularly users view articles on Wikipedia. It can be seen that Wikipedia was popular with respondents and widely accepted as all the respondents reported their experience in viewing Wikipedia articles. Almost two thirds of the respondents view articles daily or regularly, while one third occasionally view the articles. Furthermore, all of those respondents find the articles useful on Wikipedia in general.



**Figure 37 Percentage of Viewing Articles in Wikipedia**

Figure 38 shows the answers to how reliable does a respondent find the content of articles material on Wikipedia. It can be seen that more than half of the respondents find the material generally reliable. More than 13% of the respondents think the reliability of content depends on the author while there is one respondent thought it depends on the topic. 10.8% think the material is very unreliable while 8.1% think the material is sometimes reliable. There is one respondent reported that the material is very reliable. The more a user reads, the more chance he or she can evaluate the reliability of content. It can be deduced that it is those two thirds of the respondents, who read articles regularly and daily, that find the article content generally reliable. It could be concluded that content of articles on Wikipedia is generally reliable.

<b>How reliable do you find the content of articles material on Wikipedia?</b>		
<b>Answer Options</b>	<b>Response Percent</b>	<b>Response Count</b>
Very Unreliable	10.8%	4
Sometimes Reliable	8.1%	3
Generally Reliable	62.2%	23
Depends on the authors	13.5%	5
Other (please specify)	5.4%	2

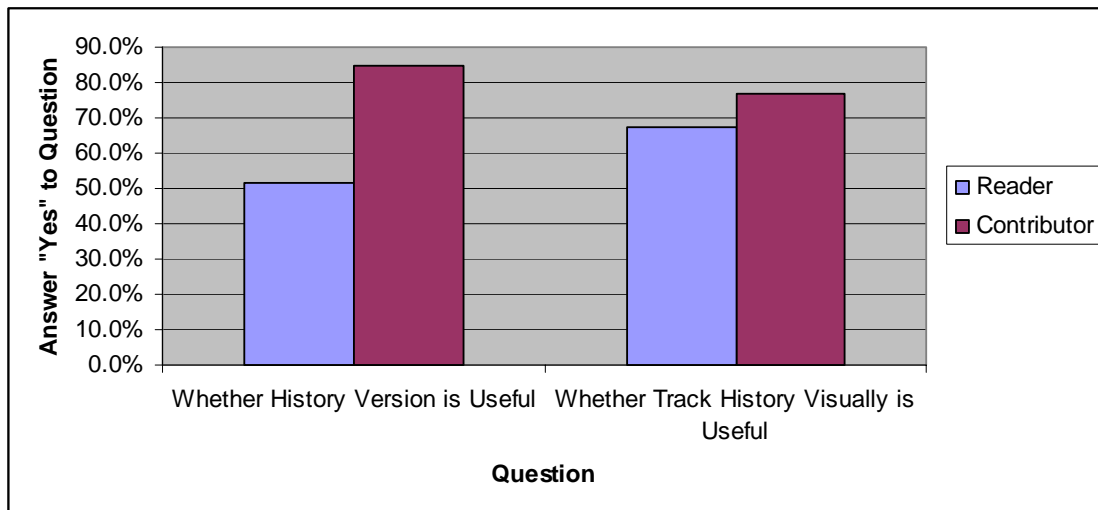
**Figure 38 Percentage of Reliability of Article Content on Wikipedia**

Similar results can be seen when asking the quality of articles on Wikipedia shown in Figure 39. It is surprising no negative feedback on the quality of Wikipedia articles is returned. More than 67% think the quality is good or very good. 8.1% report the quality is excellent while 18.9% think it depends on the authors. One respondent thinks the quality depends on the topic while another one thinks it depends on the popularity of articles – the more popular an article is, the higher quality it achieves.

How do you find the quality of articles on Wikipedia?		
Answer Options	Response Percent	Response Count
Very Poor	0.0%	0
Poor	0.0%	0
Good	40.5%	15
Very Good	27.0%	10
Excellent	8.1%	3
Depends on the authors	18.9%	7
Other (please specify)	5.4%	2

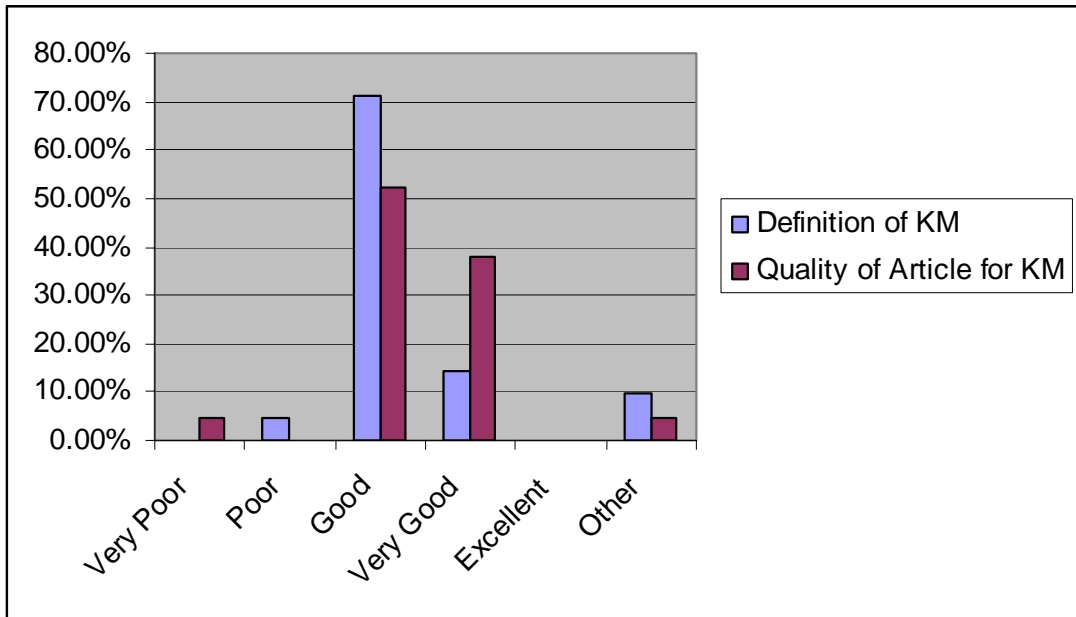
**Figure 39 Percentage of Quality of Articles on Wikipedia**

