

# **Web Portals: How to Edit the Web.**

*Donald S. Pearson, Jr.*

*June 19, 2003*

*MIS 685, E1FF*

## *Table of Contents*

Table of Contents .....	2
Abstract.....	3
The Problem.....	4
Infoglut .....	4
Quality/veracity .....	5
Database Access .....	5
Presentation .....	5
Web Portals: An Introduction .....	6
Relevance to Information Systems .....	6
Literature Review.....	6
Infoglut .....	7
Quality/veracity .....	8
Database Access .....	10
Presentation .....	12
Table I: A Typology of Web Consumers .....	13
Benchmarking Analysis and Sample Results.....	14
Benchmarking a Typical User's Experience.....	15
Summary – The Current State of Affairs.....	15
Search engines and the Benchmarking Report.....	15
Existing Portal Solutions .....	16
No standards.....	17
Conclusion .....	17
Appendix A - Benchmarking Ratings of Search Engines or Portals.....	19
Readability scores .....	20
Findings.....	20
Table II: Search topic: The Death of Lucius Annaeus Seneca .....	22
Chart I: The Number of Sites Returned by the Search Engine.....	23
Chart II: Percentage of Sites with Usable Information having to do with the Search Terms as specified .....	24
Chart III: The Amount of Words to read until the Answer is found.....	25
Chart IV: Flesch Reading Ease Score.....	26
Chart V: Flesch-Kincaid Grade Level Scores .....	27
Appendix B – Digital Object Identifier.....	28
What is a DOI?.....	28
Appendix C – Sponsor's Comments.....	29
Gayle DeGennaro.....	29
Ron Morgan .....	30
Mike Powers .....	31
References .....	32

### *Abstract*

This paper identifies some problems with current methods of web searching and classified them into four types: Infoglut, Quality/Veracity, Database Access and Presentation. It uses a benchmarking study of a search for a topic of general information to quantify and qualify how currently popular search engines handle the same search. The results of the study are used to suggest how search engines may become Web Portals.

Web Portals are defined, as are some of their emerging characteristics. The top vendors are identified as is the ongoing work to move the industry into a truly useful, commonly accessible web portal standard.

“What is the use of having countless books and libraries, whose titles their owners can scarcely read through in a whole lifetime? The learner is not instructed, but burdened by the mass of them, and it is much better to surrender yourself to a few authors than to wander through many.”

– Lucius Annaeus Seneca (3 B.C.E – 65 C.E.), *On Tranquillity of Mind*

### *The Problem*

The Internet has become a popular medium for knowledge transfer. Copious information on almost any form of human endeavor is available on the World Wide Web (WWW), which is built on top of the Internet. (“The Difference Between,” 2003) The range and depth of the WWW is both its power and its primary problem. One current estimate puts the total number of web pages at 3 billion! (Barker, 2001) This leads to several problems one encounters when searching the WWW: Infoglut, Veracity, Access and Presentation.

### Infoglut

Infoglut is not necessarily a modern problem, as evidenced by the above quote of the Roman philosopher Seneca. Nevertheless, the problem exists given the current state of web searching. For example, the search engine Google turned up “about 818,000” results for the word “Seneca,” including information about Seneca College, Seneca Rocks, WV, the Seneca Nation of Indians, the Seneca Park Zoo, Seneca County, New York and Seneca Foods. Only one result of the first ten displayed information about the Roman philosopher Lucius Anneaus Seneca.



Lucius Annaeus Seneca  
(Image from [www.locutio.com](http://www.locutio.com))

### Quality/veracity

This leads to the second problem with web searching – the quality and veracity of the results. Knowing the full name of Seneca and enclosing it in quotes narrows Google's results to a manageable 28 (down from 818,000). However, Google lists as the first result a site from Germantown Academy in Pennsylvania written by Liane S., class of '03, in which Liane writes a fictionalized account of the life of Seneca's wife, Pompeia Paulina. Although impressive, this high school senior's effort is probably not the first source one would like to use to learn about Seneca! To Google's credit, it does have enough intelligence to catch a spelling error in Seneca's name, and suggest a search under "Lucius Annaaeus Seneca," which yields 6,290 hits – back to Infoglut! (As an aside, when the research first began, there were only 5,670 pages on Seneca, the increase in 3 months was 620 new hits, or 11%!)

### Database Access

While much literature is in the public domain, another problem arises from proprietary or copyrighted information. While this sort of information may be available on the WWW, it may be only accessible through a subscription. If a user wanted peer reviewed information on Seneca, he or she could use a service like Proquest. With the proper credentials to access Proquest's database (a library card, for example), the user would find 18 recent articles on Seneca, several from *The Journal of Roman Studies* in London. While this information is obviously of much better quality, no search engine could find it. The user has to have the proper credentials (a library card), know it is there, and know how to search for it.

### Presentation

Finally, presenting the information found on the WWW often requires advanced computing skills and a high reading level. One may find information in .pdf, .doc, .jpg, .gif, .mov, .html forms, among many others. Also, search engines do not format or

present the information – often they do not even inform the user if the Uniform Resource Locator (URL) they have provided is still active. The existing search engines know very little about an individual user and his or her proficiency with computers. Their output is not tailored to various ability levels. However, Google does provide some rudimentary assistance in that it has a tab on its search page that separates out “images” from websites in order to browse through pictures of Seneca.

#### *Web Portals: An Introduction*

To address problems with web searching, librarians and others have come up with the concept of web portals. Briefly, a web portal is a “super discovery tool that specializes in high-quality content.” (Jackson, 2002) Ideally, a web portal would automatically compile diverse electronic resources in various formats from both public domain and copyrighted sources into a single web page. According to Mary Jackson (2002) writing in the Library Journal, librarians “are just beginning to define the requirements for portal products.”

#### *Relevance to Information Systems*

The concept of web portals brings up many issues in information systems. Data mining, database access, querying, data warehousing and other issues come into play. The issue of user interface and programming also must be addressed. Internet use, network bandwidth and client-server architecture must all be tuned to this new way of compiling and accessing large quantities of data. Legal and ethical complications arising from access to copyrights must be resolved. Design issues and formats must be resolved so that the presentation of the information gathered is useful and intelligible.

#### *Literature Review*

My four categories of Infoglut, Veracity, Access and Presentation are meant to address not only the problems, but also the possible features of a portal. The term Infoglut speaks to the culling, sifting or searching through a vast amount of

information. Veracity has to do with the quality, authority and legality of the information that is returned. Access addresses the permissions, credentials and even the cost of using that information. Presentation discusses how such information can be formatted so that it is intelligible and ultimately useful to the intended audience.

