University of Northern Iowa

UNI ScholarWorks

Graduate Research Papers

Student Work

2005

# Keyword or Hierarchical Searching? A Quantitative Comparison of World Wide Web Searching Methods by High School Juniors in North Iowa

Harold Price

Let us know how access to this document benefits you

Copyright ©2005 Harold Price

Follow this and additional works at: https://scholarworks.uni.edu/grp

Offensive Materials Statement: Materials located in UNI ScholarWorks come from a broad range of sources and time periods. Some of these materials may contain offensive stereotypes, ideas, visuals, or language.

Keyword or Hierarchical Searching? A Quantitative Comparison of World Wide Web Searching Methods by High School Juniors in North Iowa

Find Additional Related Research in UNI ScholarWorks

To find related research in UNI ScholarWorks, go to the collection of School Library Studies Graduate Research Papers written by students in the Division of School Library Studies, Department of Curriculum and Instruction, College of Education, at the University of Northern Iowa.

# Keyword or Hierarchical Searching? A Quantitative Comparison of World Wide Web Searching Methods by High School Juniors in North Iowa

This Graduate Research Paper

Submitted to the

Department of Curriculum and Instruction

Division of School Library Media Studies

In Partial Fulfillment of the Requirements for the Degree

Master of Arts

University of Northern Iowa

By
Harold Price
December 28, 2005

This Research Paper by: Harold K. Price, Jr.

Titled: Keyword or Hierarchical Searching? A Quantitative Comparison of World Wide Web Searching Methods by High School Juniors in North Iowa

has been approved as meeting the research paper requirements of the degree of Master of Arts

Barbara Safford

1/19/06
Date Approved

Graduate Faculty Reader

Karla Krueger

1-19-06
Date Approved

Graduate Faculty Reader

Greg P. Stefanich

1-19-06
Date Approved

Head, Department of
Curriculum and Instruction

# Abstract

In the spring of 2005, twenty 11th grade students from north central Iowa were given 5 topics to search on the Internet. All students were allowed 5 minutes to search for each of 5 topics and used identical equipment to perform the searches. The population consisted of 10 male and 10 female students. As there is a marked lack of ethnic and cultural diversity in the community, all students were Caucasian with English being their first language. All students were of similar socio-economic class since there is little variance in the community. Students were chosen based primarily on a range of grade point averages. Students who were selected to participate in the study had a cumulative grade point average of between 2.5 and 3.5 for the current school year. Students who participated had a working knowledge of the hardware and software being used.

The participants were divided into 2 groups of 10 with 5 boys and 5 girls in each. Each group searched for 5 topics using Google. One group searched using Google Search, a text based search engine utilizing keywords to locate related websites, the other group used Google Directory, a hierarchically organized directory of websites catalogued by topic. Since the directory search was new to most of the participants, a short lesson was given to all participants on the basic concept and usage of hierarchically organized directories. Each group was given 5 minutes to search for each of the 5 topics.

A search log kept by each participant asked questions relating to the participant's experience with the search of each topic. The search log allowed participants to rate each search according to overall difficulty, difficulty determining keywords or search path, difficulty finding information, and the speed that the information was found. The log also contained items that the researcher addressed by analyzing the search histories of each participant to determine the number of inappropriate sites found, the number of non-functional sites found, and the number of sites found unrelated to the search. Only the top 20 sites returned at the end of a search were considered.

While the data showed that both types of searching returned no non-functional sites within the top twenty returns. Both also did exceptional jobs of returning relevant sites with Google Search returning a minimum of 19.5% and Google Directory 22.2% of sites that did not pertain directly to the search query. There were larger differences between the two methods when considering participant perceived speed and difficulty. In nearly all instances the participants chose Google Search over Google Directory as faster and easier. Factors that may have influenced these results include lack of prior experience with the directory form of searching, the cognitive form of the topics used, and lack of prior knowledge of the topics. Keyword searching showed a definite advantage in these circumstances as a term from the topic question could be searched even if the searcher did not have knowledge of the topic itself. Where as without prior knowledge or understanding of a topic, a searcher would not know where to start using a topical or hierarchical directory. These factors and others combined to create a situation where participants favored the path of least resistance or at least the path that was more familiar. Keyword searching, with modern relevancy ranking technology was definitely the favorite of these students.

## Table of Contents

Page

Lists of Diagrams, Tables and Charts v

Chapter 1: Introduction 1
Background 2
Problem Statement 12
Purpose 12
Hypotheses 12
Assumptions 13
Limitations 13
Definitions 14
Significance 15

Chapter 2: Review of Related Literature 17

Chapter 3: Methodology 26

Chapter 4: Data Analysis 30

Chapter 5: Summary, Conclusions and Recommendations 40

Reference List 50
Appendix A Computer and Internet Experience Survey 52
Appendix B Human Participants Review-Informed Consent/Assent 53
Appendix C Search Log 56
Appendix D Lesson Plan-Internet Searching 57
Appendix E Data Set 58

Lists of Diagrams, Tables and Charts

## Diagrams

Diagram 1: Yahoo Search Advanced Search Screen 6
Diagram 2: Hotbot Search Results Showing Commercial Ads and Sponsored Links 7
Diagram 3: HTML Code Showing Various Search Descriptors 8
Diagram 4: HTML Code Showing Example of Unrelated Keyword Descriptors to Increase Search Result Volume 9
Diagram 5: Google Web Directory Search Result Showing Topic Path, URL, Summary and Relevancy to Search. 11

## Tables

Table 1: Number of Irrelevant Sites Returned in the Top 20 Returns from Google Search 31
Table 2: Number of Irrelevant Sites Returned in the Top 20 Returns from Google Directory 33
Table 3: Number of Inoperative or 'Dead Links" found in the Top 20 Returns from Google Search 35
Table 4: Number of Inoperative or 'Dead Links" found in the Top 20 Returns from Google Directory 36
Table 5: Participant Perception of Overall Speed of Searching 37
Table 6: Participant Perception of Overall Difficulty of Searching 38

## Charts

Chart 1: Participant Perception of Overall Speed of Searching 42
Chart 2: Participant Perception of Overall Difficulty of Determining Keywords or a Directory Search Path 43
Chart 3: Perception of Overall Difficulty of Searching 44
Chart 4: Total Number of Sites Returned 45
Chart 5: Unrelated Sites Returned in the Top 20 Results 46

# Chapter One

## Introduction

> "Knowledge is of two kinds: we know a subject ourselves, or we know where we can find information upon it."
> Samuel Johnson
> (1709 - 1784)

How do typical high school students find information on the Internet? At the Rudd, Rockford, Marble Rock High School in Rockford, Iowa they usually go to a search engine on the World Wide Web. That sounds simple enough, but what happens when they are faced with literally thousands or even hundreds of thousands of choices for their query? They often become overwhelmed by the number of sites and find themselves hopelessly attempting to view and analyze information that may or may not be relevant; or they may choose one of the first sites, whether it is relevant or accurate or not. Another concern is the limited time that students have to execute their searches. Teachers sometimes make assignments and assume that students can use study halls for research or do the research at home. The reality is that many students don't have access at home, and with the limited computer/student ratio at school their time to search is cut and efficiency becomes paramount. If a teacher does schedule time for a class to do research, they may even be restricted to a portion of a class period.

Once the opportunity to do research affords itself, and assuming that all the equipment is working and students get to a search engine, they now have to think of appropriate keywords for searching and wade through the results. Consider that in 1997 search engines contained about 50 million sites in their databases (Feldman, 1997), and in 2003 a search too broad in scope may have returned over a billion hits. This means that students must look at each of the results and using the miniscule amount of information provided, determine if the site is of value to their search. If this cannot be determined easily, they either have to visit the questionable site to determine its relevancy or skip it and possibly miss valuable information, (not to mention the sites that look promising only to find that they are nonfunctional). Even when search engines sort by relevancy, (a mechanical process to determine if a site contains information that matches the query), the result can be very frustrated students.

### ***Background***

In 1969, the Advanced Research Projects Agency, an agency under the United States Department of Defense, set up an experimental computer network named ARPANET. The purpose of this network was to facilitate communication between the military, defense contractors and universities. A major motivation of this network was its ability to allow communication even if some of the computers were offline (Howe, 2000). The concern was the fear of nuclear attack knocking out part of a network. This concern was later found to be irrelevant as the military has now determined that in the event of a nuclear attack, an electro-magnetic pulse generated by a nuclear detonation would knock out 95% or more of all electronic equipment (Gromov, 2002).

The Internet evolved from the desire to connect various research networks in America and Europe. First DARPA, (Defense Advanced Research Projects Agency; formerly ARPA), established a program to investigate the interconnection of heterogeneous networks. This program was called Internetting and was based on the concept of open architecture networking, in which networks with defined standard interfaces would be interconnected by gateways (*Encyclopedia Britannica*, 2003). The Internet is the world's largest computer network and enables computers of different types, sizes and operating systems for the purpose of information sharing (Carnahan, 1998).

As a simple network intended to transmit data from one institution to another, the Internet was adequate. SMTP, (simple mail transfer protocol) and FTP, (file transfer protocol), were the languages of choice because all that was being done was the transmitting of electronic messages, (email) and the transfer of data files (*Encyclopedia Britannica*, 2003). In the 1980s the trend for the Internet shifted from data transfer to communication and interaction.

From these relatively simple beginnings, the Internet began to explode. What started in 1969 with four host computers, (computers with registered IP addresses), grew to 80,000 hosts in January of 1989. In that year alone, the number of host computers multiplied to 313,000 in just 10 months. It was in 1989 that domain registration brought commercialism to the Internet. Between January and July of 1989, 3,900 domain names came into being. In the incredibly short space of the next 8 years the Internet increased in size exponentially, to 19,540,000 hosts and 1,301,000 domains. By 1997 there were 171 countries connected to each other via the Internet, and information sharing would never be the same (Zakon, 2003).

### *Birth of the World Wide Web*

From its inception the goal of the World Wide Web was to be a shared information space through which people (and machines) could communicate. The intent was that this space should span from a private information system to a public information forum. A standard for accessing remote data did exist in the File Transfer Protocol (FTP). But this was not optimal for the web. While FTP was sufficient to simply transfer data files, it was quite slow and not rich in features, so a new protocol designed to operate with greater speed and the ability to traverse hypertext links, Hyper-Text Transfer

Protocol, (HTTP) was designed. With the idea of hypertext firmly in place, the next level for the World Wide Web was to address user interfaces. To facilitate the graphics and hypertext, a new programming language was developed called Hyper-Text Markup Language, or HTML. This development brought about the Web we know today (Berners-Lee, 1996).

Programs called browsers were written to facilitate navigation of the Web. Browsers locate and display Web sites by using what was initially called a URI or Uniform Resource Identifier. The Uniform Resource Identifier gave organization to the Internet and the World Wide Web. As stated by Berners-Lee (1996), the power of a link in the Web is that it can point to a document or resource of any kind anywhere on the Internet. The ability of a link to do this requires global identifiers. These identifiers are the primary element of Web architecture. The now well-known structure starts with a prefix such as http: to indicate into which part of the Internet the rest of the string points. The URI is universal in that any new space on the Internet has some kind of identifying, naming or addressing syntax and can be mapped into a printable syntax and given a prefix. The properties of any given URI depend on the properties of the space into which it points. Depending on these properties, some spaces tend to be known as name spaces, and some as address spaces, but the actual properties of a space depend not only on its definition, syntax and support protocols, but also on the social structure supporting it and defining the allocation and reallocation of identifiers. The web architecture, fortunately, does not depend on the decision as to whether a URI is a name or an address, although the phrase URL (Uniform Resource Locator) was coined in IETF circles to indicate that most URIs actually in use were considered more like addresses than names. The world still awaits the definition of more powerful name spaces (Berners-Lee, 1996).

Being a part of Internet, the World Wide Web was destined to be an integral part of communication and commerce growth. The number of web servers in 1993, (keeping in mind that one server can house several web sites with multiple web pages each), was a modest 130. In less than 10 years, in 2002, the number of web servers was 35,543,105. So how many web pages are housed on 35.5 million servers? While it is nearly impossible to accurately determine the number of pages due to constant growth and variations in response, the estimate in the year 2000 surpassed 1 billion indexable pages (Zakon, 2003). A search on just one of the major search engines today for a common search term would easily surpass the 1 billion mark.

## ***Searching the Web***

But what good is access to the world's largest computer network and all of its vast resources if one can't find what you are looking for? Unless you already knew the URL of the site you wished to visit, you had no way of finding it.