The next series of questions focus on the usefulness of article histories on Wikipedia from a reader perspective view. 51.4% have read the history version of articles while 48.6% have never read the history versions. 14 respondents (37.8%) confirmed to have contributed to Wikipedia content. Compared to readers of Wikipedia, contributors regarded the history revisions of articles more valuable shown in Figure 40. Most of the contributors (84.6%) think the history versions of articles are as useful a resource as the most updated version when considering updates to content while about half of the readers (51.4%) agreed. 76.9% contributors think a mechanism to track the history of Wikipedia articles visually is useful while 67.6% readers agreed. There is a gap on the usefulness of history between Wikipedia contributors and readers. It could be deduced that in order to contribute, contributors need to understand both current and past community views on the concept. They need to see the evolution of articles so that they could bring some content back from the history, support and backup their contribution with the history revisions.



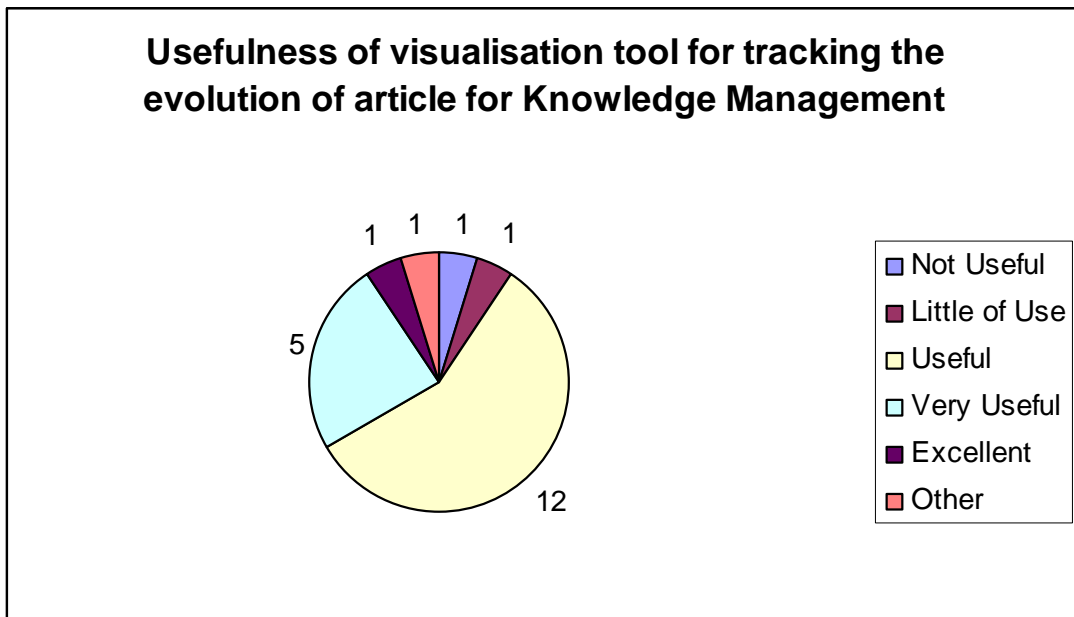
**Figure 40 Comparison of Usefulness of History between Reader and Contributor**

The next series questions focus on the quality of Knowledge Management article on Wikipedia. 24 respondents (66.7%) have read the Knowledge Management article on Wikipedia. Within those readers, 81% think the Knowledge Management article on Wikipedia has a neutral point of view. Figure 41 shows the responses to the quality of Knowledge Management article on Wikipedia. As can be seen that most respondents think the definition of Knowledge Management is good or very good, with 71.4% and 14.3% respectively. More than 90% respondents think the quality of article is acceptable, with 52.4% voted for good and 38.1% for very good. It could be deduced that as the article is well organised and substantial, people give positive feedback on it. However, the article is not distinguishing enough to gain excellent voting. This can be verified in that the Knowledge Management article had not become “featured” yet by the end of this research.



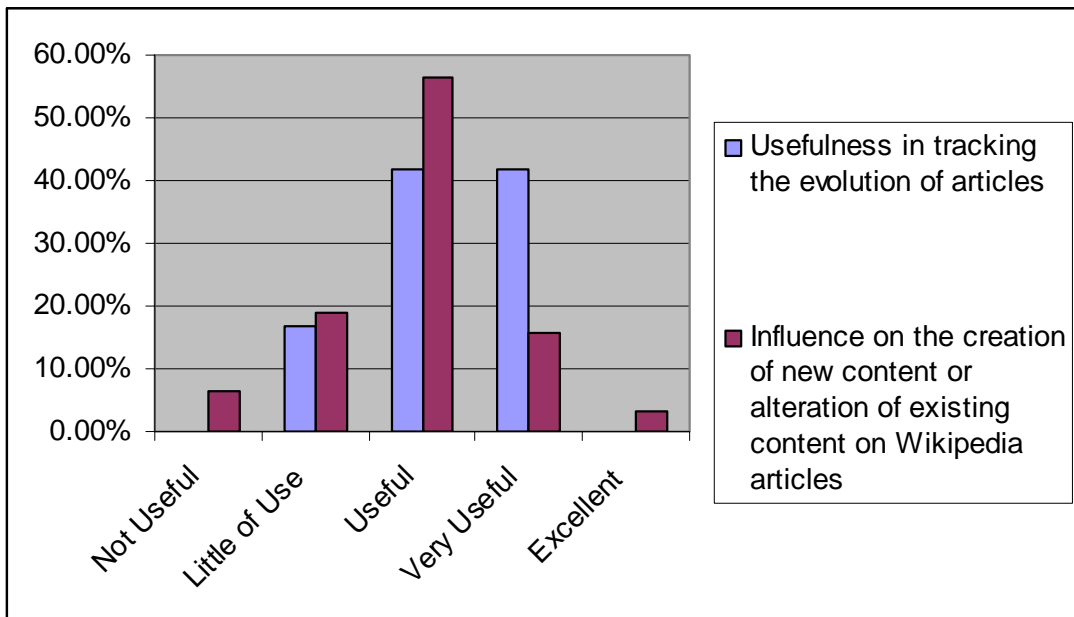
**Figure 41 Response to Quality of Knowledge Management Article on Wikipedia**