### Infoglut

About 2.5 billion terabytes of information exists on the world's computers. (Vogel, 2001) Currently Infoglut is managed by searching. Search engines like Google, Altavista or Yahoo! can reduce the amount of information presented depending on the sophistication and specificity of the search terms. However, studies have shown that the average user will spend only about half an hour looking for information and even



Source: [www.google.com](http://www.google.com)



Source: [www.yahoo.com](http://www.yahoo.com)



Source: [www.altavista.com](http://www.altavista.com)

then chances are only one in five of finding relevant information on the first page of search results. (Feighan, 2001) Fortunately, many powerful searching capabilities are in place today – what is missing is a way to do a single simple search and get a list of integrated, relevant results. There are multiple standards and metadata schemes (Z39.50, XML, MARC, Dublin Core, CIMI and EAD, (Jackson, 2002)) – but none of these is a de facto standard for database search and retrieval. This results in many powerful, but non-integrated databases lying mostly unused. In an information economy this is more than just a waste, it is a competitive disadvantage for the owners of the databases. Realizing this, two-thirds of businesses have plans to build enterprise portals. (Vogel, 2001)

Eventually web portals will become our virtual libraries. However, in order for this to happen, searches must be fast, return full information in multiple formats, ranked according to multiple relevancies. Such integrated information must be portable to multiple software packages in multiple formats. (Jackson, 2002)

In the next generation of search engines (or portals), according to Jackson (2002),

“The core feature ... will be integrated, cross-database searching of a local catalog, other library catalogs, selected web sites, locally licensed full-text and abstracting/indexing databases, and public domain or publicly accessible abstracting and indexing services.”

In addition, some portals will have to access less structured data from e-mail systems, word processing and spreadsheet software and HTML files. (Vogel, 2001) Audio and visual files from multiple sources will also have to be catalogued and presented. This cannot possibly be done manually and will require advances in “machine learning” which today is at best about 85% accurate in the classification of data. (Vogel, 2001)

### Quality/veracity

Currently, the popular search engines on the web sacrifice authority and depth for convenience. For search engines to become true information portals their results must possess higher authority and a more comprehensive scope. (Jackson, 2002)

This will probably be done without government interference. In fact, the FCC has issued a statement that “the Internet is dynamic precisely because it is not dominated by monopolies or governments.” (Charatan, 2002) For example, there is no law in the United States that regulates web content of healthcare sites, so it is up to individual users or independent agencies to look after the quality of information available on the web. Charatan quotes Dr. Arthur Caplan, director of the Center for Bioethics at the University of Pennsylvania:

“In the USA, freedom of speech and the responsibility of users of the Internet to find and evaluate information remain the dominant norms of quality control. In other words, “caveat emptor” has been and remains the American policy ...” (Charatan, 2002)

**WebMD**

Source: [www.webmd.com](http://www.webmd.com)



Source: [www.webmd.com](http://www.webmd.com)

Because of this, M. Clare Feighan (2001) suggests that a panel of clinical experts should review health care information on the web.

There are already several agencies that exist for

this purpose: Health on the Net (HON), American Medical Association (AMA) Guidelines, Internet Healthcare Coalition, eHealth Ethics Initiative, MedCertain and the American Accreditation HealthCare Commission (URAC). (Feighan, 2001 and Charatan, 2002) The Internet Healthcare Coalition has developed an International Code of Ethics for health care sites and services on the Internet (available at <http://www.ihealthcoalition.org/ethics/ehealthcode0524.html>). This code seeks to encourage sites to differentiate between advertising and healthcare content. It has sections on candor, honesty, quality, informed consent, privacy, professionalism, responsible partnering, and accountability. The URAC also runs a website accreditation program. One of the most popular healthcare portals on the web, WebMD, advertises its accreditation with URAC and HON. (WebMD, 2003) Each industry that has a web presence can follow the lead of healthcare to self-regulate.

Although the government does not regulate veracity of content, it does regulate copyrights. Jane Ginsberg, of the Columbia University School of Law, has created a fictional case study of a homemade portal of a university professor. Her fictional professor created a webpage of magazine articles, book excerpts, film clips, recorded music and scanned cartoons. In essence, she has put her college course pack online, putting her in competition with the authors of the works, their publishers and with Kinko's. (Ginsberg, 2001) The professor has not requested permission from the authors and has not restricted access to her materials. Issues having to do with the content and the audience must be addressed for the professor and the university that employs her to remain within the law. Any development of a web portal will have to include processes



Source: [www.bugjam.com/Recent.html](http://www.bugjam.com/Recent.html)

to take into account such legal issues stemming from copyrighted information. More complicated still are issues of filesharing portals like Napster, which federal courts declared illegal in 2002. Litigation is ongoing as Universal Music and EMI are still trying to recoup damages – now from the venture capitalists who invested

Figure 4. Napster Logo

in Napster! (Liedke, 2003) The lesson of Napster is that any portal must abide by copyright laws, or face forced dissolution.

### Database Access

Cooperation is a key attitude when constructing portals. Obviously, one must cooperate with the copyright holders and owners of information. Rachel Heery, Assistant Director of the UK Office for Library and Information Networking (UKOLN), says cooperation must go even further: "the close relationship between the content of the gateway [or portal] and the business model, delivery of service, and marketing direction cannot be ignored." (Heery, 2000) M. P. Evans, writing in the journal *Internet Research*, describes how websites may be able not only to interact with human beings, but also with each other. Content with some computational intelligence will change the WWW from a distribution system to a distributed system. (Evans, 1999)

Heery describes several topics of discussion at the 1999 International Collaboration on Internet Subject Gateways (IMESH) Workshop. IMESH exists for the discussion of subject-based resource discovery services. Heery, and her UK colleagues, use the term "subject gateway" to refer to a specialized portal that is centrally funded, deals with a particular area of subject matter and thus guarantees academic quality and reliability. (HyLife, 2000) In contrast, the more general term portal refers to 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." (Webopedia) Heery likens subject gateways to university departmental libraries that specialize in a limited area of information. Here users benefit from the boundaries placed upon the information space by the specialized subject area. (Heery, 2000) However, in order to avoid being merely an inaccessible island of information, subject gateways need to develop a method of interdisciplinary cooperation with other gateways and portals. This cooperation is best focused on developing standard searching techniques and metadata formats.

Standardization brings many benefits. Users can access a wider range of collections with a consistent, user-friendly interface. Service providers can save money

or concentrate on providing services that are more sophisticated because they do not have to devise the basic cataloging structure and programming from scratch. New “information brokerage” services can evolve to take care of access to, payment for and delivery of intellectual property.