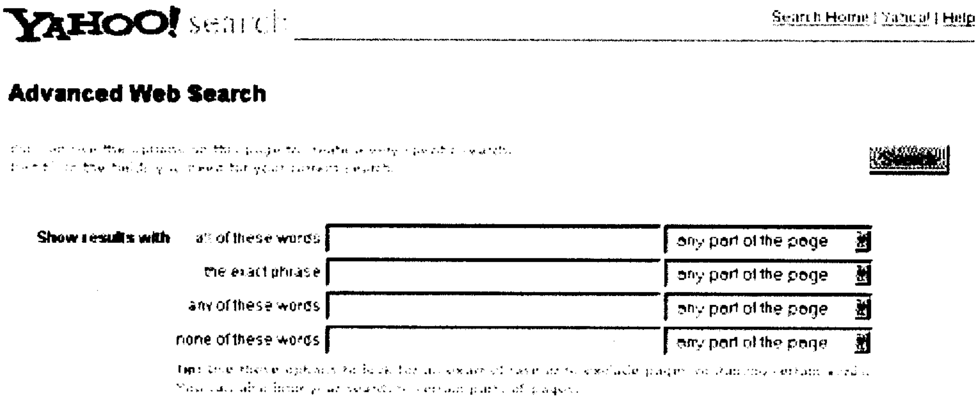
### ***Search engines.***

Search engines are tools for locating information on the Internet. They search the Internet using keywords or phrases designated as search terms by the user. More accurately, a search engine doesn't search the Internet itself, it does search a data base of information (Pealer, 1998). There are several ways that these databases are compiled. One method is to have the authors of the web pages, or their representative, submit the information. This is akin to asking an author of a book to catalog his own book, and it allows for inconsistency and inaccuracy. Another method is to have computer applications, (called spiders or robots or bots), search the Internet for sites and index them by keyword. These keywords can be contained in the keyword or title tags of the web page, or depending on the type of robot application used, anywhere in the content of the web page (Introna, 2000).

The weakness of the search engine is in the simplistic way it arrives at and ranks hits or web sites that match the search criteria. All they really do is match text strings. The greater amount of text matched, the more relevant the site is ranked (Introna, 2000). Therefore a search for the word breast is equally likely to find the web site for the American Breast Cancer Society as it is to find a pornographic site. While better and more specific search criteria do aid in narrowing the information returned, there are still a huge number of irrelevant or unwanted sites to wade through.

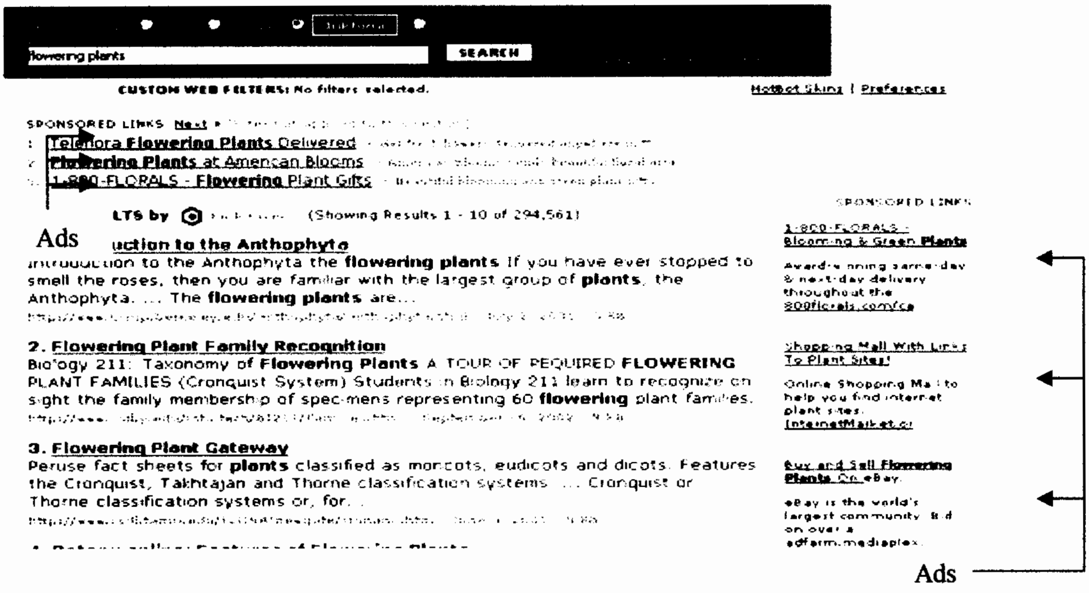
Some search engines offer power or advanced search modes using Boolean search criteria. This search strategy is named after a 1850s English Mathematician George Boole. The crux of the logic in Boolean searching is to add operators to the search terms in the form of AND, OR, and NOT. These operators allow users to narrow search results and weed out some of the unwanted sites by adding additional terms (Pealer, 1998). For instance, a search phrased as, Lord AND of AND the AND Rings NOT game, would find many sites about the book and movie Lord of the Rings, but would exclude sites containing the word game thereby not returning hits regarding the computer or video games of the same name. Most search engines have refined these Boolean terms to a choice of searching for the exact phrase, all of the words or any of the words in a given search, (Diagram 1).

**Diagram 1: Yahoo Search Advanced Search Screen**



It must be pointed out that search engines are, for the most part, not in the business of providing free internet search systems. Rather, they have created or purchased those systems as a means of doing other things such as promoting a brand name, selling advertising space, or advertising a product and so on. Because of this, a search engine tends to be only as good as is necessary to attract potential customers for its sponsors (Clyde, 2000). In a search using the search engine *Hotbot*, (Diagram 2), a search for the keywords flowering plants returned many good sites but also six advertisements on each page of 10 search results. Unfortunately, consumerism is not the only byproduct of this scenario. Politics also enters into the fray. Search engines have the ability to rank sites by relevancy. It is becoming common to find bias in searching to steer users to specific sites for various reasons (Introna, 2000).

**Diagram 2: Hotbot Search Results Showing Commercial Ads and Sponsored Links**


flowering plants
SEARCH
CUSTOM WEB FILTERS: No filters selected.
HotBot Skins | Preferences
SPONSORED LINKS Next
SPONSORED LINKS
Ads
Ads

Telefiora Flowering Plants Delivered

Flowering Plants at American Blooms

1-800-FLORALS - Flowering Plant Gifts

LTS by (Showing Results 1 - 10 of 294,561)

uction to the Anthophyta

introduction to the Anthophyta the **flowering plants** If you have ever stopped to smell the roses, then you are familiar with the largest group of **plants**, the Anthophyta. ... The **flowering plants** are...

2. Flowering Plant Family Recognition

Biology 211: Taxonomy of **Flowering Plants** A TOUR OF REQUIRED **FLOWERING** PLANT FAMILIES (Cronquist System) Students in Biology 211 learn to recognize on sight the family membership of specimens representing 60 **flowering** plant families.

3. Flowering Plant Gateway

Peruse fact sheets for **plants** classified as moncots, eudicots and dicots. Features the Cronquist, Takhtajan and Thorne classification systems ... Cronquist or Thorne classification systems or, for. .

1-800-FLORALS - Blooming & Green Plants

Award-winning same-day & next-day delivery throughout the 800florals.com/fc

Shopping Mall With Links To Plant Sites!

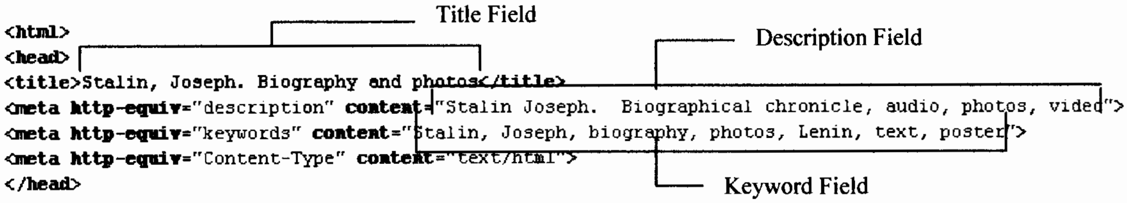
Online Shopping Mall to help you find internet plant sites. InternetMarket.cc

Buy and Sell Flowering Plants On eBay.

eBay is the world's largest community. Bid on over a adfarm.mediaplex.

Web page authors and designers also contribute to the practice of misrepresentation for the sake of increasing traffic to a particular site by padding fields within the web page with selected text strings. It is common to find keywords in the meta or title field of the web page unrelated to the subject of the site (Introna, 2000). The title field is simple to understand. It is a section in the HTML code that defines a document's title, which is displayed in the top of the browser window. Search engines commonly scan the web page title field when searching for text to match. The description meta field is a place in the code of a page that allows the author to insert a line describing the page. This text is usually displayed as part of the information shown as a result of a search along with the title. The keyword meta field is slightly more complex. This field is the place within the html code of the web page that search engines look for keywords to match the search criteria, (Diagram 3).

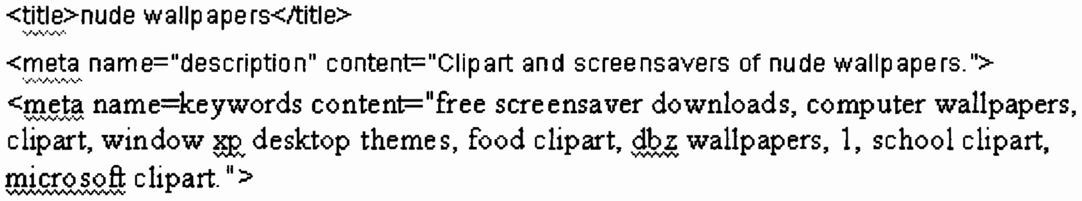
**Diagram 3: HTML Code Showing Various Search Descriptors**



```
<html>
<head>
<title>Stalin, Joseph. Biography and photos</title>
<meta http-equiv="description" content="Stalin Joseph.  Biographical chronicle, audio, photos, video">
<meta http-equiv="keywords" content="Stalin, Joseph, biography, photos, Lenin, text, poster">
<meta http-equiv="Content-Type" content="text/html">
</head>
```

The author of the web page enters keywords that he or she feels would be representative of the content of the site. This can be a powerful and useful tool when used appropriately. The chances that a web site containing plans for building a birdhouse being returned as a hit would increase by adding keywords in the meta field such as, craft, woodworking, birds, houses, or hobbies. Therefore, anyone searching for any of these words would have that site returned as a potential resource (Fleming, 1997). Problems arise when web authors enter keywords into the meta field regardless if those terms have anything to do with the content of the site. The purpose of this tactic is to include commonly searched terms in the field so that the site will be returned as a hit even if it does not pertain to the topic the search intended. An example would be to add the keyword cancer into the meta field of a web site selling nutritional products so that anyone searching for information on cancer would see the site in the results of a search. Authors of adult web sites also use keywords as a powerful tool to get the site seen by as many people as possible. In the source code for a site displaying nude images, shown below (Diagram 4), one can easily see how keywords are used to increase the number of times a site is returned as a search result.

**Diagram 4: HTML Code Showing Example of Unrelated Keyword Descriptors to Increase Search Result Volume**



```
<title>nude wallpapers</title>
<meta name="description" content="Clipart and screensavers of nude wallpapers.">
<meta name=keywords content="free screensaver downloads, computer wallpapers,
clipart, window xp desktop themes, food clipart, dbz wallpapers, 1, school clipart,
microsoft clipart. ">
```

This example for nude images would be displayed for any search containing the keywords free, screensaver, window, downloads, computer, wallpaper, desktop, themes, food, school, clipart, Microsoft and any combination of these terms. Some authors have even resorted to typing these terms in white text so they do not show up under scrutiny. This is inspired by simple greed, the more hits a site with a sponsor receives, the more money the author makes (Introna, 2000).

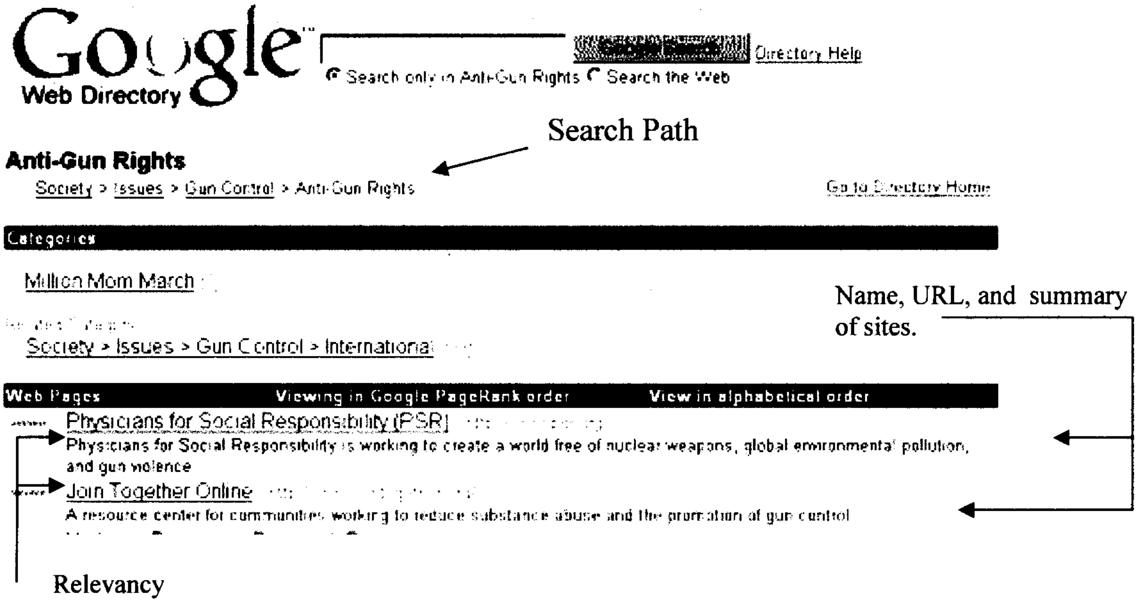
***Hierarchical directories.***

An alternative to the commercialism and politics of search engines is an hierarchical directories, (Diagram 4), also called web rings, portals, directories, and indexes. Web directories, and their more user specific variants portals, “allow users to search through predetermined categories until a site of interest is found. Web directories are assembled and maintained by people, not spiders or bots, and often contain reviews or recommendations to assist users through the content of the site" (Pealer, 1998). This system, while using a hierarchical search logic rather than the text based search engines, can be very useful and fast as the sites usually list resources by subject starting with a broad subject and then narrowing the topic until information that is specific enough is located (Clyde, 2000). Known as hierarchical searching this top-down method of locating information by subject can be very fast and quite accurate.