The next series of questions focuses on the usefulness of the visual representation for tracking the content change. 57.1% respondents, who have viewed Knowledge Management article, gained better understanding of the topic after seeing the evolution. Figure 42 shows the responses to usefulness of visualisation tool for tracking evolution of article for Knowledge Management. 12 respondents (57.1%) think the tool is useful for tracking evolution for Knowledge Management article while 5 respondents (23.8%) think it is very useful.



**Figure 42 Responses to Usefulness of Visualisation Tool for Tracking Evolution of Article for Knowledge Management**

9 out of 12 respondents (75%) think it would contribute to better understanding on topics covered on Wikipedia after viewing the visualisation tool for the particular Knowledge Management article. Figure 43 shows the response to the extensibility of the visualisation tool to other articles. 41.7% think it is useful in tracking the evolution of articles in Wikipedia while another 41.7% think it is very useful. On the other hand, 56.3% think the tool is useful for influencing the creation of new content or alteration of existing content on Wikipedia articles while 15.6% think it is very useful. However, around 20% think the tool is little of use for either tracking the evolution of articles or influencing on the creation of new content or alteration of existing content on Wikipedia articles. While there is no more than 10% think the tool is useless for tracking the evolution of article for the Knowledge Management article, there are more people (around 20%) think the tool would not be useful for tracking other articles on Wikipedia. One of the explanations is that as the visualisation for other articles are not available; people show little confidence before they actually see the tools. Another possible reason is that readers rather than the contributors voted useless of the visualisation tool as they show less interest in the article histories than the current version of articles.



**Figure 43 Response to the Extensibility of the Visualisation Tool**

14 out of 37 comments are collected during the survey. One feedback was the ease of use for casual users. While the dynamic visualisation gives rich information in terms of time, word count and contribution count, one responder thought it might be too complex for casual users. The tool is definitely interesting for detailed research on how an article evolved. For the casual user, more simple visualisations are required such as showing on one image on how often the article was edited, which sections are most active. That's because the Wikipedia users have specific information needed and they need to find it quickly, so any other extra meta-information such as article history is not very interesting to them.

Another feedback is on more integration with article content. It is suggested that the tool can be improved by showing key words/points from historical revisions in a quickly viewable manner (expandable as the user wishes to hone in on certain points and consider re-establishing them into current articles). A ratings system could be added so that old data can be rated in terms of importance from multiple users so that vandalism is quickly and easily separated from data that is potentially useful or important but absent due to bias or personal opinion or misunderstanding. In addition, displaying the rearrangement of contents within a certain article could add great value to the visualisation.

Other feedback varies from one to another. It is suggested that in addition to visualise the number of words and contributions by section, it would be interesting to view a history of changes by contributor. The requirement of visualisation on any Wikipedia article is also proposed in the feedback. The dynamic visualisation could be useful for new students of the MSc Knowledge Management course as an assignment to track the changes and relate the changes to something that occurred around that time.

### ***6.7 Conclusion***

This chapter discussed the requirements of Wikipedia visualisation and how dynamic visualisation can address the weakness and problems in static visualisation. It detailed the process of dynamic visualisation from parsing wiki notation, summarising and visualising with Google Motion API. It then presented the result of visualisation and concluded that the tool is useful to improve knowledge sharing and creation in Wikipedia.



## **7 CONCLUSIONS AND EVALUATION**

### ***7.1 Introduction***

The final chapter of this dissertation presents the conclusions and recommendations formed from performing this research project. The aim of the research was to investigate the usefulness of a visual representation of article history as a tool for Wikipedia users as well as the usefulness of this representation as a tool to improve knowledge creation and sharing in Wikipedia. The dynamic visualisation is built on the Knowledge Management article on Wikipedia as test bed with the word count and contribution count metrics. This chapter summarise the dissertation by describing how the research aims and objectives were achieved. The chapter discusses the contributions to the body of knowledge in this research as well as any limitations to experimentation or evaluation within the research project. The chapter also discusses the potential areas for future research.

### ***7.2 Research Definition & Research Overview***

Wikipedia is a successful and popular web-based platform of collaborative content creation of bodies of knowledge. Wikipedia is composed of encyclopaedias in different languages; each of them covers a very wide range of knowledge, from arts to biography, from geography to history, from mathematics to science and from society to technology. Anyone can contribute to Wikipedia and articles are created and revised by communities with shared interest. All the revisions are kept and available for each article. Contributions can be undone by reverting an article to a previous version.

Knowledge Management is a systemic and organisationally specified process for acquiring, organising, and communicating knowledge of employees so that other employees may make use of it to be more effective and productive in their work (Alavi and Leidner (1999)). A knowledge base is a collection of data, information and knowledge with an implied organisation and links to provide navigation among items within the organisation (Knowledge Base 2002). According to this definition, Wikipedia is a knowledge base in that it is a collection of articles with links to other

knowledge. Wikipedia is also a good example of community of practice (CoP) as articles are contributed by a disparate group of individuals, with a shared interest in a topic.