Standardized cataloging of information aids these activities by providing an unambiguous identification of resources. For example, the use of a Digital Object Identifier (DOI) allows a certain piece of intellectual property to be uniquely identified, regardless of its physical location (as referenced by its URL). A DOI is like a bar code for intellectual property. (International DOI Foundation, 2003) Appendix B provides a FAQ on DOIs from the IDF. Other metadata fields allow for accumulation of information about each piece of intellectual property keyed by its DOI. Improved dissemination of information would also result from close collaboration on metadata standards with information providers, i.e. the publishers of web resources. Heery also relates that IMESH contributors discussed keeping an inventory of services, gateway activities and resource description metadata in a data repository. Such an authoritative registry would in essence contain a directory and profile or schema of subject gateways. This would enable users and information brokers “to negotiate and select gateways for satisfying queries.” (Heery, 2000)

Another issue having to do with database access is the coordination of classification schemes, vocabularies and thesauri. (Heery, 2000) The DESIRE Project aims “to build large scale information networks for the research community.” Its first phase, completed in 2000, focused on distributed Web indexing, subject-based Web cataloguing, directory services, and caching. (Desire, 2000)

Searching across multiple databases can be accomplished by software that is compliant with the Library of Congress’s Z39.50 protocol. This is a “client/server-based protocol for searching and retrieving information from remote databases.” (Library of Congress, 2003) Heery states a cross-resource searching capability implies more than adherence to a protocol – it also takes agreements between database owners to use common indexing and resource descriptions. (Heery, 2000)

The IMESH discussions at the end of 1999 proposed several provocative solutions to the problem of database access. Again, what remains is for the international owners of information to cooperate and agree on standards to make their products available.

### Presentation

As of December, 2001, 89 million people were using the Internet. (Feighan, 2001) These users are not homogeneous – any portal will have to pay attention to traditional marketing strategies to be successful.

For example, several studies have shown that much of the information available on the Web is too difficult for the average adult, who reads at an eighth grade level. Benchmarking of popular search engines (see Appendix A) shows that the average reading level of the information returned in response to a sample search was almost at an eleventh grade level.

From the point of view of content, Gerry McGovern, writing in the *Irish Marketing Review*, offers advice on how to develop quality content with primary interest towards the reader. One must view the WWW as a publishing medium. McGovern distinguishes between Altavista, which he says merely dumps information at the users' feet, and Yahoo, which employs human editors to select the best of the Web. "Yahoo has become a huge success because people want limited choice ... If what we're looking for is not in the top twenty, forget it." (McGovern, 2000)

Daintry Duffy, (2000), writing in CIO Magazine cites a study by McKinsey & Co. and Media Metrix which grouped web consumers into six basic behavioral groups (see Table I).

**Table I: A Typology of Web Consumers**

<b>Simplifiers</b>	29% of users Spend seven hours per month on the Web Account for more than half of all online transactions Want to perform specific tasks simply and easily Turn-ons: convenience and reliable customer service Turnoffs: pop-up windows, unsolicited e-mails and chat rooms
<b>Surfers</b>	8% of users Account for 32% of time spent online Access four times as many pages as the average user. Activities: shop, explore and find information and entertainment. Turn-ons: cool design, strong brand and features like games, chat rooms and streaming video Turnoffs: old content and boredom
<b>Bargainers</b>	8% of users Active deal-seekers, enjoy searching for great prices and participating in the community of bargain-hunters Account for 52% of all visits to eBay. Turn-ons: Priceline.com and Beanie Babies Turnoffs: paying full price
<b>Connectors</b>	36% of users New to the Internet – looking for ways interact and explore Only 42% have made an online purchase (versus the average of 61%) Turn-ons: chat rooms, novelties such as e-greetings, well-known brands Turnoffs: complex and intimidating sites
<b>Routiners</b>	15% of users Like to read and research online but buy offline 80% of their time is spent at their top 10 domains Look for superior and timely content Turn-ons: the latest from news and financial sites Turnoffs: old news
<b>Sportsters</b>	4% of active users Act like Routiners but focus largely on sports and entertainment sites Challenge: move them to a revenue-generating model where they pay for content Turn-ons: ESPN.com, colorful sites with interactive features like polls Turnoffs: The Financial Times

Source: *Know Thy Customer*, by Daintry Duffy, available at [http://www.darwinmag.com/read/100100/online\\_sidebar1.html](http://www.darwinmag.com/read/100100/online_sidebar1.html)

The McKinsey marketing study shows that users will not tolerate old content, poor design, too much text or complex or intimidating interfaces. (Duffy, 2000) Multimedia information found in a search must be cleverly, simply and conveniently displayed. It must be written concisely in a language and style that middle-schoolers can understand. It must have easy to use interfaces and its design must be novel, cool, colorful and interactive.

In addition to these user must-haves, portals must provide user security, recognition and customization in the form of pre-programmed alerts, saved searches, ranking choices, etc. (Jackson, 2002) They must be flexible enough to tailor their services to specific targeted consumers. They must provide a host of support services like information capture, integration, manipulation and distribution. They must enable communication with other users for consultation, collaboration and information sharing. (Jackson, 2002) They might learn from a user's past searches and identify other databases with similar information, or even other users who are doing similar searches. Vogel estimates that such capabilities could boost a company's profits by one-third! (Vogel, 2001)

For the construction of a successful portal, all the skills of marketing, advertising and design will have to be employed. Maintenance and editing of content will be ongoing, important tasks. As McGovern says, "The Web is about publishing." (McGovern, 2000) Although automated tools will help manage Infoglut, the human touch is what will distinguish a vibrant online community from an information dump.

### *Benchmarking Analysis and Sample Results*

Already several methods exist to search and weed out information on the web. Although none of these is a pure web portal, several are approaching such a title. Search engines such as Google, Altavista, Yahoo! and others use keywords and phrases to search their databases for "hits." Ratings of a site by percentages or relevance help the user decide which result is most likely to contain useful information. About.com has compiled topical pages on many items of general knowledge. Other websites such as Ask Jeeves allow questions to be asked in natural language and matches to be made with other questions people have asked. Browsing a database of such questions offers another way to narrow one's search. Information may be loaded into these databases either by automated "crawling" of the web or by a human editor manually entering the

uniform resource locators (URLs) of relevant sites, or both. (Webopedia, "How Web Search," 2003)

Still other web resources allow mining databases with search terms. Thane Paulsen of BrightPlanet.com estimates that information in web databases, called the "deep Web," is five hundred times larger than what is targeted by most search engines (McDonough, 2002). Unfortunately, knowledge of and permission to use resources such as Lexis-Nexis and Proquest are not part of a typical user's searching repertoire.