This hierarchical strategy is easier to use than the various search strategies of search engines. Since web directories use topic or subject classification, they lead to specific information and sites without the clutter, irrelevant, unwanted or inappropriate sites typically returned with search engines (Clyde, 2000). As an example, a search in *Bomis* for Australian art would start by selecting the category of Regions. Australia would be selected from the available list of regions. Then the category of Arts and Culture would lead to sites only related to art and culture in Australia.

Some politics and commercialism still may enter into the web directory concept, but is more obvious and less intrusive and misleading. While little if any advertising may appear, the content of the search in not affected, (Diagram 5). It is usually fairly easy to determine the purpose or target audience of a web directory or portal. A portal that is maintained by a school will obviously contain links related to academics or a topic that the school feels strongly about. Web directories such as *Yahoo* and *Bomis* clearly label topics and sometimes provide warnings about content. Each step of the hierarchical search, from broad subject to specific topic is clear (Clyde, 2000).

**Diagram 5: Google Web Directory Search Result Showing Topic Path, URL , Summary and Relevancy to Search.**



◉ Search only in Anti-Gun Rights ○ Search the Web

**Anti-Gun Rights**

Society > Issues > Gun Control > Anti-Gun Rights

Categories

Million Mom March

Society > Issues > Gun Control > International

Web Pages Viewing in Google PageRank order View in alphabetical order

Physicians for Social Responsibility (PSR)
Physicians for Social Responsibility is working to create a world free of nuclear weapons, global environmental pollution, and gun violence

Join Together Online
A resource center for communities working to reduce substance abuse and the promotion of gun control

So how do web directories build their database of sites? One way is for an author or designer to submit the site, and another is to suggest someone else's site. *Yahoo* has a form to submit in order for a site to be listed. The author can suggest possible topics for a site to be listed under, but a reviewer will attempt to make sure that the site has content that supports the suggestion. *Bomis* not only allows authors and designers the opportunity to submit their own site, but also allows anyone to submit sites that they feel are relevant and allow the creation of new categories called rings, (Hubbard, 1999).

Web directories with a hierarchical search structure are efficient in finding information as they provide a human factor that helps direct content by subject and filter out unwanted and unrelated results. Hubbard makes a profound statement by stating, " I would argue that what really makes indexing and search retrieval difficult to automate are two things that human indexers do and machines do not. One is to consider the audience for a document, whether book or Web page. The other is to keep a mind map or syndetic structure in mind as a document is indexed" (Hubbard, 1999, p.6).

### *Problem Statement*

High school students are not effective web searchers. Search engines that use text-based keyword searching are simple in concept but provide too many sidebars of information and irrelevant returns due to their semantic text limitations and ease of manipulation. Many sites returned are not functional. Many others have content that has absolutely nothing to do with the desired search. Others are related to the search intention but may have incomplete or inaccurate information. Web directories use a topically classified list of sites arranged in a hierarchical structure in order to provide a greater chance of finding information without nonfunctional sites and clutter. They depend on a more complex thought process that forces a searcher to start with a broad concept and successively narrow the search through a series of choices until the desired information is presented.

### *Purpose*

The purpose of this research is to assess information gathered from a group of high school juniors who will execute searches for information using both a text-based search engine and a web directory utilizing a hierarchical indexing method. Upon comparison, data and opinions from the group will be used to determine if the web directory or the search engine leads searchers more efficiently to relevant information.

### *Hypotheses*

(1) Students using a text-based search engine when searching for information on the World Wide Web will have no irrelevant sites returned in the search results.

(2) Students using a hierarchically indexed web directory on the World Wide Web will locate only sites directly related to the intent of the search.

(3) Students using a text-based search engine when searching for information on the World Wide Web will encounter inoperative sites within the sites returned.

(4) Students using a hierarchically indexed web directory when searching for information on the World Wide Web will encounter inoperative sites within the list of sites related to the search.

(5) Students will find that a text-based search engine will provide a slower return of information than a hierarchical indexed web directory when searching for information on the World Wide Web.

(6) Students will find that when searching for information on the World Wide Web a hierarchically indexed web directory will be less difficult than using a text-based search engine.

### *Assumptions*

It is assumed that the high school students participating in the search comparison have experience and some level of competence with text-based search engines and hierarchically indexed web directories. It is also assumed that the students will be of comparable intelligence and ability. This researcher also assumes that the hardware being used to complete the searching will be as equal as possible considering processing speed, random access memory size, and Internet connection type and speed. The researcher will assume that the students and the researcher will not have prior knowledge of the topics to be searched. A final assumption is that the text-based search engine and the hierarchically indexed web directory used are properly maintained by their sponsors.

### *Limitations*

This study will measure responses from high school juniors from a small high school in North Central Iowa. The number of students will be limited as the entire junior class is made up of only 65 students. The ethnic and cultural backgrounds as well as the socioeconomic status of the group will also be limited as the geographic area where the study will be done is not culturally or ethnically diverse nor is there a large variation in socioeconomic class. The study will be further limited by the experience or lack thereof on the part of the participants in the use of hierarchically indexed web directories as they are not commonly taught or suggested for use by the teaching staff.

### *Definitions*

Hierarchical Index

Refers to an index that is organized in the shape of a pyramid, with the top being a broad subject area branching down in to successively narrower topics. (Webpedia, 2003)

Hit

Any time a piece of data matches criteria you set. For example, each of the matches from a Yahoo or any other search engine search is called a hit. (Webpedia, 2003)

HTML

Hyper Text Markup Language is a language used to create webpages. It is not a pure programming language. HTML defines the structure of a page including paragraphs, headings, graphics, and audio and video. (Fleming, 1997, p. 48)

HTTP

Hyper Text Transfer Protocol is the standard communications protocol used on the Internet. It is very flexible and fast. (Fleming, 1997, p. 48)

Internet

A global network connecting millions of computers enabling communication throughout much of the world. (Margolis, 1999, p.283)

Meta Data

For the purposes of this paper Meta Data are strings of text located in a special section of the HTML code of a Webpage used by search engines to index and search the content of the page. Used for titles, descriptions and keywords.

Meta-searcher

For the purposes of this paper, the term meta searcher shall refer to an online search tool that performs a search for a user selected term using several search engines and returning an aggregate of the results. The advantage is the ability to type in the search term once and submit the query once, but benefit from multiple searches.

Portal

A Web site or service that offers a broad array of resources and services, such as e-mail, forums, search engines, and on-line shopping malls. The first Web portals were online services, such as AOL, that provided access to the Web, but by now most of the traditional search engines have transformed themselves into Web portals to attract and keep a larger audience. (Webpedia, 2003)

Search Engine

A tool designed to search the Internet for keywords or phrases designated as search terms by a webpage author. By indexing these search terms or keywords a database is created to identify search terms to specific pages. (Pealer, 1998, p.346)

URI

Uniform Resource Identifier or URL, Uniform Resource Locator. The global address of documents and other resources on the World Wide Web. The first part of the address indicates what protocol to use, and the second part specifies the IP address or the domain name where the resource is located. (Webpedia, 2003)

Web Directory

For the purpose of this paper, a web directory is defined as a listing of World Wide Web sites organized or classified by topic in a hierarchal structure.

World Wide Web

A system of Internet servers that support specifically formatted documents. These documents are formatted in a language called HTML, or Hypertext Markup Language, that supports links to other documents, as well as graphics, audio and video files. (Margolis, 1999, p.617)

### *Significance*

This research will compare two popular web-searching tools. This comparison should yield an answer to the question of which searching method, text-based search engines or hierarchically structured web directories, is the most efficient in locating relevant information for high school students. If one or the other proves to be significantly more efficient, then anyone who has anything to do with assisting students with research would do well to emphasize the skills involved with that search method. Students are expected to learn more, faster than anytime in previous history. When time is at a premium, students should use the most efficient method available to them to locate information. The time savings afforded by searching the World Wide Web more efficiently would allow students to redirect more of their time on synthesizing the information found and have a profound affect on the quality and cognitive rewards of their work.

# Chapter Two

# Literature Review

It can be said that today's high school students literally have the world at their fingertips. Thanks to the Internet and the World Wide Web information can be obtained with a minimum of equipment and knowledge and on a global basis. But even with fast computers and unlimited Internet access, searching for information can be a challenge. Many times a teacher assigns topics to research and students are expected to find credible information in mere minutes. Although high school students of today are vastly more prepared and capable than any who came before them, they still must navigate through the flotsam and jetsam of information available to find what they are looking for.

## *Searching the World Wide Web*

According to the problem statement of a study done by Vansickle in 2002, there is a concern that high school students lack research skills and are not being adequately prepared for college studies. This study explored how research skills were being learned by high school students to determine if their skill and efficiency level were adequate to be conducive to finding information on the Internet. The study consisted of a survey administered to 136 tenth grade students and focused on an attempt to measure general knowledge and search knowledge. In general terms, the results of the survey indicated that nearly three-fourths of the participants felt that they had taught themselves to use the Internet and received no formal instruction on searching the Internet, (p.35). Less than one-fourth indicated that they have used advanced search options and most did not feel that they were efficient in their searching, (p.35). The students did think that individual search assistance and posted search tips would be helpful: however, the majority stated that they would not be interested in an elective class in search skills. To this researcher, this study indicates that students are not able to use adequate search terms and not only lack the knowledge to search effectively, but they also are unaware that they need to improve their skills.

Possibly one of the most serious problems that exists related to Internet or World Wide Web searching is that of what tool to use to perform the search. According to Feldman (1997), at the time of the study there existed more than 1,800 search services on the Web. Search services include general and meta-search engines as well as specialized searchers. The number of search options and the complexity of their operation combined with the huge variations in the skills possessed by the people doing the searching, creates a serious problem for those attempting to find information. The study consisted of performing searches for seven test questions on several search engines. Results of the search were tabulated and the top ten sites returned were checked for relevancy. It was found that while no single search engine distinguished itself among all the others for every type of search, a few did fare better than the others in certain categories, (p.36). It was noted that small variances in search terminology had substantial affect on the results. Something as simple as searching in singular form as opposed to plural gave very different results. The two factors that contributed to poor results were the way web crawlers or spiders search and index sites, and the limitations caused by text-based search terms, (p.36).

But even the best-designed research tools have one common problem, the researcher. It has long been accepted that the more knowledge a person has on a subject, the more efficient they are at locating information for it. A study by Minkel, (2000), addressed the argument of whether or not a student's prior knowledge of a topic has an effect on his searching efficiency while online. In this study two groups of students were given search topics. One group consisted of 12 year-olds and the second group was made-up of 16 year-olds. All members of both groups were considered experienced Web users. Both groups were observed as they searched and the results were tracked. When literally thousands of hits or returns of a search query are shown, it was found that the participants, possibly because of being overwhelmed by the number of returns, tended to skim through results looking for something familiar, (p.22). Due to their lack of knowledge on the subject, they skipped over important information that could have led them to valuable sources. The study also found that animations and graphics supposedly designed to attract attention to a site or a particular part of a site, were regarded as somewhat of a nuisance and students scrolled past them to get them out of view bypassing valuable information. This study concludes that middle school and high school students who know about a subject beforehand are best able to use the results of a Web search on that topic to answer questions, (p.22).

### *Search Engines*

Search engines revolutionized the way people search the Web. As the World Wide Web exploded in content, it became increasingly difficult to locate information. In a study by M. Chau, D. Zeng, and H. Chen in 2001 the credit for increased efficiency in Web searching goes to search engines and their indexing software. According to this study, by 2001 there were "more than 1 billion unique indexable web pages" in what was called "the biggest digital library available", (p.79). Before the advent of indexing utilities, known as crawlers, spiders and bots, the search for information on the Web was plagued by low recall rates and outdated indexes. The study tested two spiders, CI Spider and Meta Spider. These indexers build and update databases of Web pages for search engines. They search the Web and take a snapshot of a Web page and then index it according to text contained in specific areas of the page. Some even consider all of the text on the page including the HTML tags. Searches were performed and the results were compared to Lycos, a general search engine, (p.85). The results indicated that both of these spiders performed better than Lycos in the areas of precision, recall, time and usability. While the use of indexers like CI Spider and Meta Spider will improve search engine results, the same problems of relying on non-human indexing and lack of search skills exist.