Knowledge visualisation examines the use of visual representations to improve the creation and transfer of knowledge between at least two people. Knowledge visualisation designates all graphic means that can be used to construct and convey complex insights. Knowledge visualisation aims to transfer insights, experiences, attitudes, values, expectations, perspectives, opinions and predictions, and this in a way that enables someone else to re-construct, remember and apply these insights correctly (Eppler & Burkhard 2005).

Although each article in Wikipedia has many revisions, it is difficult to get a clear picture of how an article has reach its current state being. Revision history is hard for processing. Articles are created by a section of a community which may not reflect the true views of the discipline to which article refers. There are potential bias as articles are contributed by a particular section of people who have particular bias about the topic. As articles are always evolving - there is no final version of articles. To understand how the knowledge is changing within an article requires a visualisation process.

The premise of this dissertation is to investigate the usefulness of a visual representation of article history as a tool for Wikipedia users. The dissertation is also to investigate the usefulness of this representation as a tool to improve knowledge creation and sharing in Wikipedia.

The research began by performing a literature review on Wikipedia. In particular, it examined the cooperation, quality, application of Wikipedia. The project then performed a literature review on Knowledge Management and assessed Wikipedia from knowledge management perspective. It then reviewed the literature for both static and dynamic visualisation as well as how knowledge visualisation can help for the spiral of knowledge and knowledge management process. The project addressed the weakness of static visualisation and identified the gap for the current research on Wikipedia visualisation.

The project explored several tools to prepare the data for visualisation. A network file transferring tool cURL (1996) was used to download article revisions from Wikipedia. XML SAX (Harold 2002) parser was used to extract metadata from the revision history while java-wikipedia-parser (Steven 2007) was used to extract article content. The extracted data was stored in MySQL (1998) database for querying and further analysis.

The project then statically visualised the content of articles by encoding sections and hyperlinks for the Knowledge Management article into an image. The GraphViz (Ellson et al. 2002) visualisation tool was employed to build a network diagram by linking sections and hyperlinks together.

Nothing particular was discovered in the static visualisation for the Wikipedia article. The project continued by visualising the content change in word count and contribution count metrics with Motion Visualisation (2007). The dynamic visual representation was then published together with an online survey as well as two videos explaining purpose and usage of the tool for evaluation.

As a result, the following objectives have been achieved in this dissertation:

1. Performed a literature review on Wikipedia from the cooperation, quality and application perspective view. Explored how Wikipedia can be regarded as a knowledge base and an example of community of practice (CoP) from the Knowledge Management perspective.
2. Performed a literature review on knowledge visualisation. Reviewed different formats of knowledge visualisation and how knowledge visualisation can help to transfer insights, experiences, attitudes, values, expectations, perspectives, opinions and predictions in knowledge management.
3. Identified an appropriate tool chain, - file transferring, parsers and visualisation tool - for creating both the static and dynamic visualisation. Developed a tool in house to summarise data for both static and dynamic visualisation.

4. Identified the key aspects of content change in Wikipedia articles for developing the visualisation. Identified the metrics and data of interest for Wikipedia article revisions for the purpose of content change visualisation.
5. Created a static and dynamic series of visualisations for the Knowledge Management article on Wikipedia. Published the result of dynamic visualisation to academic staffs and students as well as Wikipedia users for evaluating the usefulness of the visualisation.
6. Compared and analysed the feedback from the survey on the dynamic visualisation. Evaluated the result and concluded that the visualisation is helpful for tracking the content change in Knowledge Management article on Wikipedia.

### ***7.3 Contributions to the Body of Knowledge***

The research outlined in this dissertation has achieved a number of contributions to the body of knowledge. Firstly an update literature review on Wikipedia, its usage, creation and quality has been presented. Wikipedia is a mirror of society. Articles are created by online community with no centralised control. Wikipedia itself shows the capability of supporting a broader range of structures and activities than other collaborative platforms (Butler et al. 2008). Despite the potential for anarchy, the Wikipedia community places a strong emphasis on group coordination, policy, and process (Viegas 2007). What's more, there is a high level of inequality in the total number of contributions to each Wikipedia language edition, with less than 10% of the total number of authors being responsible for more than 90% of the total number of contributions (Ortega et al. 2008). While there is no centralised control mechanism for Wikipedia, implicit coordination and authority still exist in the community. Contributions are self motivated in the community and it is the flexibility that achieves the success of Wikipedia regarding the huge population in the community.

A finding emerged from the literature survey is the relationship between the processes of constructing Wikipedia articles with their quality. While Wikipedia owes its incredible growth to its openness, it is exactly the same feature that makes the quality

of articles hard to control. It is difficult to distinguish good articles from bad ones due to a number of reasons such as large number of articles for quality judgement, diverse content among articles, unknown contributors and abuse. Thus, there is a variety of research investigating Wikipedia from a variety of perspectives which contribute to considerations of Wikipedia quality. High-quality articles in Wikipedia are distinguished from the rest by a larger number of edits and distinct editors (Wilkinson & Huberman 2007). Pages edited in the very beginning by authors with high reputation have a higher chance to get featured in the future (Stein & Hess 2007). Article length is a very good predictor of whether an article will be featured on Wikipedia (Blumenstock 2008). While all those metrics can measure the quality of articles in some aspects, they cannot give a qualitative view as humans do. For example, it is difficult for a machine to judge whether an article is comprehensive. Nor can a machine determine how accurate the article is. The problem is that the quality of article is largely dependent on readers, who have complex criteria to judge the article in their knowledge context.