### Benchmarking a Typical User's Experience

To study the current state of search engines and other web tools a simple method of rating search engines has been devised. These benchmarking ratings consist of a series of questions that an average user might turn to the web to answer. According to Robyn Greenspan (2002) of Internet.com:

"the average American Internet user is young, white, employed, well-educated, wealthier, and suburban. Gender is balanced equally among Internet users."

This "web questionnaire" might include questions on some of the most common topics, such as health care, entertainment, news and basic topical research. Questions posed to the search engines should have clearly "right" or "wrong" answers. Appendix A provides a sample question and its results.

*Summary – The Current State of Affairs*

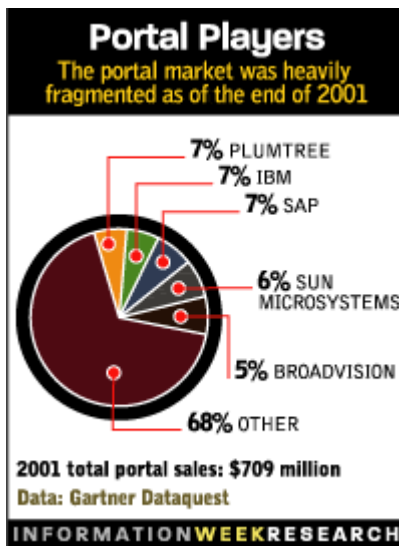
### Search engines and the Benchmarking Report

Because of their convenience, on-line single-step search engines like Google or Altavista have supplanted databases that are more authoritative. (Jackson, 2002) The result has been a glut of questionable information presented in an unorganized, confusing fashion. Gerry McGovern has referred to the results of such searches as "information dumps." (McGovern, 2000) Moreover, these simple search engines,

despite the volume of information they return, do not search information-rich “deep web” of on-line databases, accessible only with special credentials or costly subscriptions. The benchmarking study of some popular search engines in this paper shows that there are several useful features in existence today, but that no one search engine combines all of them into what might be termed a web portal for general information. All of the search engines returned information at a level too advanced for the general reader, and none allowed fine-tuning of search results based on user preferences or abilities.

### Existing Portal Solutions

According to Tony Kontzer (2002), writing in Information Week, the portal market is heavily fragmented. Although total portal sales for 2001 amounted to \$709



million, no single vendor controls more than 7% of the market. In fact, 68% of the portal market is made up of vendors with less than a 5% market share. Kontzer’s statistics suggest that there are at least 20 portal vendors in existence. Jackson (2002) cites an Association of Research Libraries list of 16 vendors of library-oriented portal software. The top three vendors of portal software, with a 7% market share each are Plumtree, SAP and IBM. Sun Microsystems and Broadvision account for 6% and 5% of the market respectively. (See Figure 5. The 2001

Figure 5. The 2001 Portal Market Portal Market)

Jon Surmacz (2003), writing in CIO Magazine, notes that in twelve percent of portal deployments, the system “failed to meet expectations, and in nearly half (46 percent) of those cases, non-technological factors were to blame.” Most of the problems were not with the portal itself, but with company culture and processes. Despite this, Surmacz expects that the portal market will climb to \$957 million by the end of 2003 and to \$1.13 billion by 2004. (Surmacz, 2003)

### No standards

The portal market is clearly in need of standardization and cooperation. The very definition of what a portal should be is still being formulated. Data storage and retrieval standards are in flux. Much of the information available on the web is buried behind non-communicating layers of authentication. Multiple copies of the same information lie on remote servers. For the most part, governments and industry organizations have left the user to be his or her own editor. Legal battles are still being fought over copyright infringements and concepts of ownership of digital intellectual property. The good news is that each of these issues is being addressed and there is promise that they can be resolved so that a true "dream portal" will soon come into existence.

### *Conclusion*

What is most needed is cooperation amongst the various human organizations that possess pieces of the future dream portals. Current search engines possess many powerful features, but none possesses all of them. Amongst the current search engines, Google is able to catch spelling errors in search terms and is able to separate out images from textual information. Proquest is able to return a limited number of peer-reviewed, authoritative answers. AskJeeves is able to accept natural language queries and allows users to browse what others have asked. Yahoo uses both web crawling and human editors to classify information into a familiar directory structure. If these existing features were simply amassed into a single application, the search engine could nearly become a portal.

Various industries have begun to self-regulate the quality of websites in their areas of interest. The health care industry, perhaps because of the life or death importance of its information, seems to have taken the lead in developing codes of ethics and methods of accreditation. Legal issues such as copyrights

and ownership of and payment for digital intellectual property need to be worked out.

Cooperation in the IT field is also necessary. Metadata standards need to be worked out. There needs to be a method of identifying digital information as a first class object rather than by location, i.e. replacing URL's with DOI's. Objects need to have some computational intelligence so that they can relate with each other and with the human users. Presentation needs to be focused on the user and how to accommodate diversity in the web audience.

In short, the technology for extremely useful, powerful and fun web access is there, but the human cooperation to make it come into being for the average Internet user is not yet. Human cooperation and agreements will replace simple search engines with true web portals.

*“The ideal enterprise portal should combine the power of computerized classification with the subtleties of human intelligence ... ” (Vogel, 2001)*

## *Appendix A – Benchmarking Ratings of Search Engines or Portals*

In order to quantify and qualify the efficacy of some of the currently available search engines, I have compiled a table of statistics on the results of a particular search. I have tried to make my search terms specific enough to return a manageable number of results, without being so advanced so as to be out of reach of the average user.

To score the search engine, I progressed from the first (topmost) result displayed to the last in sequential order. I did not follow more than 10 links total. I did not follow links within any returned results, instead I returned to the original search results page of the search engine. For each result, I cut and pasted into Microsoft Word the text of the article, excluding advertisements, tables of contents, etc. I did this until I got a final answer or had followed 10 result links. I then used the Word Spelling and Grammar Check feature to get a word count of the text I had read and two readability scores.