In today's fast paced, high expectation world, it seems that the majority of people searching for information are more concerned with speed and ease of use than qualitative results. Few can argue that search engines are the kings of speed in the Web searching world. By typing in a one or two-word search term, one can expect dozens, thousands or even millions of returned sites to consider. Problems such as dead sites, and sites unrelated to the intent of the search can be overlooked when one considers that perhaps billions of sites were searched in a few seconds. But are some search engines better or easier to use or faster than others? Gowan and Spanbauer (2001), of *PC World* did a comparison of twelve of the most popular search tools. This included general search engines, meta-searchers, and directories. They discovered that there are very big differences in the way searches are performed, classified and ranked. Some are just better than others. They do specify in their comparison that user knowledge and willingness to use advanced search features greatly increases the efficiency of the search. It was also found that in order to access some search engine's advanced features, one must first perform a basic search, (p.112). This takes longer and more effort on the part of the user and so they tend to ignore the feature. However, many search engines refine searches for the user by displaying the results by relevancy or rank, and some even place the sites fitting the search term into topical categories, (p.114). Search engines have come a long way in the last ten years. In comparing these twelve search tools, the researchers performed a set of searches on each tools and compared the results with emphasis on relevance, advanced features, ease of use, percentage of dead links, and freshness of the links returned. Their favorite was *Google*, which had high marks in the areas of relevancy and ease of use. The least liked was *Ask Jeeves*, which suffered from too few direct links to information and its lack of advanced features. Two other search tools that were ranked nearly as good as *Google* were *Lycos* and *Yahoo*, both of which are classified as directories by the researchers, (p.115).

### *Other Approaches*

Perhaps it is society's desire for instant gratification in all things that pushes us into an emphasis on speed and simplicity resulting in a disorganized quagmire of information so immense that it overwhelms us and makes finding relevant information very difficult. Organizing the Web may be the next step. In 2000, Chen and Dumais stated that it has become more and more difficult to find information because Web search systems return a ranked list of pages in response to a user's search request. These pages on different topics or different aspects of the same topic are mixed together in the list along with sites that are no longer functional making a difficult and arduous task of finding what the user is looking for. Chen and Dumais developed a new user interface that organizes Web search results into hierarchical categories. It uses text classification algorithms to automatically classify arbitrary search results into an existing category structure on-the-fly. During the study eighteen adult subjects of intermediate web ability from Seattle, Washington, performed 30 searches on the Web in two different sessions using a list type interface similar to a general search engine, and a category interface that grouped results using a text classification system, (p.145). To ensure that results from different subjects were comparable, the keywords or search terms were fixed. Restated, the subjects were given search terms they were required to use and were not allowed to enter search terms or keyword that they themselves thought of. This was done to effectively eliminate the variable created by the variation in the user's prior knowledge of a subject. The study found that searches were faster and provided more relevant results. Participants indicated that the category interface was easy to use and that they liked using it. They also indicated that they felt that they were confident that they could find the information if it was there while using the category interface. The strongest sentiment from the participants was that they preferred the category interface over their usual search engine and the list interface also used in the study, (p.149).

A study was done in Iowa (Flanigan, 2001), which addressed the concern that high school students do not make effective use of their time searching for information. In this study 23 high school students who were enrolled in a composition and perception class submitted the topics for their writing assignment to the Media Specialist who researched the topics using *Google* as a search tool and reviews from professional journals. Sites that were found to be of value to the student's research were then cataloged into the school library's automated catalog system using the 856 tag in the MARC records, (p.54). It is assumed that the researcher used standard Sears subject headings to catalog the sites by topic. Students then were told to research their topics using online public access catalog stations. Students were able to quickly locate material in the library's collection pertaining to their topic as well as the cataloged Web sites that were entered by the researcher. The students were able to use a hyper link to go directly to the Web sites that they found. A survey was completed by all participants at the end of the study in which the students indicated that their satisfaction level was high in the areas of ease of use and efficiency, (p.62). Anecdotal statements from the participants made it clear that they felt the cataloged sites were very useful and that they could find them quickly. While this study did use a hierarchical structure in its search, i.e. the Dewey Decimal Catalog, it circumvented the Web searching problem of evaluating search query returns. While Flanigan's study uses a closed and catalogued set of data instead of the Internet, it still demonstrates that students should have the ability to find information using a classification system.

One way to evaluate search results is to use a tool that has been experimented with for the past 2 or 3 years known as interactive searching. Bruza, McArthur, and Dennis (2000), did a comparison study focusing on three types of searching, standard internet query search (*Google*), directory browsing (*Yahoo*), and interactive searching, also known as phrase-based query reformulation (*Hyperindex*). Phrase-based query reformulation is a concept involving a hybrid search tool that uses a directory search but then displays the results with a twist. Instead of immediately showing the results ranked by relevance or text matching, the results are grouped by content using phrase comparison, (p.280). The content groups, (a directory structure), or groups of sites relating one or another based on phrases contained in those sites, are displayed and the user can select a group that appears to be related to the intent of their search. Once that group is selected, sites are displayed that contain content related to that group. By using a directory to perform the base search, it is known that the sites in those groups have been reviewed and classified by human reviewers. This minimizes the number of dead sites. Since the sites are reviewed and classified search terms and keywords will be more accurate. This allows the grouping program to assign sites to groups using short phrases as descriptors. The user then compares the phrases to their original search concept, (p.284). For the study, fifty-four subjects were recruited from the undergraduate psychology pool at the University of Queensland. Each subject was given a set of six topics to search using one of the three search methods. The subjects were given 5 minutes to locate what they felt was relevant information on each question. They were to then bookmark the relevant sites they found. This allowed the researchers to track the time spent searching, and the number of relevant sites found through the search. The results showed that Yahoo handled searches best related to shopping, but the *Hyperindex* browser provided more relevant sites faster than the other two search methods for all information based queries, p285. The conclusion was that involving a cognitive factor in the search results, i.e. the users selection of a content group from a phrase based selection method, significantly increased search efficiency based on time, number of sites and relevance, (p.286).

### *Summary*

The research discussed here points to some very pertinent conclusions. According to Vansickle (2002), Minkel (2000) and to a lesser extent Chen and Dumais (2000), prior knowledge of a topic is an important factor to search efficiency. Searchers who understood a topic well are more likely to find relevant information faster since they know what to look for. Searchers who did not know the topic or were not able to determine valid keywords had trouble locating relevant sites and often skipped over valuable information because of lack of recognition or the attempt to look at more information that they could process. Gowan and Spanbauer (2001) stated that in their comparison of twelve search tools, one of the biggest factors affecting search efficiency was user knowledge and willingness to use advanced search features.

While prior knowledge was certainly a factor in efficiency, the methodology of the search tools was equally important in the return of valid search results. Feldman (1997) showed that not only did appropriate keywords in context make a difference in search accuracy, but even the singular or plural forms of words made a difference in the results of a search. In the study by Chau, Zeng and Chen (2001) the researchers felt that searching the more than one billion web pages required an improved approach over the standard search engine. Using spider indexers modified to catalog pages by content instead of relying on the author's choice of keywords, was found to greatly enhance search efficiency. This begins to explore the concept of categorizing web pages by content.

While Gowan and Spanbauer (2001) seemed to be focusing on search engines, their results stated that two of the three most efficient search tools were directories. While the comparison did not use a hierarchical search path, the keyword searching in the directories were considered very efficient since instead of relying on spiders or bots to catalog pages, the directories instead rely on human decision on content to catalog the pages. Possibly the most eye-opening study presented was from Bruza, McArthur, and Dennis (2000) who compared a standard searcher, a directory browser, and a hybrid searching mechanism called *Hyperindex*. *Hyperindex*, searched a directory database of web pages, but then attempted to categorize the results into content-based groups. By using content to separate the pages into topical categories, the searcher can select the grouping closest to their intended search topic, thereby bypassing the sites that do not pertain to the search.

The recurring theme of these studies seems to be that Web searching needs to be more efficient. There seem to be two implied solutions to the problem of inefficient Web searching. One is to increase the searcher's knowledge base and the instruction in the use of advanced searching. The other, which seems to be on the horizon, is to combine page content with human cataloging decisions to organize pages by topic. The use of current general searchers using spiders to create data-bases is far from over. Bruza, McArthur and Dennis (2000) even admit that consumer searches were handled better by *Yahoo* and not their *Hyperindex*. But it is obvious that instruction in searching methods would greatly enhance search performance for high school students.

# Chapter Three

# Methodology

High school students are not effective web searchers. Search engines that use text-based keyword searching are simple in concept but provide many sidebars of information and irrelevant returns because of their semantic text limitations and the resulting ease of manipulation. Many sites that are returned are not functional. Many others have content that has absolutely nothing to do with the desired search. Others are related to the search intention, but may have incomplete or inaccurate information. Web directories that use a topically classified list of sites arranged in a hierarchical structure or which provide a greater chance of finding relevant information. They depend, however, on a more complex thought process that forces a searcher to start with a broad concept and to then successively narrow the search through a series of choices until the desired information is presented.

## *Research Design*

An experimental research design was used for the study. Experimental research is a powerful quantitative research method for establishing cause-and-etlect relationships. In the case of this study, the participants were split into two groups. The first group searched for information on pre-specified topics using the search engine, Google. The second group searched for the same topics using the hierarchically based web directory available also from Google. Students tracked the results of their search using a search log, in respect to time, ease of use, difficulty in determining which keywords or search path to use and total number of sites returned. The searches were tracked and saved so the researcher could study the data to determine relevancy and number of non-functional, irrelevant and inappropriate sites.

### *Population to Be Studied*

The student participants in the study were in the 11th grade at the Rudd, Rockford, Marble Rock Senior High School in Rockford, Iowa. Only 20 of the 65 eleventh grade students participated. This sample was selected in order to provide a consistent experimental group. The students were selected with the assistance of the school counselor who had access to student grades and background knowledge of the students. The population consisted of 10 male and 10 female students. As there is a marked lack of ethnic and cultural diversity in the community, all students were Caucasian with English being their first language. All students were of similar socio-economic class since there is little variance in the community. Students were chosen based primarily on a range of grade point averages. Students who were selected to participate in the study had a cumulative grade point average of between 2.5 and 3.5 for the current school year. Students who participated had a working knowledge of the hardware and software being used. This knowledge and the other requirements were determined with a pre-study survey (Appendix A). This survey was given to all 11th grade students who have a current GPA between 2.5 and 3.5 and only students who indicated that they currently use the Internet, use the Internet at least 2 times per week, were at least somewhat comfortable using a computer and indicated that they are competent enough to do things on their own were selected to participate in the study.

### *Preparation*

Permission forms were sent to the parents of all potential participants. Along with the form, a letter of explanation will outlined the study's purpose and scope. Only participants whose parent returned a signed form expressly giving permission for the student to participate were considered (Appendix B).

In an attempt to ensure that all participants had a similar knowledge base of the two searching tools to be used, they attended a 30-minute class which reviewed the concept and searching method of both a text-based search engine and a hierarchically structured web directory. The lesson plan for this class, (Appendix D), included definitions of the two search tools, a brief outline of the mechanics of how each search tool locates information, search logic, and some general search examples. The class was taught by an objective instructor who was familiar with both of the search methods.

### *Equipment*

The computers used in the study were limited to those located in the computer lab in the media center of the school. These MacIntosh G3 computers had 256 megabytes of random access memory, 600 Mhz processors and had no peripherals installed that would affect processing speed. They were connected to the local area network through 10-base network cards to a common server. The Internet and World Wide Web were accessed through this server using a DSL connection.

All of the computers had the same software installed and should have performed very similarly. The operating system used was Mac OS 9. The Internet browser was Microsoft Internet Explorer 5.5. Settings relating to the performance of web browsing were set identically on all of the computers being used.

### *Search topics*

Topics were selected by Barbara Ripp Safford, an associate professor and Program Coordinator of the Division of School Library Media Studies at the University of Northern Iowa in Cedar Falls, Iowa. Topics were not disclosed prior to the start of the study to ensure a reasonable level of impartiality on the part of the researcher.

### *Search tools*

Both the text-based search engine and the web directory came from the same source, *Google.com*. *Google* has been heralded as the premier search engine by several sources and was commonly used by the students participating in the study. Search results included the approximate number of sites found, the site name, the contents of the meta name tag, the URL, and an option to view similar pages. *Google* also offered a topically based, hierarchically structured directory with 15 main categories. Web sites were shown beginning at the third hierarchical level and expanded with each successive level visited. Related sites were also shown at each level and popular topics were in bold type.

### *Process*

The participants were split into two groups equal in gender. Each group searched the web for 5 topics. The first group searched using a text-based search engine, *Google*, and the second group searched using a hierarchically structured web directory, *Google Directory*. Students were allocated 5 minutes to search for each topic. At the conclusion of each search, the students filled out a search log, (Appendix C) about the results of the search. Students ranked the search in respect to time, ease of use, difficulty in determining which keywords or search path to use and total number of sites returned. Search queries were be tracked; the results were be checked for relevancy and number of non-functional, irrelevant and inappropriate sites by the researcher.

# Chapter Four

## Data Analysis

High school students are not effective web searchers. Search engines that use text-based keyword searching are simple in concept but provide too many sidebars of information and irrelevant returns due to their semantic text limitations and ease of manipulation. Many sites returned are not functional. Many others have content that has absolutely nothing to do with the desired search. Others are related to the search intention but may have incomplete or inaccurate information. Web directories use a topically classified list of sites arranged in a hierarchical structure in order to provide a greater chance of finding information without nonfunctional sites and clutter. They depend on a more complex thought process that forces a searcher to start with a broad concept and successively narrow the search through a series of choices until the desired information is presented.