Wikipedia articles can be accessed with a web browser. However, the research in this dissertation also considered how Wikipedia can be used as a knowledge resource for integration for building various applications. Milne et al. (2007) use extracted thesauri from Wikipedia to automatically and interactively facilitate query expansion. Banerjee et al. (2007) proposes a method of improving the accuracy of clustering short texts by enriching their representation with additional features from Wikipedia. Wikipedia is also capable of using Wikipedia knowledge to construct a global ontology (Pei et al. 2008). Sinclair et al. (2007) develop a system that extracts information from the free text descriptions and try to identify the respective Wikipedia article describing each entity extracted from the text. Wikipedia is an attractive resource for different research areas as well as for organisations and individual users. In the knowledge management area, ontology can be built upon the collection of articles with relationships connected via hyperlinks. In the information retrieval area, thesauri can be extracted to improve the accuracy of clustering short texts as well as to facilitate query expansion for search engine. Hyperlinks to the Wikipedia articles can be injected to enrich the web pages. Similar items in the feed reader can be clustered with extracted thesauri from Wikipedia to make the information more manageable for a user. Wikipedia deserves more attraction for researchers and organisations for investigation.

Another contribution is assessing the Wikipedia from Knowledge Management perspective. Wikipedia can be treated as an invaluable knowledge base. It is a collection of articles with internal links to other articles within Wikipedia website itself as well as to the external hyperlinks to other web pages and documents. Articles are encoded with Wikipedia notation in both machine readable and human readable format. Meanwhile, Wikipedia shared the characteristic of community of practice. Wikipedia is an example of what can be accomplished by a disparate group of individuals, with a shared interest in a topic, working on such a foundation. In Wikipedia, each article has an association group of people - a community that are interested in contributing their knowledge to the content of article. Contributors engage in joint activities and discussions through "Talk Page" to share knowledge and opinions on the topics. Wikipedia users share repertoire of tools to ensure the quality of Wikipedia article, to add enhanced text processing functions to Wikipedia and to monitor and detect vandalism.

The project also contributes by assessing the usefulness of static and dynamic visualisation for content change. Knowledge visualisation aims to transfer insights, experiences, attitudes, values, expectations, perspectives, opinions and predictions, and this in a way that enables someone else to re-construct, remember and apply these insights correctly (Eppler & Burkhard 2005). A static visualisation allows exploring data by offering different methods such as overview, zooming in and filtering and then showing details on demand to achieve the cognition. On the other hand, dynamic visualisation helps to explore large time-varying datasets with reoccurring data objects that alter in time.

The project built a static visualisation by linking different level sections and hyperlinks together to gain a summarisation of the article content. However, the visualisation exposes too much detail from the article and cannot show the difference between revisions.

The project chose two simple metrics - word count and contribution count - for tracking content change of Wikipedia articles. By retrieving revision dump from Wikipedia website, parsing the article and its metadata, and summarising the data, the

project built a dynamic visualisation for the Knowledge Management article in Wikipedia. The visualisation representation gave the community a valuable insight into the evolution of the Knowledge Management concept. The community gained better understanding on the topic after seeing the article evolution. The community agreed that the visualisation could be applied to other Wikipedia articles to influence the creation of new content or alteration of existing content.

Finally, by publishing the tool to academic staff and students in Knowledge Management area as well as the broader community of Wikipedia users for evaluation, various promising feedback was collected by an online survey. The consensus of this survey is that the tool would primarily be of use to Wikipedia sysops and editors to quickly take a picture of the current structure and evolution over time of a certain article. Visualising the article edit history is very useful and interesting. The tool can help people who are always editing Wikipedia articles, or who is responsible for maintaining Wikipedia will benefit dynamic visualisation tool very much. The dynamic visualisation may encourage people to check previous versions before adding updates which may have pre existed.

Current researches on Wikipedia visualisation are mainly focusing on the meta-data of revision histories, the editing patterns of articles and cooperation patterns. The promising feedback showed that tracking the article content change is an area worth further investigation.

#### ***7.4 Experimentation, Evaluation and Limitation***

The project uses word count and contribution count metrics for dynamic visualisation. However, the metrics are limited to the syntax level, which has a bias based on the assumption the more words a section contains, the better quality it is. However, a user can easily change the meaning of the content by simply reordering the words within sentences. This leads to the requirements of semantic visualisation for content change. For example, one needs to figure out how the content is semantic identical but syntax differential and how to encode the change for processing. Furthermore, it is difficult to figure out the best way for visualisation to see the semantic content change in both

richness and simplicity. The semantic visualisation is a good research line for Wikipedia visualisation.

The dynamic visual representation of content change is then published for evaluation. One of the focuses in the feedback was the ease of use for casual users. While the dynamic visualisation gives rich information in terms of time, word count and contribution count, more simple visualisations are required for casual users. Another focus is the requirement of more integration with article content. For example, displaying the rearrangement of contents within a certain article could add great value to the tool.

### ***7.5 Future Work & Research***

This visualisation project mainly targeted at more advanced users and researchers for tracking content change. For the casual user, the dynamic visualisation could be too complicated for use. This gives the research topic that more simple visualisations are required. Rather than focusing on the dynamic visualisation in this project, the research simple but novel static visualisation is a good research area for Wikipedia visualisation.

Although this project only employed word count and contribution count metrics for dynamic visualisation, it has proved that by utilising a set of tool kits such as file downloading, parsing and visualising tools other metrics could also be added. For instance, the project could be extended by visualising the behaviour of contributors. It'll be great if more metrics can be added and the user would be able to combine those metrics see the correlation between any two metrics.

This project uses the Knowledge Management article as a test bed for the research purpose. When published the dynamic visualisation for evaluation, one responder expressed an interest in being able to use the tool to visualise any other article. Future work on automatically generating visualisation script is required. However, the problem is that building the dynamic visualisation for a particular article needs a lot of network file transferring to retrieve the article revisions. It could take a couple of minutes for the request to come back, which is not a tolerable amount of time. One of the possible solutions is to queue the request and send back the users an email with the



script for dynamic visualisation. It is promising if the visualisation can be integrated into the official Wikipedia for each article and put a link on top of the page similar to “edit this page” and “history” links. The Wikipedia website will be able to directly query the data from its local database and generate the visualisation script.