On Table II, I recorded:

- The number of results returned by each search engine.
- The number of hits (results with information actually on the subject), out of the first 10.
- The number of misses (results returned that were not about the subject), out of the first 10, dead links were counted as misses.
- A percentage of hits in the first 10 results.
- Whether the answer was found in less than 10 results (“yes” or “no”).
- The correctness of the answer (“yes” or “no,” any incorrect information results in a score of “no”).
- The total number of words in the text of the returned results, until the answer was found. This is meant to represent a measure of how much reading a person would have to do to find the answer.
- The Flesch Reading Ease Score of the text, up to the answer.
- The Flesch-Kincaid Grade Level Score of the text, up to the answer.
- What credentials or special privileges were needed to access the information.
- The types of formats of information available in the results.

### *Readability scores*

Textual results from the search, up to the point of definitive answer were pasted into Microsoft Word. After a spelling and grammar check, Word displayed two scores: a Flesch Reading Ease score and a Flesch-Kincaid Grade Level score. These scores are both based on the average number of syllables per word and words per sentence. (Microsoft Word, 2002) The Flesch Reading Ease score is a 100-point scale with higher scores signifying easier comprehension. Word suggests writers aim for a score of 65. The Flesch-Kincaid Grade Level score returns a U.S. grade-school level corresponding to the difficulty of the text. Word suggests writing for a mid-seventh grade reader for most standard documents (score of 7.5).

### *Findings*

Earlier in this paper, I mentioned a few features of Google that help a user search more efficiently. Google suggests alternate spellings that help correct a spelling error on Seneca's name. It also provided a separate results page for images. These features are a great help to novice users. The following results of my benchmarking study also point to some strengths and weaknesses of some popular search engines.

As can be seen, the number of sites returned in response to my search for information on how Seneca died ranged from 11 to 1,428. For my purposes, a limited number of results is better and less overwhelming. Most of the search engines, with the exception of About.com, were able to provide a correct answer within the first 10 results anyway. Chart I graphically depicts the number of sites returned for the same search terms. Chart II depicts the percentage of the first 10 results that actually had information having to do with Lucius Annaeus Seneca.

Chart III shows the disparity in the amount of reading necessary to find the answer to my question. Yahoo only required 142 words of reading, while Proquest, being a more authoritative source, perhaps required about 40 times more reading!

Chart IV shows the difficulty of the reading required from the materials each search engine returned. Here again, Yahoo provided the most accessible text (Flesch

Reading Ease Score of 51.2) and Proquest the least (Flesch Reading Ease Score of 34.8). All search engines provided text that was much harder to read than is appropriate for general audiences (minimum Flesch Reading Ease Score of 60). Finally, Chart V shows the reading grade level of the search engine results. Again, the Yahoo results were the easiest to read with a grade level score of 9.4 (mid-Ninth grade level) and the Proquest and About.com results were the most difficult (Twelfth grade level, the maximum score). Again, all search engines returned material that was too difficult for general audiences who read at an average seventh to eighth grade level.

This study suggests that although Google has some convenient features and Proquest and About.com return more authoritative answers, it is Yahoo! that provides the best general reference material. The amount of material it returned and its reading difficulty are the most appropriate for general WWW users. This study has shown that choice of search engine will affect the amount, difficulty and authority of information returned. While Yahoo! may be perfect for general reference, About.com is better for more advanced users, and Proquest is best for professional researchers. More benchmarking would show a clearer picture of these distinctions and the addition of more search engines or portals would offer a more complete picture of the current offering of Internet search options.

**Table II: Search topic: The Death of Lucius Annaeus Seneca**

Search terms: "death of Seneca" "Lucius Annaeus Seneca"

For this search, I was looking for how Seneca died specifically, not just that he died by suicide in 65 C.E. I counted a correct answer as one that specified that he committed suicide by slitting his wrists.

	number of sites returned	number of hits	number of misses	% hits	answer found?	correctness	number of words	flesch reading ease score	flesch-kincaid grade level score	credentials needed	number of formats of information
google	11	10	0	100	yes	yes	1265	41.0	9.4	none	.html, .doc
about	464	10	0	100	no	yes	3465	49.9	10.3	none	.html, .gif, .jpg
yahoo	9	9	0	100	yes	yes	142	51.2	9.6	none	.html, .doc, .jpg
altavista	1428	9	1	90	yes	yes	2737	46.7	10.8	none	.html, .gif, .jpg
askjeeves	no count	9	1	90	yes	yes	1221	34.8	12.0	none	.html, .gif,
proquest*	37	1	9	10	yes	yes	5532	37.5	12.0	cml library card	.html
* Proquest could not find articles unless I omitted the "death of" phrase											

Chart I: *The Number of Sites Returned by the Search Engine. Note: AskJeeves did not provide a count of its results.*