The purpose of this research was to assess information gathered from a group of high school juniors who executed searches for information using both a text-based search engine and a web directory utilizing a hierarchical indexing method. Upon comparison, data and opinions from the group were used to determine if the web directory or the search engine led searchers more efficiently to relevant information.

Six hypotheses structured the analysis of the study. The data set is Appendix E. Hypothesis number one stated, "Students using a text-based search engine when searching for information on the World Wide Web will have no irrelevant sites returned in the search results." The fifth question on the search log under the Google Search section dealt directly with the number of irrelevant sites or sites that were unrelated to the topic being searched for within the first 20 results. The participants were instructed not to answer this question as it would be completed by the researcher. After student searching was completed, the researcher accessed the histories of each computer used to observe the search results and determine the number of irrelevant sites the participants encountered. The possible answers for that question ranged from a score of zero for no unrelated sites encountered to a score of five, if five or more sites were found. Accordingly, a score of one was tallied for one site, two for two sites, three for three sites and four, if four unrelated sites were encountered. Table 1 illustrates the data used to address Hypothesis number 1.

**Table 1: Number of Irrelevant Sites Returned in the Top 20 Returns from Google Search**

| Score | No Sites<br>0 | 1 | 2 | 3 | 4 | Five or More Returns<br>5 | Total Score | Average Score | Percent |
|---|---|---|---|---|---|---|---|---|---|
| Responses | 0 | 0 | 11 | 9 | 4 | 26 | 195 | 3.9 | 19.5% |

There were ten participants searching for five topics. This resulted in fifty searches with the top 20 returns being considered resulting in 1000 possible sites returned. As the table indicates, there were no topics that returned zero or only one irrelevant site. Eleven responses encountered two sites; nine found three sites; five returned four sites; and twenty-six encountered five or more sites unrelated to the search topic. This indicates a minimum of 195 out of 1000, or 19.5% of the sites that did not pertain to the search topic.

It should be noted that most of the sites returned were somewhat close in relation to the topic. This resulted in sites returned, which if explored, may have provided a link that would have led to other sites with information on the history of the city and the information wanted. This relationship of sites that are not directly related to the search topic but may still prove valuable is more relevant than sites that are completely removed from the subject. For instance, one search for the topic “What about the allies capture of

Antwerp in 1944 turned out to be a disaster? " included the keywords 'Antwerp' and '1944'. This resulted in sites returned regarding tourism in Antwerp. Following this link may have led to other sites with information on the history of the city and the information wanted. Likewise, the search for the topic "Describe the ethical issues in organ donations." using the keywords 'organ donation' resulted in sites describing organ donation organizations which, if followed, may have led to ethical issues. In these two examples the sites returned were not directly related to the search topic in that the information on that site did not answer the questions sought. However, they may still prove valuable if explored, as related information or links to other sites may have led the searcher to the correct information. The presence of these sites that were not directly related to such a specific search but could still prove useful may be attributed to the search relevancy technology used by Google which attempts to rank sites by comparing phases in order to determine what information the user is attempting to find. Even so, the data are clear. It is apparent that a minimum of 19.5% of the sites returned by searches done by the participants were unrelated to the topic; therefore, hypothesis number 1 is rejected.

Hypothesis number 2 stated, "Students using a hierarchically indexed web directory on the World Wide Web will locate only sites directly related to the intent of the search." Question number 5 under the Google Directory section of the search log dealt directly with the number of irrelevant sites or sites that were unrelated to the topic being searched for within the first 20 results. The participants were instructed not to answer this question. After the study, the researcher accessed the histories of each computer used to observe the search results and determine the number of irrelevant sites the participants encountered. The possible answers for that question ranged from a score of zero for no unrelated sites encountered to a score of five, if five or more sites were found. Accordingly, a score of one was tallied for one site; two for two sites; three for three sites; and four, if four unrelated sites were encountered. Table 2 illustrates the data used to address Hypothesis number 2.

**Table 2: Number of Irrelevant Sites Returned in the Top 20 Returns from Google Directory**

| Score | No Sites<br>0 | 1 | 2 | 3 | 4 | Five or More Returns<br>5 | Total Score | Average Score | Percent |
|---|---|---|---|---|---|---|---|---|---|
| Responses | 0 | 2 | 8 | 2 | 2 | 36 | 212 | 4.24 | 21.2% |

There were ten participants searching for five topics. This resulted in fifty searches with the top 20 returns being considered resulting in 1000 possible sites returned. As the table indicates, there were no topics that returned zero irrelevant sites. Two responses encountered one site; eight returned two sites; two found three sites; two returned four sites; and thirty-six encountered five or more sites unrelated to the search topic. This indicates a minimum of 212 out of 1000, or 21.2%, of the sites did not pertain to the search topic.

In the evaluation of hypothesis number 2, irrelevant sites returned by Google Directory, first appearances would make it seem this high percentage of irrelevant sites, (21.2%), especially with 36 responses in the 5 or more category would not bode well for the directory form of searching. However, during the data analysis, the researcher noted that many of the responses in that category appeared to stem from the participants following the wrong path on their way from the general directory categories to the specific information they were in search of. For example, when looking for information on the Chaco War, it appeared that the some of the participants followed a logical progression starting with the category 'Society' and chose from the sub categories the displayed topic, 'Politics'. In the 'Politics' topic screen the participants correctly chose the item labeled ‘Wars and Conflicts’. At this point the participants were at a loss as there was no selection labeled ‘Chaco War’. At this point they began a hit and miss method of searching. It appears the participants ran out of time before locating the correct information which resulted in these being counted as more than 5 irrelevant sites. In similar, but more severe occurrences, the search pattern was interrupted at the beginning by a lack of a participant’s knowledge on a topic. It appeared that if the participant had no base knowledge of a topic he or she had no idea where to start to look for information. In the case of looking for information about the capture of Antwerp during the Second World War, one participant did not know where Antwerp was or that it was connected to World War 2. Their lack of base knowledge with no other clues from the question resulted in only being able to do what appeared to be random selections in the category list with no success. It is important to note when participants did have a solid base knowledge of a subject, they easily followed a logical progression through the topic lists and found relevant information with very few, if any, unrelated sites. With the topic involving the Cubists, (a topic recently covered in an art class), nearly all of the participants started with the category ‘Arts’, and then followed a fairly direct path through ‘Art History’, then ‘Periods and Movements’ arriving at ‘Cubism’. While there were a few attempts that resulted in dead ends, the participants quickly backed out to the last topic heading and redirected themselves to the correct category. The raw data in table 2 makes it apparent that a minimum of 21.2% of the sites returned by searches done by the participants were unrelated to the topic; therefore, hypothesis number 2 is rejected.

Hypothesis number 3 stated, “Students using a text-based search engine when searching for information on the World Wide Web will encounter inoperative sites within the sites returned.” Question number 3 in the Google Search section of the search log dealt directly with the number of inoperative or dead-link sites that were found within the first 20 results. These were sites that when the links shown were selected, the browsers either displayed an error stating that the requested site could not be found, or that the requested site could not be loaded. The participants were instructed not to answer this question. After the study, the researcher accessed the histories of each computer used to observe the search results and determine the number of inoperative or dead-link sites the participants encountered. The possible answers for that question ranged from a score of zero for no inoperative or dead-link sites encountered to a score of five, if five or more sites were found. Accordingly, a score of one was tallied for one site; two for two sites; three for three sites; and four, if four inoperative or dead-link sites were encountered. Table 3 illustrates the data used to address hypothesis number 3.

**Table 3: Number of Inoperative or 'Dead Links" found in the Top 20 Returns from Google Search**

| Score | No Returns 0 | 1 | 2 | 3 | 4 | Five or More Returns 5 | Total Score | Average Score | Percent |
|---|---|---|---|---|---|---|---|---|---|
| Responses | 50 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.0% |

As table 3 indicates, the researcher found no inoperative or dead-link sites. Every site that Google Search displayed in the top 20 returns successfully connected to the requested site. This may be attributed to the relevancy ranking software's efficiency, the method of database management on Google's part, and the nature of the search itself. If the relevancy ranking software ranks sites and displays them according to whether the site had previously been visited during a similar search, the chance that a site would still be operative would be much higher than if the site was selected and displayed simply because the search terms appeared in the keyword tag of the HTML heading or in the text. It is apparent no inoperable sites were encountered; therefore, hypothesis number 4 is rejected.

Hypothesis number 4 stated, "Students using a hierarchically indexed web directory when searching for information on the World Wide Web will encounter inoperative sites within the list of sites related to the search." Question number 3 in the Google Directory section of the search log dealt directly with the number of inoperative or dead-link sites that were found within the first 20 results. These were sites that when the links shown were selected, the browsers either displayed an error stating the requested site could not be found, or that the requested site could not be loaded. The participants were instructed not to answer these questions. After the study, the researcher accessed the histories of each computer used to observe the search results and determine the number of inoperative or dead-link sites the participants encountered. The possible answers for that question ranged from a score of zero for no inoperative or dead-link sites encountered to a score of five, if five or more sites were found. Accordingly, a score of one was tallied for one site; two for two sites; three for three sites; and four, if four inoperative or dead-link sites were encountered. Table 4 illustrates the data used to address hypothesis number 4.

**Table 4: Number of Inoperative or 'Dead Links" found in the Top 20 Returns from Google Directory**

| Score | No Returns<br>0 | 1 | 2 | 3 | 4 | Five or More Returns<br>5 | Total Score | Average Score | Percent |
|---|---|---|---|---|---|---|---|---|---|
| Responses | 50 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.0% |

As table 4 indicates, the researcher found no inoperative or dead-link sites. Every site Google Directory displayed in the top 20 returns successfully connected to the requested site. While this result seems to mirror Table 3, there are different reasons for the lack of inoperable sites. Google Directory's efficiency can only be the result of the decision to place a particular site in a category and the maintenance of the database. This indicates that the method Google uses to catalogue sites by topic, and then maintain those sites, shows accuracy in topic matching and efficiency in database management. At the time of the study, the researcher was not privileged to the specifics in the methods that Google Directory used to maintain and catalogue the database. Whether checked by automated means or by human labor, it is evident, considering the lack of inoperable sites, that Google Directory is very efficient. No inoperable sites were encountered, therefore, hypothesis number 4 is rejected.

Hypothesis number 5 stated, "Students will find that a text-based search engine will provide a slower return of information than a hierarchical indexed web directory when searching for information on the World Wide Web." Question number 7 in the Google Search and Google Directory sections of the search log dealt directly with the overall speed the information was found as perceived by the participants. The possible answers to this question range from a score of 1 for a response of very fast to a score of 5 for a response of very slow. Scores of 2, 3, and 4 were included to give participants midrange options. Table 5 illustrates the data used to address hypothesis number 5.

**Table 5: Participant Perception of Overall Speed of Searching**

| Score | Very Fast 1 | 2 | 3 | 4 | Very Slow 5 | Total Score | Average Score |
|---|---|---|---|---|---|---|---|
| Google Search | 18 | 17 | 11 | 4 | 0 | 101 | 2.02 |
| Google Directory | 0 | 0 | 0 | 16 | 34 | 234 | 4.68 |

As table 5 indicates, participants found Google Search nearly twice as fast as Google Directory. Participants perceived that Google Search was very fast in 18 out of 50 responses, or 36% of the time. Looking at a score of 3 or less when 3 indicates the median between 5 points (very slow), and 1 point (very fast), there were 46 out of 50 responses or 92% of the responses. Only 8% thought any of the searching fell in the 4 point range, and none felt that it was very slow with a score of 5 points. Considering the Google Directory searches, the table indicates that the participants decidedly felt the searches were slow to very slow with 16 responses, or 32%, in the 4 point range, and 34 responses, or 68%, in the very slow category. This totals 100% in either slow or very slow perception. It is apparent the majority of participants perceived Google Search to be fast, to very fast, and the majority perceived Google Directory to be slow, to very slow, therefore, hypothesis number 5 is rejected.

Hypothesis number 6 stated, "Students will find that when searching for information on the World Wide Web a hierarchically indexed web directory will be less difficult than using a text-based search engine." Question number 6 in the Google Search and Google Directory sections of the search log dealt directly with the overall difficulty of finding the information as perceived by the participants. The possible answers to this question range from a score of 1 for a response of very easy, to a score of 5 for a response of very difficult. Scores of 2, 3, and 4 were included to give participants midrange options. Table 6 illustrates the data used to address hypothesis number 5.

**Table 6: Participant Perception of Overall Difficulty of Searching**

| Score | Very Easy 1 | 2 | 3 | 4 | Very Difficult 5 | Total Score | Average Score |
|---|---|---|---|---|---|---|---|
| Google Search | 3 | 30 | 9 | 6 | 2 | 124 | 2.48 |
| Google Directory | 1 | 2 | 6 | 14 | 26 | 209 | 4.18 |

As table 6 indicates, participants found Google Search between 2 and 3 times less difficult than Google Directory. Participants perceived that Google Search was very easy in 3 out of 50 responses, or 6% of the time. They found the searches easy with a score of two, 30 times out of 50, or 60% of the time. The median score was selected 9 times, or 18% of the time. Participants found Google Search difficult with a score of four, 6 times, or 12% of the time and very difficult 2 times, with a score of 5, 4% of the time. The table indicates participants perceived that Google Directory was not as easy to use as Google Search. Participants perceived that Google Directory was very easy only 1 time out of 50 responses, or 2% of the time. They found the searches easy with a score of two, 2 times out of 50, or 4% of the time. The median score was selected 6 times, or 12% of the time. Participants found Google Directory difficult with a score of four, 14 times, or 28% of the time and very difficult 26 times, with a score of 5, 52% of the time. It is apparent the majority of participants perceived Google Search to be fairly easy, and the majority perceived Google Directory to be difficult, to very difficult; therefore, hypothesis number 6 is rejected.