This project focused on the syntax of the article content change by utilising word count and contribution count metrics. The weakness of the syntax visualising is that it only reflects a limited view on the content change. For example, in terms of the word count metrics, a user can easily rewrite the whole sections by totally changing the content meaning with the same number of words. This leads to the requirements of semantic measurement of the content change by comparing and contrasting the text of sections in revisions. This could be both a promising and challenge research area as one need to figure out how the content is semantic identical but syntax differential and how to encode the change for processing.

The wiki technology and platform that powers the Wikipedia website has been widely spread within modern organisations. While departments of modern organisations are scattering out all over the world, employees are gradually relying on wiki technology to share, transfer and create knowledge to improve their performance and effectiveness. Large volume of knowledge is embedded in the organisations wiki database, which gives the chance for visualising the path of knowledge creation. One of the essential requirements for visualisation on the content change is the availability of article histories. While Wikipedia is an open content collaboration platform, organisation wiki is an invaluable and sensitive assert. This leads to the requirements of fair use for article revisions on organisation wikis.

## **7.6 Conclusion**

The purpose of this research project was to visualise the content change of Wikipedia article and to evaluate the usefulness of the visualisation. This objective was achieved. Whether the representations of visualisations are useful as a tool for Wikipedia users as well as a tool to improve knowledge creation and sharing in Wikipedia is open for debate. During the time taken to conduct this research, no other similar visualisation existed, especially focusing on the content change rather than the metadata for the

article revisions. Although the metrics used for visualisation are still on syntax level, the research showed that tracking content change is an interesting research line worth further investigation in the Wikipedia visualisation area.

## **BIBLIOGRAPHY**

Alexa, Wikipedia Site Information, 2008

Website: <http://www.alexa.com/data/details/main/wikipedia.org>

Accessed August 27, 2008

Alavi, M. & Leidner, D.E., 1999. Knowledge management systems: issues, challenges, and benefits. *Communications of the AIS*, 1(2es).

Ann M. Lally and Carolyn E. Dunford., May/June 2007. 'Using Wikipedia to Extend Digital Collections', *D-Lib Magazine*, Volume 13, Number 5/6.

ASYMPTOTE, The Virtual New York Stock Exchange, 1998

Website: <http://www.asymptote.net/>

Accessed: September 5, 2008

Banerjee, S., Ramanathan, K. & Gupta, A., 2007. Clustering short texts using wikipedia. *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, 787-788.

Butler, B., Joyce, E. & Pike, J., 2008. Don't look now, but we've created a bureaucracy: the nature and roles of policies and rules in wikipedia. In *Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*. Florence, Italy: ACM, pp. 1101-1110.

Best, D., 2006. *Web 2.0: Next Big Thing or Next Big Internet Bubble*. Lecture Web Information System. Technische Universiteit Eindhoven.

Berners-Lee, T., 1996. WWW: Past, Present, and Future. *IEEE Computer*, October, 69-77.

Bixler, C.H., 2002. Applying the four pillars of knowledge management. *KMWorld Magazine*, 11(1).

Blumenstock, J.E., 2008. Size matters: word count as a measure of quality on wikipedia. In Proceeding of the 17th international conference on World Wide Web. Beijing, China: ACM, pp. 1095-1096.

cURL, cURL and libcurl, 1996

Website: <http://curl.haxx.se/>

Accessed August 5, 2008

Ellson, J. et al., 2002. Graphviz— Open Source Graph Drawing Tools. In Graph Drawing. pp. 594-597.

Evans, M.N., Bassuk, N. & Trowbridge, P., 1990. Sidewalk design. Landscape Architecture, 80(3), 102-3.

Eppler, M.J. & Burkhard, R.A., 2005. Knowledge Visualisation. Towards a New Discipline and its Fields of Application. Schwartz, DG Idea Group, Wiley.

Gapminder, About Gapminder, 2007

Website: <http://www.gapminder.org/about/about/>

Accessed: August 7, 2008

GNU Free Documentation License, 2002

Website: <http://www.gnu.org/licenses/fdl-1.2.txt>

Accessed: August 4, 2008

Harold, E.R., 2002. Processing Xml with Java, Addison-Wesley Longman Publishing Co., Inc. Boston, MA, USA.

HousingMaps, Craigslist-GoogleMaps combo site, 2008

Website: <http://www.housingmaps.com/>

Accessed: September 5, 2008

Harms, I. & Schweibenz, W., 2001. Evaluating the Usability of a Museum Web Site. Papers Museums and the Web 2001.

Hanrahan, P., Eick, S., Ebert D., Hansen, C., Joy, K., Ward, M., Billingham, M., White, D., 2005. Visual Representations and Interaction Technologies. Illuminating the Path: The Research and Development Agenda for Visual Analytics, IEEE Computer Society.

Joinson, A.N., 2008. Looking at, looking up or keeping up with people?: motives and use of facebook. In Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems. Florence, Italy: ACM, pp. 1027-1036.

Knowledge Base, Definition of Terms, 2002

Website: <http://www.kwbsolutions.com/kbsterms.htm>

Accessed: August 21, 2008

KartOO, KartOO Visual Meta Search Engine, 2008

Website: <http://www.kartoo.com/>

Accessed: August 23, 2008

Lim, E.P. et al., 2006. Measuring qualities of articles contributed by online communities. Proc. of WI, 6, 81–87.

Liss, K., 1999. Do we know how to do that? Understanding knowledge management. Harvard Management Update, 1-4.

Lewicki, R.J., Saunders, D. M. & Minton, J. W. 1997. Essentials of Negotiation. Boston: Irwin Mc Graw-Hill.