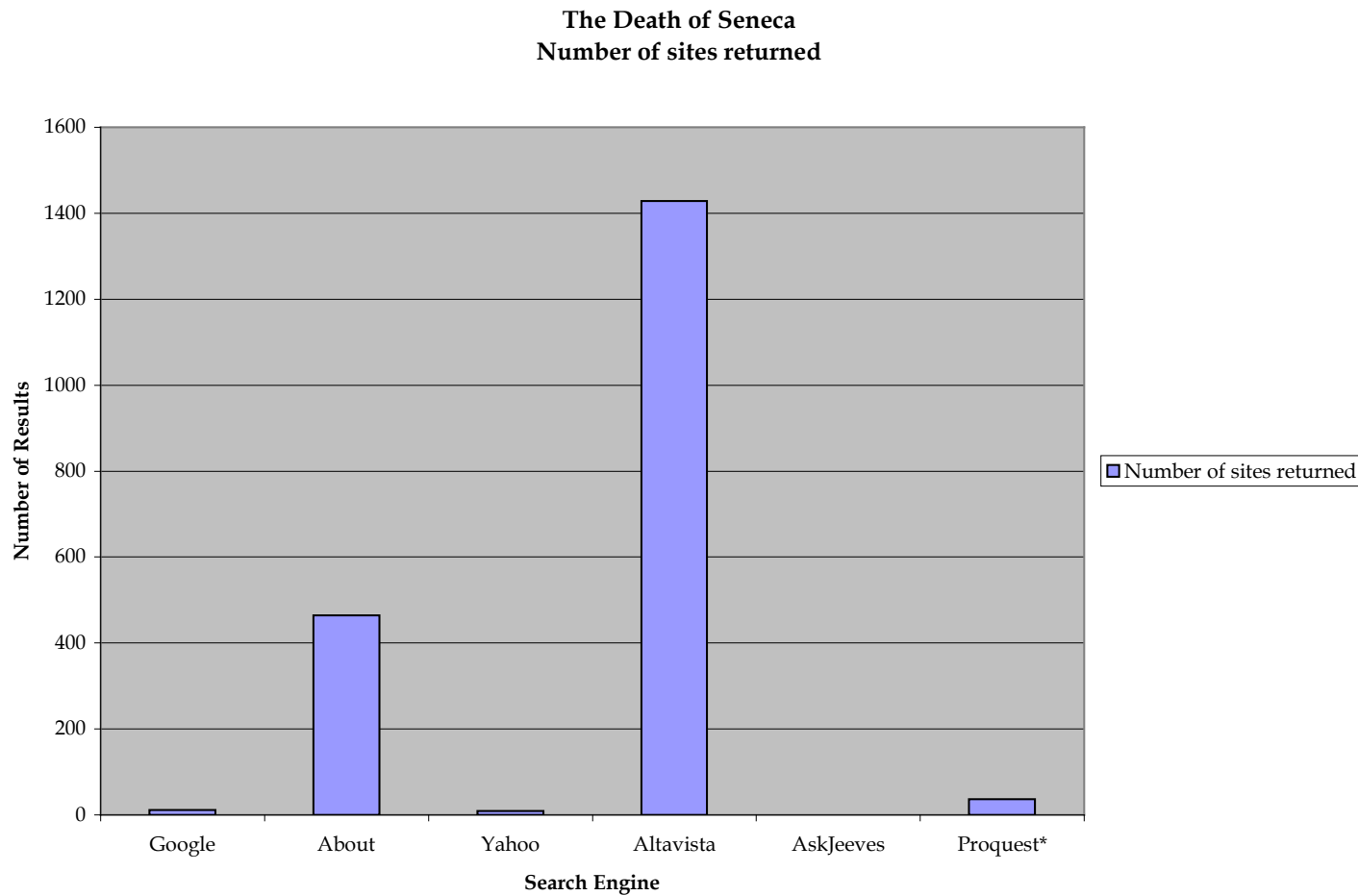


Chart II: *Percentage of Sites with Usable Information having to do with the Search Terms as specified*