# Chapter Five

## Summary, Conclusions and Recommendations

### *Summary*

In the spring of 2005, twenty 11th grade students from north central Iowa were given 5 topics to search on the Internet. All students were allowed 5 minutes to search for each of 5 topics and used identical equipment to perform the searches. The population consisted of 10 male and 10 female students. As there is a marked lack of ethnic and cultural diversity in the community, all students were Caucasian with English being their first language. All students were of similar socio-economic class since there is little variance in the community. Students were chosen based primarily on a range of grade point averages. Students who were selected to participate in the study had a cumulative grade point average of between 2.5 and 3.5 for the current school year. Students who participated had a working knowledge of the hardware and software being used.

The participants were divided into 2 groups. The first group consisted of 5 boys and 5 girls who searched for the 5 topics using Google, a text based search engine utilizing keywords to locate related websites. The second group consisted of 5 boys and 5 girls who searched for the 5 topics using Google Directory, a hierarchically organized directory of websites catalogued by topic. Since the directory search was new to most of the participants, a short lesson was given to everyone on the basic concept and usage of hierarchically organized directories. Each group was given 5 minutes to search for each of the 5 topics.

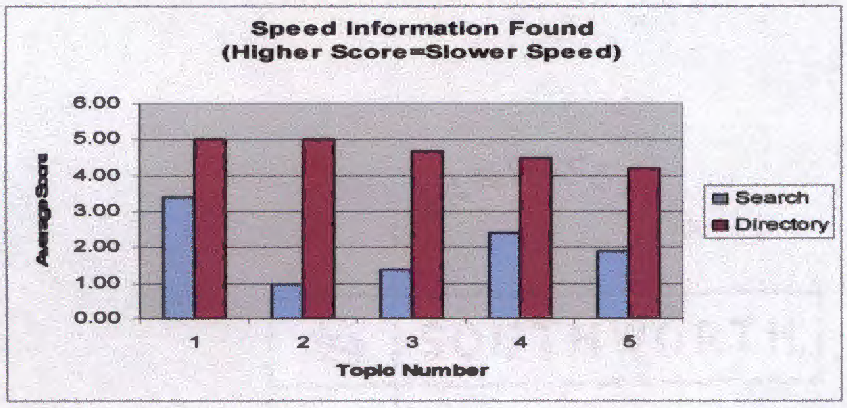
A search log kept by each student asked questions relating to the participant's experience with the search of each topic. The search log allowed participants to rate each search according to overall difficulty, difficulty determining keywords or search path, difficulty finding information, and the speed that the information was found. The log also contained items that the researcher addressed by analyzing the search histories of each participant to determine the number of inappropriate sites found, the number of non-functional sites found, and the number of sites found unrelated to the search. Only the top 20 sites returned at the end of a search were considered.

### *Conclusions and Recommendations*

Upon completion of the study, it was determined that neither type of search method encountered non-functional sites within the first 20 returned, and irrelevant websites were kept to a minimum of 19.5% for Google Directory and a minimum of 21.2% for Google Search. The participants, after considering overall difficulty, determining keywords or search path, difficulty finding information, and the speed that the information was found, showed a definite preference towards the keyword method. Results were consistent with the participants preferring keyword searching over the directory in all instances.

Several factors contributed to the participants' preference towards keyword searching. A large part of the results may be attributed to the fact that keyword searching was by far the more familiar of the search methods. None of the participants had used a hierarchically structured directory prior to the study. This lack of experience and other factors probably contributed to the results illustrated in Chart 1.

**Chart 1: Participant Perception of Overall Speed of Searching**

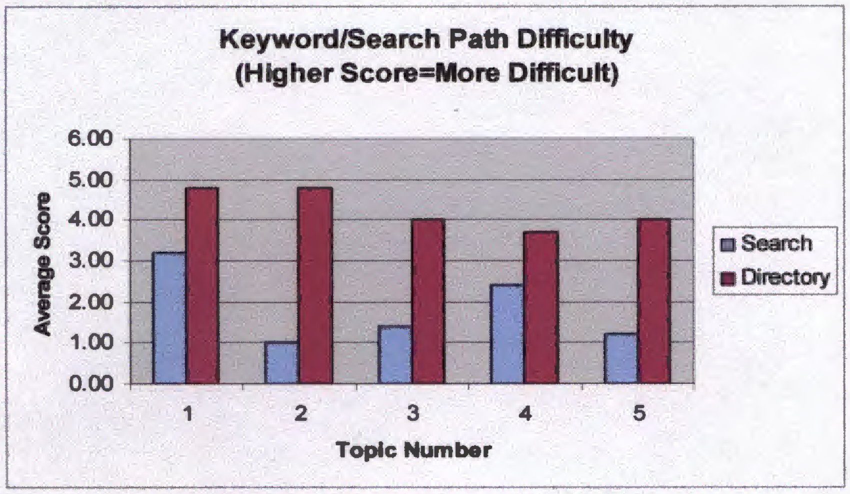


This unfamiliarity may have contributed to a bias towards keyword searching and certainly the lack of experience probably made it difficult for the participants to focus solely on the search without concern about the method. Overwhelmingly, the participants felt that keyword searching was faster. But, it could be argued that this may be because of a combination of their familiarity of the method, lack of comfort with Google Directory, and stress induced by the time limit imposed on the search.

This study should be repeated with participants who had gone beyond a brief introduction of an hierarchical directory so they are more comfortable and practiced in its use and concept with the same students searching for a topic using both search methods as a comparison.

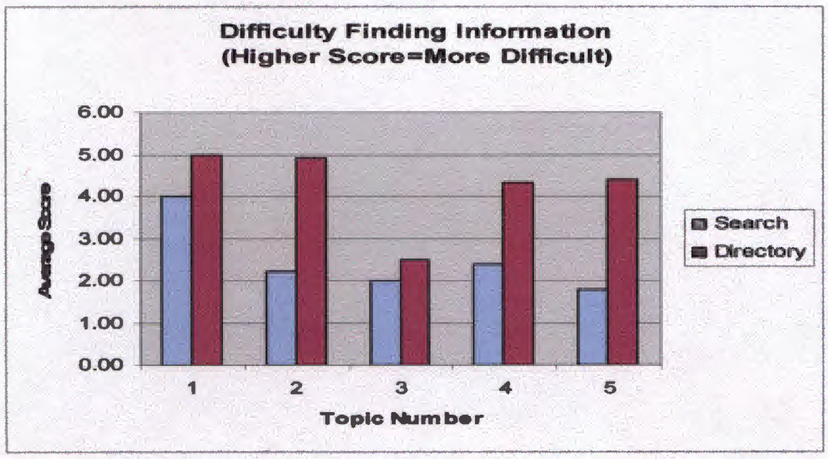
Another important contributing factor in this study was the participants' lack of knowledge on the topics chosen. The topics provided were of a sort, and worded in such a way, that unless a person had some background knowledge on the topic, it would be difficult to follow a search path in a hierarchical directory. This may have been a contributing factor to the participants' perception of difficulty as evidenced in Chart 2 below. The issue of lack of pre-searching and knowledge about a topic is of major concern. While this was apparent in students' difficulty in hierarchical searching in this study, it may also be of concern in students' ability to choose sites wisely in keyword searching.

**Chart 2: Participant Perception of Overall Difficulty of Determining Keywords or a Directory Search Path**



Keyword searching, on the other hand, allowed the participants to type in search terms that they may have been unfamiliar with but would still be able to get search returns on. An example would have been the question about the capture of Antwerp. It was worded as follows: "What about the Allies' capture of Antwerp in 1944 turned out to be a disaster?" The only clues that would lead someone unfamiliar with this area of history to World War 2 would be the date 1944 and the word 'Allies.' If these two bits of information wouldn't lead a searcher to the correct path, they would be lost. As observed, the students did not make this connection. A keyword search by comparison would easily return several sites using any combination of search terms such as Allies, Antwerp, 1944, or capture. Without a doubt, keyword searching is superior to a hierarchically structured directory when the user lacks background knowledge on a topic. Even with students who were unfamiliar with hierarchically structured directories prior to the study, when a subject that was familiar to the participants was searched the results were much different. Looking at Chart 3, the average score for difficulty was very close between the two search methods for topic number 3. This was the question regarding the Cubist movement in art. Since participants had some prior knowledge on the subject and were able to relate the term 'cubist' to art, they found a relatively unbroken path to 'Cubists' using the hierarchical directory. Chart 3 also illustrates that too vague or broad of a question can create a situation where both search methods are very difficult. Topic number 1 was the question that read, "Explain the good and bad effects of Saddam's rule in Iraq." This question is so subjective that it is difficult to locate in a hierarchical directory, and keyword searching is difficult because of the copious amount of information returned with any combination of search terms. Thus, the elevated difficulty score for both methods.

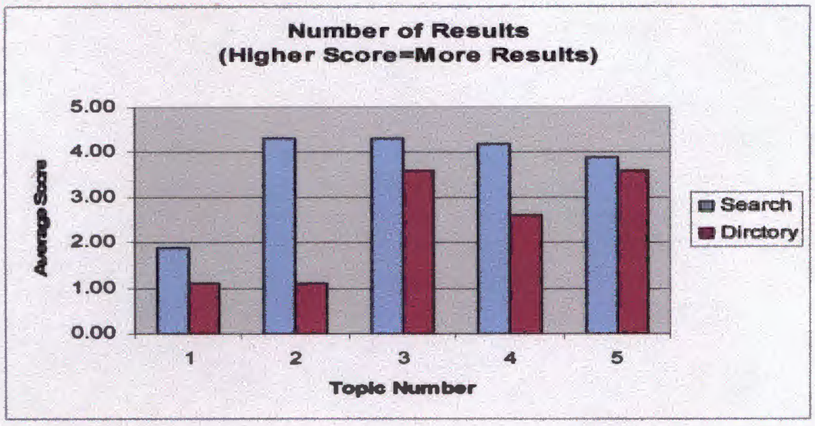
**Chart 3: Perception of Overall Difficulty of Searching**



This study should be repeated with topics that are more familiar to the participants. This would replicate a situation where students have been introduced to a subject or been exposed to some general information prior to an assignment to gather more information. In the above example, they may have started a unit on World War 2 in Europe with an assignment regarding various Allied offensives. This would give them enough background to follow a logical search path in a hierarchical directory in order to find some information of value.

The data collected regarding the number of results tended to be misleading for two reasons. The number of results that were considered for relevancy and functionality consisted of only the top 20 sites returned. Of course, a keyword search would nearly always return at least 20 sites and usually many thousands of varying relevance. A hierarchical directory, on the other hand, would return fewer sites as it would display only sites that have been cataloged either by human decision or by technological means. Because of this, results as shown in Chart 4 only show that keyword searching returns more total sites and says nothing about the value of the information of those sites.

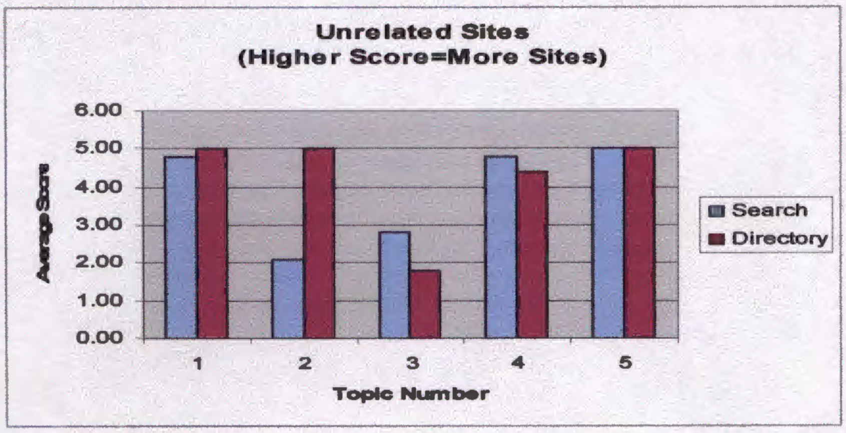
**Chart 4: Total Number of Sites Returned**



If a similar study is done in the future, the total number of sites returned by either of the two search methods should be disregarded since the information it provides is irrelevant to the efficiency of the searching. For example, a student searches for a topic that they understand or have some prior knowledge of. If they conduct a successful search with an hierarchical directory it is likely that they will find anywhere from a few to a few dozen sites that would be of some value. A similar search for the same topic using a keyword search browser may result in hundreds of thousands of total returns, however, if relevancy technology is used, the student would typically explore only the first 10 or 20 of those sites. This has an effect of equaling out the number of results since no one would take the time to check thousands of sites. Possibly a better way to approach this particular question would be to allow participant anecdotal data as to their feelings on the amount of relevant information provided by each method.