MediaWiki.org, Welcome to MediaWiki.org, 2003

Website: <http://www.mediawiki.org/wiki/MediaWiki>

Accessed: August 26, 2008

Milne, D.N., Witten, I.H. & Nichols, D.M., 2007. A knowledge-based search engine powered by wikipedia. Proceedings of the sixteenth ACM conference on Conference on information and knowledge management, 445-454.

Marwick, A.D., 2001. Knowledge management technology. IBM Systems Journal, 40(4), 814-830.

Murray, P., Poole, D. & Jones, G., 2005. Contemporary Issues in Management and Organisational Behaviour, Thomson Learning Nelson.

Many Eyes, World Map of Medals, Population and GDP, 2008

Website:<http://services.alphaworks.ibm.com/manyeyes/view/SKsauPsOtha6P6knrqlvP>  
2~

Accessed: August 23, 2008

Mosaic Wikipedia Visualisation, EN Wikipedia Visual Browser, 2008

Website: <http://scimaps.org/maps/wikipedia/20080103/>

Accessed: August 24, 2008

MySQL, The world's most popular open source database, 1998

Website: <http://www.mysql.com/>

Accessed: August 25, 2008

Motion Chart, Google Visualization API, 2007

Website:<http://code.google.com/apis/visualization/documentation/gallery/motionchart.html>.

Accessed: August 27, 2008

Nonaka, I. & Takeuchi, H., 1995. The Knowledge-Creating Company. New York, 1, 995.

Ortega, F., Gonzalez-Barahona, J.M. & Robles, G., 2008. On the Inequality of Contributions to Wikipedia. Hawaii International Conference on System Sciences, Proceedings of the 41st Annual, 304-304.

O'Reilly, T., 2005. What is Web 2.0. Design Patterns and Business Models for the Next Generation of Software, 30, 2005.

Ortega, F. & Barahona, J.M.G., 2007. Quantitative analysis of the wikipedia community of users. In Proceedings of the 2007 international symposium on Wikis. Montreal, Quebec, Canada: ACM, pp. 75-86.

Opte Project, The Opte Project, 2003

Website: <http://opte.org/>

Accessed: August 23, 2008

Pei, M. et al., 2008. Constructing a Global Ontology by Concept Mapping Using Wikipedia Thesaurus. Advanced Information Networking and Applications-Workshops, 2008. AINAW 2008. 22nd International Conference on, 1205-1210.

Ponzetto, S.P. & Strube, M., 2007. Knowledge Derived From Wikipedia For Computing Semantic Relatedness. Journal of Artificial Intelligence Research, 30, 181-212.

Porter, M.E., 2001. Strategy and the Internet. HARVARD BUSINESS REVIEW, 79(3), 62-79.

Rao, M., Eight keys to successful KM practice, 2003

Website:[http://www.providersedge.com/docs/km\\_articles/Eight\\_Keys\\_to\\_Successful\\_KM\\_Practice.pdf](http://www.providersedge.com/docs/km_articles/Eight_Keys_to_Successful_KM_Practice.pdf)

Last Accessed: September 7, 2008

Schreiber, G., 2000. Knowledge Engineering and Management: The Commonkads Methodology, MIT Press.

Shenk, D., 1997. Data Smog: Surviving the Information Glut, HarperCollins Publishers New York, NY, USA.

Su, N.M. et al., 2007. The gospel of knowledge management in and out of a professional community. Proceedings of the 2007 international ACM conference on Conference on supporting group work, 197-206.

Steven, D., java-wikipedia-parser, 2007

Website: <http://code.google.com/p/java-wikipedia-parser/>

Accessed: July 2, 2008

Stein, K. & Hess, C., 2007. Does it matter who contributes: a study on featured articles in the german wikipedia. Proceedings of the 18th conference on Hypertext and hypermedia, 171-174.

Suh, B. et al., 2007. Us vs. Them: Understanding Social Dynamics in Wikipedia with Revert Graph Visualisations. Visual Analytics Science and Technology, 2007. VAST 2007. IEEE Symposium on, 163-170.

Siegle, D., 2007. Moving beyond a Google Search: Google Earth, SketchUp, Spreadsheet, and More. Gifted Child Today, 30(1), 5.

Schonhofen, P., 2006. Identifying Document Topics Using the Wikipedia Category Network. Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence, 456-462.

Sinclair, P.A.S., Martinez, K. & Lewis, P.H., 2007. Dynamic link service 2.0: using wikipedia as a linkbase. Proceedings of the 18th conference on Hypertext and hypermedia, 161-162.

Visual Thesaurus, What is the Visual Thesaurus, 1998

Website: <http://www.visualthesaurus.com/howitworks/>

Accessed: August 23, 2008

Viégas, F.B., Wattenberg, M. & Dave, K., 2004. Studying cooperation and conflict between authors with history flow visualisations. Proceedings of the SIGCHI conference on Human factors in computing systems, 575-582.



Viégas, F.B., 2007. The Visual Side of Wikipedia. Proceedings of the 40th Annual Hawaii International Conference on System Sciences.

VandalProof, VandalProof, 2006

Website: <https://amidaniel.com/pub/VandalProof/>

Accessed: August 19, 2008

Valiati, E.R.A., Freitas, C.M.D.S. & Pimenta, M.S., 2008. Using multi-dimensional in-depth long-term case studies for information visualization evaluation. In Proceedings of the 2008 conference on Beyond time and errors: novel evaluation methods for Information Visualization. Florence, Italy: ACM, pp. 1-7.

Viégas, F.B. et al., 2007. Many Eyes: A Site for Visualization at Internet Scale. IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS, 1121-1128.

Viegas, F.B. et al., 2007. Talk Before You Type: Coordination in Wikipedia. HAWAII INTERNATIONAL CONFERENCE ON SYSTEM SCIENCES, 40(3), 1298.