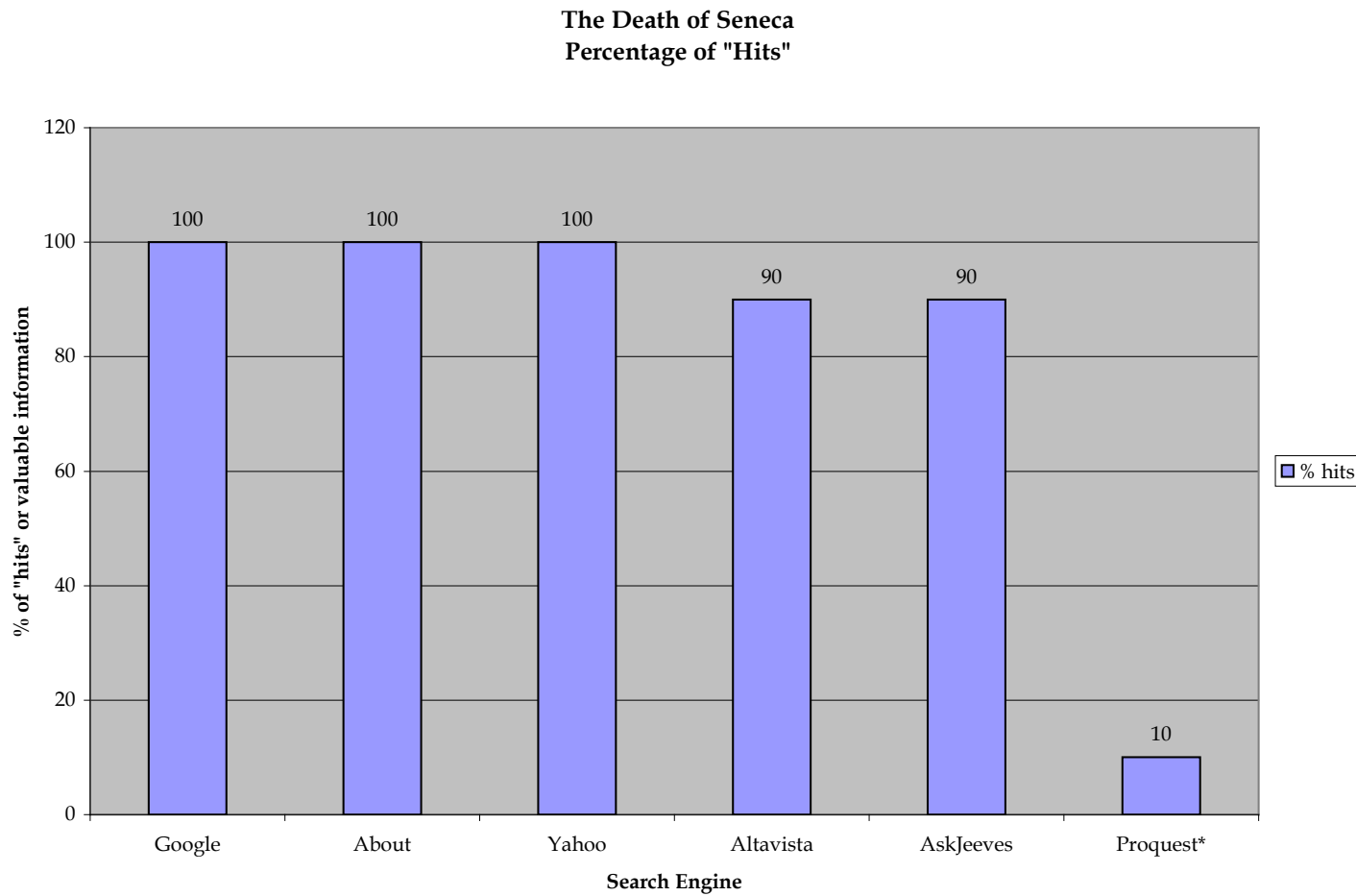


Chart III: *The Amount of Words to read until the Answer is found.*

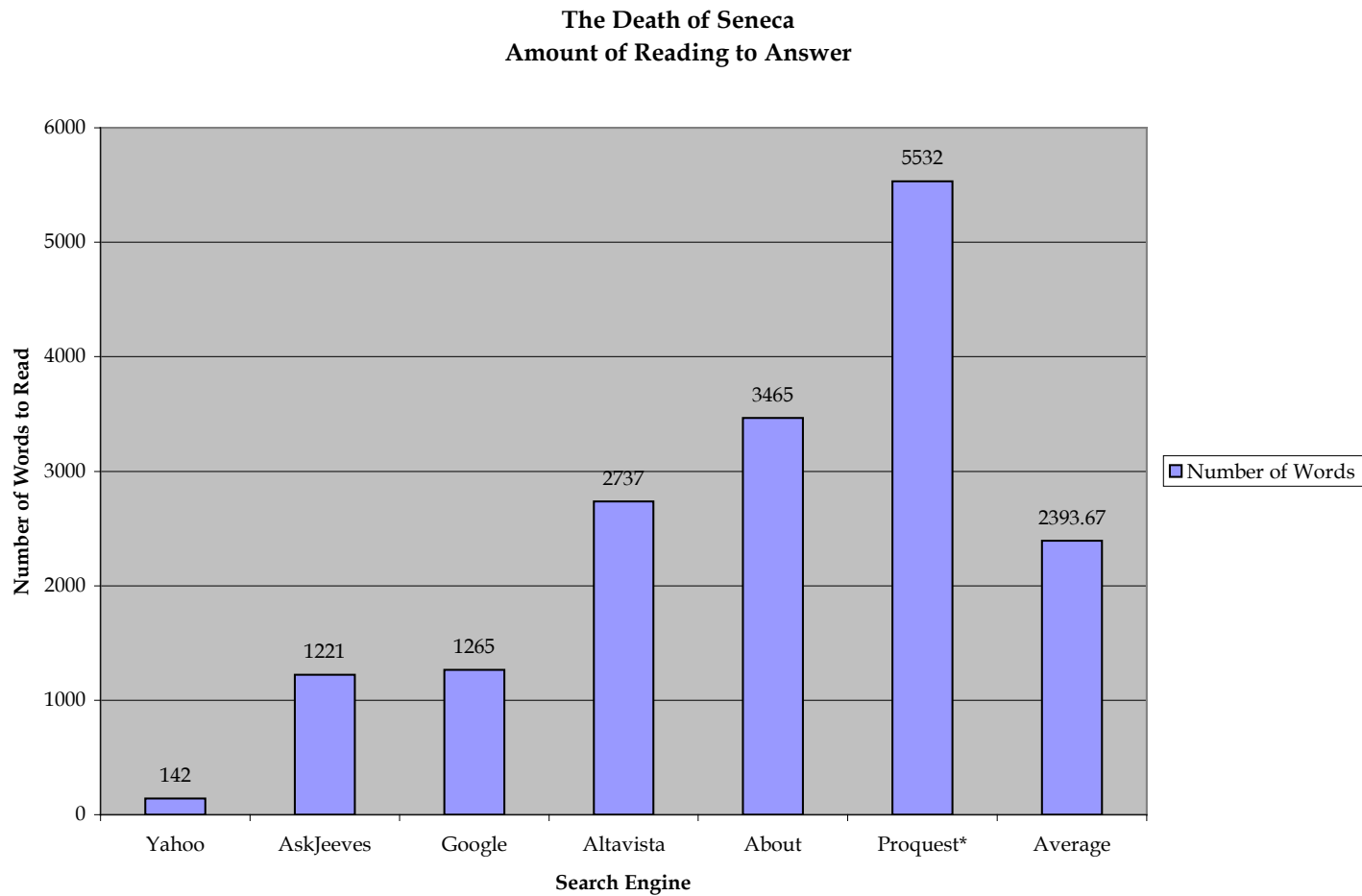


Chart IV: Flesch Reading Ease Score (Higher numbers are easier to read. Material written for general audiences should rank at least 60.)