Search time was another area where, if another study is done, the hierarchical directory may fare better. In several instances, the participants simply ran out of time trying to find the right search path. Therefore, the number of irrelevant sites appeared higher since, at the end of the search, they had not located any usable sites. Chart 5 shows that both search methods had a fair number of irrelevant sites (19.5% for Google Search and 21.2% for Google Directory), but for different reasons.

**Chart 5: Unrelated Sites Returned in the Top 20 Results**



The main reason for unrelated sites returned by Google Search was poor choice of search terms on the part of the user or the order the search terms appeared in the request. By contrast, the main reason for unrelated sites returned by Google Directory was either lack of knowledge on the user's part, or lack of adequate time for the searcher to locate the correct search path. In cases where the correct path was followed to an end, the number of irrelevant sites declined dramatically. As in the case of topic number 3 regarding the Cubist movement, where the participants had knowledge enough and adequate time to follow a logical path to an end, the number of irrelevant sites was minimal.

The relevancy ranking method of Google Search must be commended. Even though most of the topics searched were difficult or unfamiliar to the participants, the quality of the sites returned and the grouping of those sites made it relatively simple for the participants to locate enough information to gain an informed opinion on the subject. During analysis, the researcher had to go 3 or 4 pages deep, 30 to 40 returned sites, before encountering sites that were considered completely unrelated to the topic. For the terms and topics searched for this study, the top 20 sites returned by Google Search were either relevant or at least related to the topic being searched. The relevancy ranking software may track a particular site's visit history and compare the current search request with requests in the past, thus giving a higher relevancy to sites visited by others making similar searches. This would increase the chances that only functional sites would be found in the first several listings. The other factor that may contribute to a lack of non-functional sites is the method used by Google to delete dead-links, whether automated, human staff or user initiated. If Google is maintaining a high degree of effort or efficiency in identifying and deleting dead-links, the chance of a user encountering one would be diminished. While this is no guarantee that every search for every topic will return only valid links, (in fact, the researcher did several random searches to make sure that dead links do occur), it does make a powerful and positive statement about Google's ability to rank and maintain its huge base of web sites.

As discussed in Chapter 4, a related site was one that may not have contained specific information about the question asked but was of some value generally, and if viewed, may have led to more specific sites. As an example, topic number 2 asked, "Describe the ethical issues in organ donations." Many of the sites returned using the search term 'organ donations ethics' were the homepages of organ donation organizations which were biased on the pro side of the issue and mainly gave information about how one would go about making a donation, the legalities involved, and the positive results to a recipient. While these sites didn't discuss the ethics involved, many did provide links to other Internet sites that did provide good information from both sides of the debate.

There are areas where a directory may be more suitable for information searching. Perhaps a school librarian working with teachers could catalog the library's relevant holdings and other resources for specific projects so students could quickly find information on a topic and not waste time. A directory for a social studies project on the Civil War could include audio, video, and print sources in the library as well as multimedia articles, online databases like EBSCO, and Internet sites that are grouped by topic. Then students could look for the Civil War directory and then select their topic of interest, possibly slavery, and proceed to a more specific branch of the directory that discusses slavery as a cause of the war; and then, view available sources of information or jump to online sources from the directory. This would help students stay with approved or selected information as determined by the instructor and avoid sifting through material that would not relate to their research, such as fictional accounts and stories also found in the library. Of course, all of this would require time and effort on the part of the librarian and teacher to sort, catalog, and maintain the directory.

While observing the participants making their searches and looking at the data from the study that pertains to the perceived difficulty of the search methods, one thing did become very clear. With these high school students, the extra effort required by Google Directory in determining and following a search path through knowledge and/or logic, was no match for the speed and ease of typing in two or three search terms and having Google Search jump back with a plethora of sites ranked by relevancy in just a second or two. Regardless of the quality of sites returned, or whether or not time was wasted with irrelevant sites, the participants preferred the speed and ease of keyword searching.

What would be the cause of this preference? Possibly the "path of least resistance theory". It is inherent in nature that all things tend to take the path of least resistance. A water filter system works on this premise, water is directed through filters by resistance to trap impurities, flowers grow towards the sun and away from shade, and vines grow near an object that they can climb. People too, in most endeavors choose the easy path in most instances. Most of us climb a mountain following a trail instead of going up a cliff, and if there is a way to search for information on the Internet that we perceive as faster or easier, most of us will chose to use it. Both search methods evaluated in this study returned valid and accurate information, (sometimes the same sites were found). There is a place were both methods show their individual strengths. It would appear from this study that the participants felt Google Search and keyword searching is easier and faster to use, especially if they have little knowledge on a subject. However, Google Directory could be a valuable tool for upper level research intended to stretch deductive reasoning ability or for educators to use the directory to narrow the focus of a particular subject or project so students don't waste time being sidetracked to sites not directly related to the material in need. A keyword search for Abraham Lincoln would provide so much information one may not know where to start. But a narrower path through a directory would yield a better chance of focusing on a particular aspect of his life.

# Bibliography

Berners-Lee, T. *The World Wide Web: Past, present and future*. (August 1996). Retrieved on (February, 14, 2003) from http://www/w3/org/People/Berners-Lee/1996/ppf/html.

Bruza, P., McArthur, R., & Dennis, S. (2000). Interactive internet search: Keyword, directory and query reformulation mechanisms compared. *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, July 2000, New York, 280-7.

Carnahan, K. (1998). *Official Microsoft bookshelf internet directory*. Redmond, Washington: Microsoft Press.

Chau, M., Zeng, D., Chen, H. (2001). *Personalized spiders for web search and Analysis* (Report No. IR 058 355). MI: National Science Foundation. (ERIC Document Reproduction Service No. ED 459 819).

Chen, H., & Dumais, S. (2000). Hierarchical classification of web content. *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, July 2000, New York, 256-63.

Clyde, A., Search engines, *Teacher Librarian*, April 2000, Vol. 27, Issue 4.

Feldman, S., Just the answers, please: choosing a web search service, *Searcher*, May 1997, Vol. 5, (5).

Flanigan, N. M. (2001). *A better way to search the internet*. Unpublished Master's research paper, University of Northern Iowa.

Fleming, D., Demystifying HTML , *Training & Development*, September 1997, Vol. 51(9), 48-51.

Gowan, M., Spanbauer, S., Find everything faster. *PC World*, Sept 2001, Vol. 19(9), 109-16.0

Gromov, G. R., *History of Internet and WWW: The roads and crossroads of internet history*. (2002). Retrieved on February, 18, 2003 from http://www.internetvalley.com/intvall/html

*Hierarchical index*, (2003). Retrieved on March 10, 2003 from *Webopedia*, http://www.webopedia.com.

*Hit*, Retrieved on March 10, 2003 from *Webopedia*, http://www.webopedia.com.

Hubbard, J., *Indexing the Internet.* (December, 1999). Retrieved on (March 3, 2003) from http://www.tk421.net/essays/babel.shtml.

Howe, W., *A Brief History of the Internet.* (2000). Retrieved on February, 18, 2003 from http://www.walthowe.com.

*Internet.* Retrieved March 10, 2003, from *Encyclopædia Britannica Online.* http://0-search.eb.com.unistar.uni.edu:80/eb/article?eu=1460.

Introna, L, D., Shaping the web: why the politics of search engines matters, *Information Society,* July-September 2000, Vol. 16(3).

Margolis, P.E. (1999). *Random House Webster's Computer & Internet Dictionary.* Random House: New York, NY.

Minkel, W. (2000). What students know before they go online matters, *School Library Journal,* August, 2000, Vol. 46(8), 22

Pealer, L. N., Using search engines and web directories, *Journal of School Health,* October 1998, Vol. 68(8).

Portal, Retrieved on March 10, 2003 from *Webopedia,* http://www.webopedia.com.

URI, Retrieved on March 10, 2003 from *Webopedia,* http://www.webopedia.com.

Vansickle, S., Tenth grader's search knowledge and use of the Web, *Knowledge Quest,* March/April, 2002, Vol 30(4), 33-7.

Zakon, R. H., Hobbes' Internet timeline v6.0, February, 2003. Retrieved on (March 3, 2003) from http://www.zakon.org/robert/internet/timeline.

# APPENDIX A
## Computer and Internet Experience Survey

**Please circle your choice.**

Do you use a computer to find information on the Internet? Yes No

Do you use a computer to find information on the Internet at home? Yes No

What is currently your favorite way to search the Internet? ____________________

**Please circle the answer that best describes you.**

How often do you use the Internet or World Wide Web?

One day per week or less 2 or 3 times per week

4 to 6 times per week Every day

How comfortable are you about using a computer?

Very comfortable Some what comfortable

Some what uncomfortable Not comfortable at all

How would you rate your level of competency with computers?

I am kind of an expert and can teach others how to do things.

I am able to do most everything or at least figure it out myself.

I can do most things as long as someone shows me how the first time.

I can do some things on my own but need help once in a while.

I have trouble with most things and need a lot of help

## APPENDIX B
### Human Participants Review-Informed Consent/Assent

Keyword or Hierarchical Searching? A quantitative comparison of World Wide Web searching methods for high school juniors in North Iowa

Principal investigator; Harold K. Price, Media Specialist
Rudd, Rockford, Marble Rock Community School District

Some 11th grade students will be invited to participate in a research project conducted through the University of Northern Iowa. If you are a student who is 18 years of age or older the University requires that you give your signed agreement to participate in this project. If you are a student that is under 18 years of age the University requires that you and your parent or guardian both give your signed agreement to participate in this project. The following information is provided to help you make an informed decision whether or not to participate.

This project will have 11th grade students, whose grades are within a specific range, research various topics on the World Wide Web using one of two search methods for the purpose of comparing the efficiency of the two methods. The topics will be academic in nature and will include such things as locating specific answers to questions, general information about a broad topic and current events. Essentially, this study is attempting to determine if a keyword type search for information on the Internet or a an organized listing, called 'hierarchal' searching will be faster or more efficient in students obtaining the information they are looking for.

Students will first attend an informational seminar on search methods before the study. The seminar will be conducted on a separate day from the study and will take approximately 60 minutes. The study itself will require students to search the Internet using one of the two search methods for information on topics that will be provided to them by an instructor from the University of Northern Iowa. Search results will be timed and supervised and all results will be tracked. The study should last approximately 60 minutes. The data collected will be calculated to determine the efficiency of the two searching methods comparing the time involved in the search, the number of irrelevant sites returned and the participant's opinion on the quality of information found.

Risks involved in participation in the study are minimal and similar to a normal day at school. It will involve the student getting to and from the computer lab located in the Media Center of the Rudd, Rockford, Marble Rock High School and any normal risks related to computer use including eyestrain, physical discomfort from sitting. Any student whose parents have previously signed a permission to use the Internet agreement with the district is eligible.

Participation in this study will offer no direct rewards for the student. This includes physical, academic, monetary or any type of gift rewards for the student. Students who choose to participate may learn new or more efficient ways to locate information for future use in their academic or professional pursuits. The choice not to participate will have no affect on any grade or status for the student at the Rudd, Rockford, Marble Rock Community School District.

Participants will be given a unique identification number and no student will be identified by name. The only personal references in the study will be grade level and gender. Information obtained during this study which could identify the participant will be kept strictly confidential. The information may be published in an academic journal or presented at a scholarly conference.

Participation is completely voluntary. The participant is free to withdraw from the study at any time or to choose not to participated at all, and by doing so, they will not be penalized or lose benefits to which they are otherwise entitled.

If you have questions about the study, or desire information in the future regarding participation in the study, or general information about the study, you can contact Harold Price at Rudd, Rockford, Marble Rock Community Schools by phone at 641-756-3508 or at home by phone at 641-424-5176. You may also contact the project's faculty advisor, Barbara Safford at the Department of Library Science, University of Northern Iowa, by phone at 319-273-2551. You can also contact the office of the Human Participants Coordinator, University of Northern Iowa, at 319-273-2748 for answers to questions about right of research participants.

Sincerely,

Harold Price, Media Specialist
Rudd, Rockford, Marble Rock Community School District

**Agreement:**

**If the student is 18 years old or older complete section A. If the student is under 18 years old, complete section B.**

**Section A; (18 or older)**

I am fully aware of the nature and extent of my participation in this project as stated above and the possible risks arising from it. I hereby agree to participate in this project. I acknowledge that I have received a copy of this consent statement. I am 18 years of age or older.

(Signature of participant) (Date) (GPA)

(Printed name of participant) (Date of Birth)

**Section B; (Under 18 years old)**

I am under 18 years of age and I have signed this consent form with full knowledge of my parent or legal guardian.

(Signature of participant) (Date) (GPA)

(Printed name of participant) (Date of Birth)

I the parent or legal guardian of the participant am fully aware of the nature and extent of my child's participation in this project as stated above and the possible risks arising from it. I hereby give permission for their participation in this project. I acknowledge that I have received a copy of this consent statement.

(Signature of parent or guardian) (Date)

(Printed name of parent or guardian)

(Signature of investigator) (Date)

(Signature of instructor/advisor) (Date)

## Appendix C

## Search Log

Search Method Used: Google Search Google Directory
(Circle)

Topic Searched: ________________________________

Total results: ________________

Rank the following considering the top 20 results.