Wget, GNU Wget, 1996

Website: <http://www.gnu.org/software/wget/>

Accessed: August 24, 2008

Waters, N.L., 2007. Why you can't cite Wikipedia in my class. Commun. ACM, 50(9), 15-17.

Wikimedia Foundation Financial Statements, 2007

Website: [http://upload.wikimedia.org/wikipedia/foundation/4/49/Wikimedia\\_2007\\_fs.pdf](http://upload.wikimedia.org/wikipedia/foundation/4/49/Wikimedia_2007_fs.pdf)

Accessed: August 27, 2008

Wilkinson, D.M. & Huberman, B.A., 2007. Cooperation and quality in wikipedia. Proceedings of the 2007 international symposium on Wikis, 157-164.

wikEd, A full-featured in-browser editor for Wikipedia and other MediaWikis, 2006

Website: <http://userscripts.org/scripts/show/12529>

Accessed: August 19, 2008

Wenger, E., McDermott, R. & Snyder, W.M., 2002. Cultivating communities of practice (Boston, MA, Harvard Business School Press).

Wenger, E., 2006. Communities of practice: A brief introduction. North San Juan, CA:

Author.< <http://www.ewenger.com/theory/index.htm>>. June, 19, 2007.

Wordle, Wordle - Beautiful Word Clouds, 2008

Website: <http://wordle.net/>

Accessed: August 23, 2008

Yu, J., Thom, J.A. & Tam, A., 2007. Ontology evaluation using wikipedia categories for browsing. Proceedings of the sixteenth ACM conference on Conference on information and knowledge management, 223-232.

YouTube, Broadcast Yourself, 2005

Website: <http://www.youtube.com/>

Accessed: August 18, 2008

Zeng, H. et al., 2006. Mining Revision History to Assess Trustworthiness of Article Fragments. Proc. of the 2nd Intl. Conf. on Collaborative Computing: Networking, Applications, and Worksharing (COLLABORATECOM).

# APPENDIX A

## Knowledge Visualisation for Wikipedia Survey Screen Shots:

### Knowledge Visualisation for Wikipedia

#### 1. How do you find Wikipedia...

**\* 1. Do you view articles in Wikipedia regularly?**

Never

Occasionally

Regularly

Daily

**\* 2. In general do you find the articles useful on Wikipedia?**

Yes

No

Other (please specify)

\_\_\_\_\_

**\* 3. How reliable do you find the content of articles material on Wikipedia?**

Very Unreliable

Sometimes Reliable

Generally Reliable

Depends on the authors

Other (please specify)

\_\_\_\_\_

**\* 4. How do you find the quality of articles on Wikipedia?**

Very Poor

Poor

Good

Very Good

Excellent

Depends on the authors

Other (please specify)

\_\_\_\_\_

**\* 5. When reading an article on Wikipedia, have you ever read any of its history edition?**

Yes

No

Page 1

## Knowledge Visualisation for Wikipedia

**\* 6. As a reader of Wikipedia, do you think the history versions of articles are as useful a resource as the most updated version when considering updates to content?**

Yes

No

Other (please specify)

**\* 7. As a reader of Wikipedia, do you think a mechanism to track the history of Wikipedia articles visually would be useful?**

Yes

No

Other (please specify)

**\* 8. Do you contribute to the content of articles on Wikipedia?**

Yes

No

Other (please specify)

## Knowledge Visualisation for Wikipedia

### 2. If you have contributed to the content of articles on Wikipedia...

**\* 1. As a contributor to Wikipedia, do you think the history versions of articles are as useful a resource as the most updated version when considering updates to content?**

Yes

No

Other (please specify)

**\* 2. As a contributor to Wikipedia, do you think a mechanism to track the history of Wikipedia articles visually would be useful?**

Yes

No

Other (please specify)

## Knowledge Visualisation for Wikipedia

### 3. Have you viewed the Wikipedia article on Knowledge Management?

**\* 1. Have you viewed the Wikipedia article on Knowledge Management?**

Yes

No

## Knowledge Visualisation for Wikipedia

### 4. If you have viewed the Wikipedia article on Knowledge Management...

**\* 1. Do you think the article for Knowledge Management on Wikipedia has a neutral point of view?**

Yes

No

Other (please specify)

**\* 2. How do you find the definition of Knowledge Management on Wikipedia?**

Very Poor

Poor

Good

Very Good

Excellent

Other (please specify)

**\* 3. How do you find the quality of article for Knowledge Management on Wikipedia?**

Very Poor

Poor

Good

Very Good

Excellent

Other (please specify)

**\* 4. Having viewed the visualisation tool, do you think such a tool would be useful in tracking the evolution of article for Knowledge Management?**

Not Useful

Little of Use

Useful

Very Useful

Excellent

Other (please specify)

## Knowledge Visualisation for Wikipedia

**\* 5. Having viewed the visualisation of Knowledge Management, have you gained better understanding of the topic after seeing the evolution of its article on Wikipedia?**

Yes

No

Other (please specify)



## Knowledge Visualisation for Wikipedia

### 5. If you have NOT viewed the Wikipedia article on Knowledge Management...

**\* 1. Having viewed the visualisation tool, do you think such a tool would be useful in tracking the evolution of articles on Wikipedia?**

- Not Useful  
 Little of Use  
 Useful  
 Very Useful  
 Excellent

Other (please specify)

**\* 2. Having viewed the visualisation tool, do you think it would contribute to better understanding on topics covered on Wikipedia?**

- Yes  
 No

Other (please specify)

## Knowledge Visualisation for Wikipedia

### 6. Finally...

**\* 1. Having viewed the visualisation, do you think such a tool could influence the creation of new content or alteration of existing content on Wikipedia articles?**

- Not Useful
- Little of Use
- Useful
- Very Useful
- Excellent

**2. Please note any additional comments you may have that could contribute to further development of the visualisation tool.**