***Google Search*** Very Easy Very
Difficult
Difficulty determining keywords. 1 2 3 4 5
None Too
Many
Number of results. 1 2 3 4 5
None More
than 4
Non-functional sites found. 0 1 2 3 4 5
None More
than 4
Inappropriate sites found. 0 1 2 3 4 5
None More
than 4
Sites found unrelated to the search. 0 1 2 3 4 5
Very Easy
Very Difficult
Overall difficulty finding information. 1 2 3 4 5
Very Fast
Very Slow
Overall speed the information was found. 1 2 3 4 5

***Google Directory*** Very Easy
Very Difficult
Difficulty determining search path. 1 2 3 4 5
None Too
Many
Number of results. 1 2 3 4 5
None More
than 4
Non-functional sites found. 0 1 2 3 4 5
None More
than 4
Inappropriate sites found. 0 1 2 3 4 5

None More
than 4
Sites found unrelated to the search. 0 1 2 3 4 5
Very Easy
Very Difficult
Overall difficulty finding information. 1 2 3 4 5
Very Fast
Very Slow
Overall speed the information was found. 1 2 3 4 5

## Appendix D

### Lesson Plan
### Internet Searching

Purpose:
To introduce or review methods of Internet searching using keyword-based search engines and hierarchically cataloged web directories.

Introduction:
Describe searching in general.
- Overview of Internet and searching.
- Define 'keyword'.
- Look at example of a search and point out specific areas of the results.
  - Title
  - Description
  - URL
- Talk about ads and how to avoid accidentally going to misleading sites.
- Discuss dead sites, and irrelevant or inappropriate sites.

Discuss how general searchers work.
- Explain how spiders or bots catalog sites.
- Uses keyword or phrase.
- Matches keyword or phrase to text in WWW page.
  - Some only search meta fields like title, description and keyword.
  - Others search all text within the page.
- Differentiate between the use of, any of the words, all of the words and exact phrase.
- Displays results ranked by 'relevancy'.
- Show example of search and results.

Discuss how web directories work.

- Explain what hierarchically cataloged means.
- Explain how human reviewers and catalogers organize sites.
- Uses a search path, (show how to follow a path from broad to narrow.)
  - Start with a topic.
  - Think about what general or broad area the topic belongs in.
  - From the broad area select the next level towards the topic.
  - Continue until topic is reached.
- Displays results with only sites that pertain to the topic search for.

Conclusion:
Point out that the two search methods require different thought processes. Compare the differences in search method and results.

## Appendix E

## Data Set

The following data was collected in accordance with the research as described in Chapter 3: Methodology. Each of the 5 topics were searched for by 10 participants using Google Search, a text-based Internet search engine, and each of the 5 topics were also searched for by 10 participants using Google Directory, a hierarchically indexed site directory. In the tables listed the question numbers listed correspond to the questions on the search log that each participant answered. Therefore, the first question on the search log under the Google Search section corresponds to the item in the table listed as question 1 just as the first question on the search log under the Google Directory section corresponds to the item in the table listed as question 1. Below is a list of the questions as they appeared on the search log and their corresponding question numbers in the tables.

Google Search
Question 1 Difficulty determining keywords Ranked from very east to very difficult

Question 2 Number of results. Ranked from none to too many
Question 3 Non-functional sites found Ranked from none to more than 5
Question 4 Inappropriate sites found Ranked from none to more than 5
Question 5 Sites found unrelated to the search Ranked from none to more than 5
Question 6 Overall difficulty finding information Ranked from very easy to very difficult
Question 7 Overall speed the information was found Ranked from very easy to very difficult

Google Directory
Question 1 Difficulty determining search path Ranked from very east to very difficult
Question 2 Number of results. Ranked from none to too many
Question 3 Non-functional sites found Ranked from none to more than 5
Question 4 Inappropriate sites found Ranked from none to more than 5
Question 5 Sites found unrelated to the search Ranked from none to more than 5
Question 6 Overall difficulty finding information Ranked from very easy to very difficult
Question 7 Overall speed the information was found Ranked from very easy to

very difficult

**Topic #1** **Explain the good and bad effects of Saddam's rule in Iraq.**

| Google Search | 0 | 1 | 2 | 3 | 4 | 5 | Score | Responses | Mean | Mode | Median |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Question 1 | | | 1 | 6 | 3 | | 32 | 10 | 3.20 | 3.00 | 3.00 |
| Question 2 | | 2 | 7 | 1 | | | 19 | 10 | 1.90 | 2.00 | 2.00 |
| Question 3 | 10 | | | | | | 0 | 10 | 0.00 | 0.00 | 0.00 |
| Question 4 | 10 | | | | | | 0 | 10 | 0.00 | 0.00 | 0.00 |
| Question 5 | | | | | 2 | 8 | 48 | 10 | 4.80 | 5.00 | 5.00 |
| Question 6 | | | | 2 | 6 | 2 | 40 | 10 | 4.00 | 4.00 | 4.00 |
| Question 7 | | | | 6 | 4 | | 34 | 10 | 3.40 | 3.00 | 3.00 |

| Google Directory | 0 | 1 | 2 | 3 | 4 | 5 | Score | Responses | Mean | Mode | Median |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Question 1 | | | | | 2 | 8 | 48 | 10 | 4.80 | 5.00 | 5.00 |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Question 2 | | 9 | 1 | | | | 11 | 10 | 1.10 | 1.00 | 1.00 |
| Question 3 | 10 | | | | | | 0 | 10 | 0.00 | 0.00 | 0.00 |
| Question 4 | 10 | | | | | | 0 | 10 | 0.00 | 0.00 | 0.00 |
| Question 5 | | | | | | 10 | 50 | 10 | 5.00 | 5.00 | 5.00 |
| Question 6 | | | | | | 10 | 50 | 10 | 5.00 | 5.00 | 5.00 |
| Question 7 | | | | | | 10 | 50 | 10 | 5.00 | 5.00 | 5.00 |

**Topic #2 Describe the ethical issues in organ donations.**

| Google<br>Search | 0 | 1 | 2 | 3 | 4 | 5 | Score | Responses | Mean | Mode | Median |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Question 1 | | 10 | | | | | 10 | 10 | 1.00 | 1.00 | 1.00 |
| Question 2 | | | | | 7 | 3 | 43 | 10 | 4.30 | 4.00 | 4.00 |
| Question 3 | 10 | | | | | | 0 | 10 | 0.00 | 0.00 | 0.00 |
| Question 4 | 10 | | | | | | 0 | 10 | 0.00 | 0.00 | 0.00 |
| Question 5 | | | 9 | 1 | | | 21 | 10 | 2.10 | 2.00 | 2.00 |
| Question 6 | | | 8 | 2 | | | 22 | 10 | 2.20 | 2.00 | 2.00 |
| Question 7 | | 10 | | | | | 10 | 10 | 1.00 | 1.00 | 1.00 |

| Google<br>Directory | 0 | 1 | 2 | 3 | 4 | 5 | Score | Responses | Mean | Mode | Median |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Question 1 | | | | | 2 | 8 | 48 | 10 | 4.80 | 5.00 | 5.00 |
| Question 2 | | 9 | 1 | | | | 11 | 10 | 1.10 | 1.00 | 1.00 |
| Question 3 | 10 | | | | | | 0 | 10 | 0.00 | 0.00 | 0.00 |
| Question 4 | 10 | | | | | | 0 | 10 | 0.00 | 0.00 | 0.00 |
| Question 5 | | | | | | 10 | 50 | 10 | 5.00 | 5.00 | 5.00 |
| Question 6 | | | | | 1 | 9 | 49 | 10 | 4.90 | 5.00 | 5.00 |
| Question 7 | | | | | | 10 | 50 | 10 | 5.00 | 5.00 | 5.00 |

**Topic #3 What were the cubists trying to show?**

| Google | | | | | | | | | Mea | Mod | Media |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Search | 0 | 1 | 2 | 3 | 4 | 5 | Score | Responses | n | e | n |
| Question 1 | | 7 | 2 | 1 | | | 14 | 10 | 1.40 | 1.00 | 1.00 |
| Question 2 | | | | | 7 | 3 | 43 | 10 | 4.30 | 4.00 | 4.00 |
| Question 3 | 10 | | | | | | 0 | 10 | 0.00 | 0.00 | 0.00 |
| Question 4 | 10 | | | | | | 0 | 10 | 0.00 | 0.00 | 0.00 |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Question 5 | | | 2 | 8 | | | 28 | 10 | 2.80 | 5.00 | 5.00 |
| Question 6 | | 1 | 8 | 1 | | | 20 | 10 | 2.00 | 2.00 | 2.00 |
| Question 7 | | 6 | 4 | | | | 14 | 10 | 1.40 | 1.00 | 1.00 |

| Google Directory | 0 | 1 | 2 | 3 | 4 | 5 | Score | Responses | Mean | Mode | Median |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Question 1 | | | | 2 | 6 | 2 | 40 | 10 | 4.00 | 4.00 | 4.00 |
| Question 2 | | | | 4 | 6 | | 36 | 10 | 3.60 | 4.00 | 4.00 |
| Question 3 | 10 | | | | | | 0 | 10 | 0.00 | 0.00 | 0.00 |
| Question 4 | 10 | | | | | | 0 | 10 | 0.00 | 0.00 | 0.00 |
| Question 5 | | 2 | 8 | | | | 18 | 10 | 1.80 | 2.00 | 2.00 |
| Question 6 | | 1 | 3 | 6 | | | 25 | 10 | 2.50 | 3.00 | 3.00 |
| Question 7 | | | | | 3 | 7 | 47 | 10 | 4.70 | 5.00 | 5.00 |

**Topic #4** **What about the allies capture of Antwerp in 1944 turned out to be a disaster?**

| Google Search | 0 | 1 | 2 | 3 | 4 | 5 | Score | Responses | Mean | Mode | Median |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Question 1 | | | 7 | 2 | 1 | | 24 | 10 | 2.40 | 2.00 | 2.00 |
| Question 2 | | | | | 8 | 2 | 42 | 10 | 4.20 | 4.00 | 4.00 |
| Question 3 | 10 | | | | | | 0 | 10 | 0.00 | 0.00 | 0.00 |
| Question 4 | 10 | | | | | | 0 | 10 | 0.00 | 0.00 | 0.00 |
| Question 5 | | | | | 2 | 8 | 48 | 10 | 4.80 | 5.00 | 5.00 |
| Question 6 | | | 6 | 4 | | | 24 | 10 | 2.40 | 2.00 | 2.00 |
| Question 7 | | | 6 | 4 | | | 24 | 10 | 2.40 | 2.00 | 2.00 |

| Google Directory | 0 | 1 | 2 | 3 | 4 | 5 | Score | Responses | Mean | Mode | Median |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Question 1 | | | | 4 | 5 | 1 | 37 | 10 | 3.70 | 4.00 | 4.00 |
| Question 2 | | | 4 | 6 | | | 26 | 10 | 2.60 | 3.00 | 3.00 |
| Question 3 | 10 | | | | | | 0 | 10 | 0.00 | 0.00 | 0.00 |
| Question 4 | 10 | | | | | | 0 | 10 | 0.00 | 0.00 | 0.00 |
| Question 5 | | | | 2 | 2 | 6 | 44 | 10 | 4.40 | 5.00 | 5.00 |
| Question 6 | | | | | 7 | 3 | 43 | 10 | 4.30 | 4.00 | 4.00 |
| Question 7 | | | | | 5 | 5 | 45 | 10 | 4.50 | 4.00 | 4.00 |

**Topic #5** **Describe the causes of the Chaco War.**

| Google<br>Search | 0 | 1 | 2 | 3 | 4 | 5 | Score | Responses | Average | Mode | Median |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Question 1 | | 8 | 2 | | | | 12 | 10 | 1.20 | 1.00 | 1.00 |
| Question 2 | | | | 1 | 9 | | 39 | 10 | 3.90 | 4.00 | 4.00 |
| Question 3 | 10 | | | | | | 0 | 10 | 0.00 | 0.00 | 0.00 |
| Question 4 | 10 | | | | | | 0 | 10 | 0.00 | 0.00 | 0.00 |
| Question 5 | | | | | | 10 | 50 | 10 | 5.00 | 5.00 | 5.00 |
| Question 6 | | 2 | 8 | | | | 18 | 10 | 1.80 | 2.00 | 2.00 |
| Question 7 | | 2 | 7 | 1 | | | 19 | 10 | 1.90 | 2.00 | 2.00 |

| Google<br>Directory | 0 | 1 | 2 | 3 | 4 | 5 | Score | Responses | Average | Mode | Median |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Question 1 | | | | 2 | 6 | 2 | 40 | 10 | 4.00 | 4.00 | 4.00 |
| Question 2 | | | | 4 | 6 | | 36 | 10 | 3.60 | 4.00 | 4.00 |
| Question 3 | 10 | | | | | | 0 | 10 | 0.00 | 0.00 | 0.00 |
| Question 4 | 10 | | | | | | 0 | 10 | 0.00 | 0.00 | 0.00 |
| Question 5 | | | | | | 10 | 50 | 10 | 5.00 | 5.00 | 5.00 |
| Question 6 | | | | | 6 | 4 | 44 | 10 | 4.40 | 4.00 | 4.00 |
| Question 7 | | | | | 8 | 2 | 42 | 10 | 4.20 | 4.00 | 4.00 |