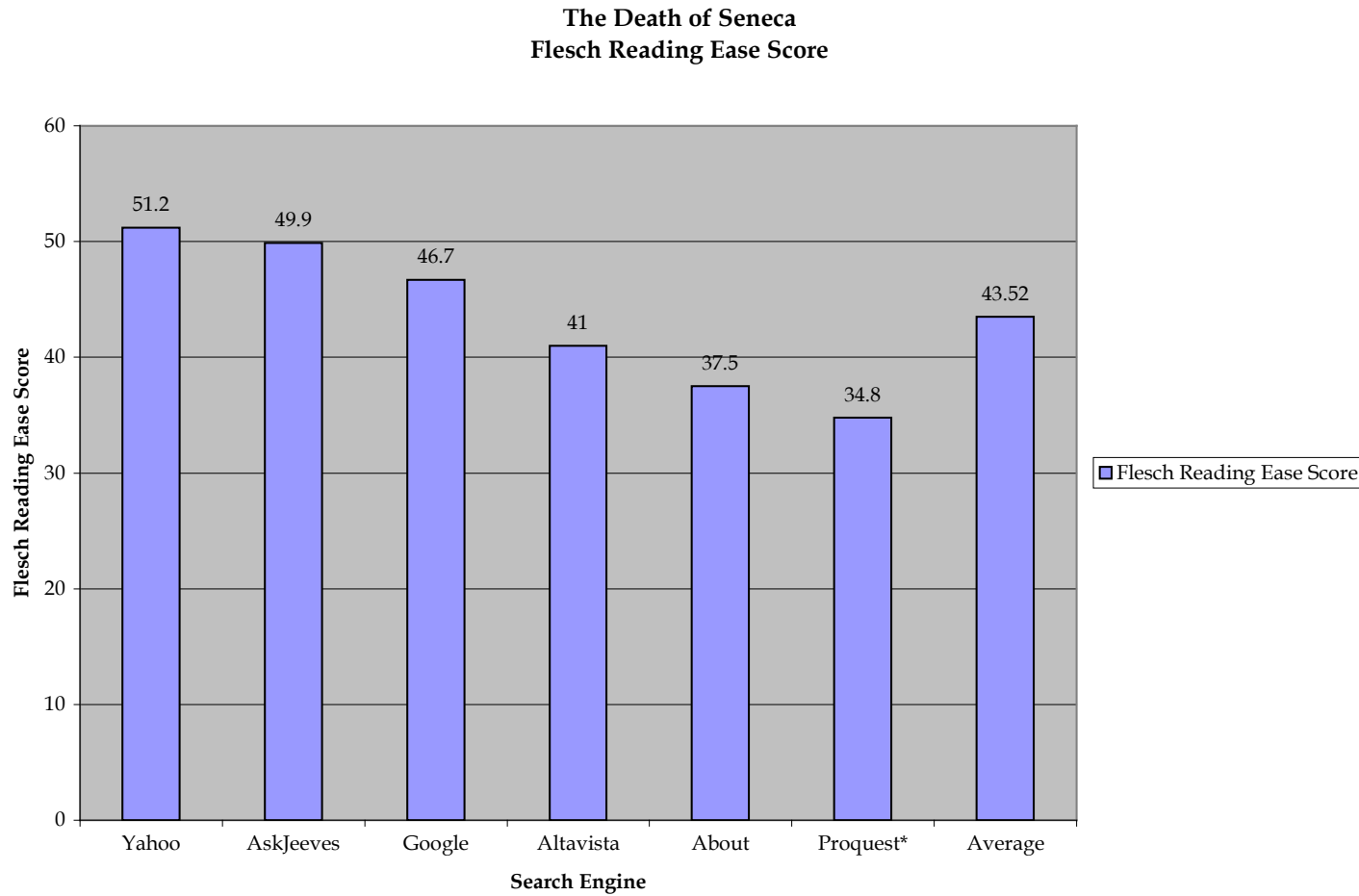
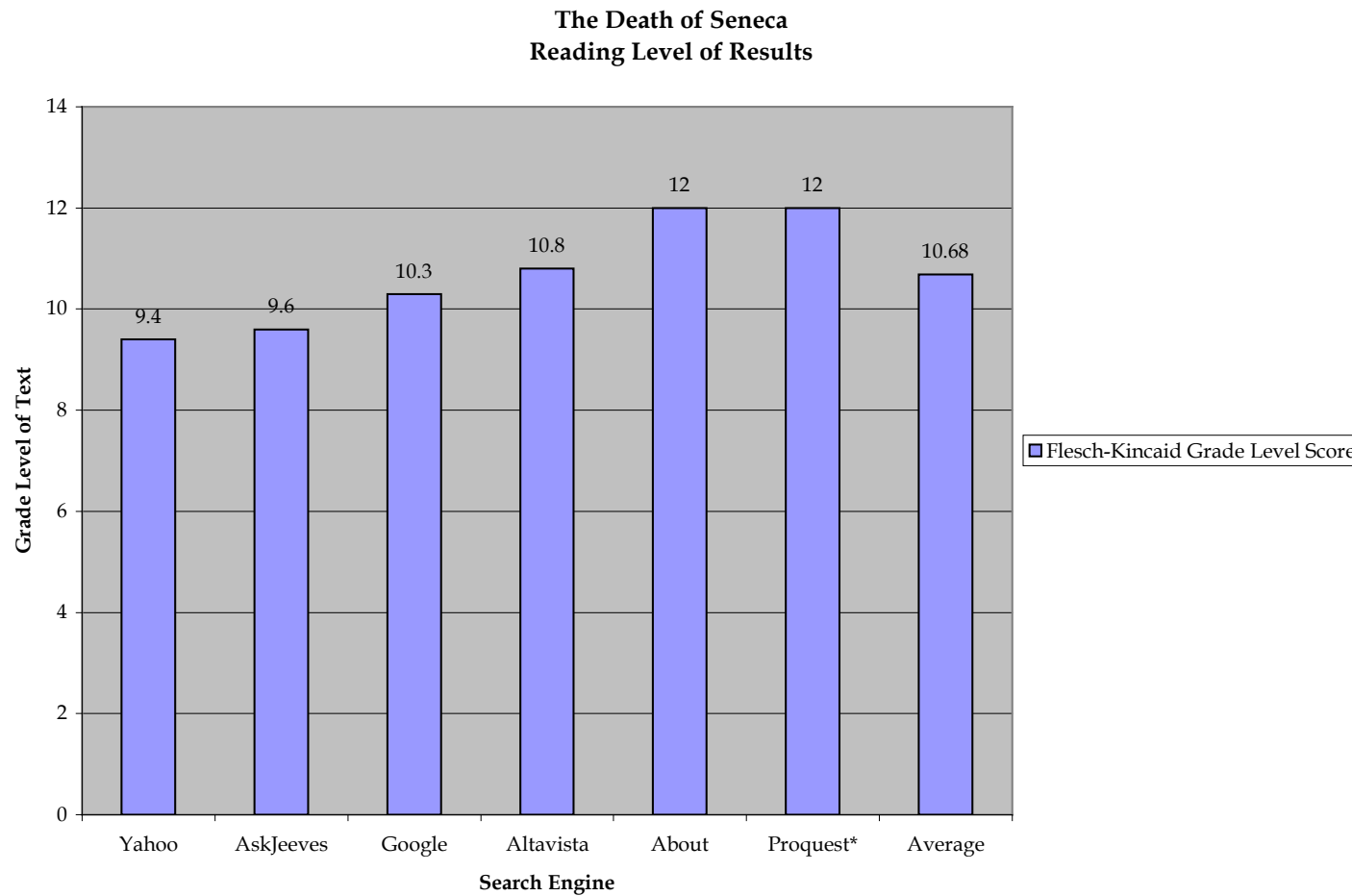


Chart V: Flesch-Kincaid Grade Level Scores. (Material written for general audiences should be at 7<sup>th</sup> grade level.)



## *Appendix B – Digital Object Identifier*

### **What is a DOI?**

Digital Object Identifier -- a *digital identifier* for any *object of intellectual property*.

A DOI provides a means of persistently identifying a piece of intellectual property on a digital network and associating it with related current data in a structured extensible way.

A DOI can apply to any form of intellectual property expressed in any digital environment. DOIs have been called "the bar code for intellectual property" – like the physical bar code, they are enabling tools for use all through the supply chain to add value and save cost.

A DOI differs from commonly used internet pointers to material such as the URL because it identifies an object as a first-class entity, not simply the place where the object is located. The DOI identifies an entity directly, not some attribute of an object (an address is an attribute of a thing, whereas the thing itself is a first class object).

A DOI also differs from commonly used identifiers of intellectual property such as standard bibliographic and related identifiers (ISBN, ISRC, etc.) because it can be associated with defined services and is immediately actionable on a network.

A DOI is an implementation of the Internet concepts of Uniform Resource Name and Universal Resource Identifier. A DOI differs from abstract naming specifications such as URI in that it is a defined implementation complete with social and technical infrastructure, ready to use.

Source: The International DOI Foundation, <http://www.doi.org/faq.html#1>

## *Appendix C – Sponsor’s Comments*

**Gayle DeGennaro**

**e-mail: [DeGennaG@franklin.edu](mailto:DeGennaG@franklin.edu)**

Comment	Changes Made
Multiple grammatical corrections	Corrections made
Dislikes use of “There is ..., There are ..., It is ..., etc.”	Corrections made
Define all acronyms	Corrections made
Need to find a place to explain some items of relevance for information systems (around page 4): data mining, database access, querying, data warehousing	A detailed analysis of these concepts is out of the scope of this paper
For the Conclusion: <ul style="list-style-type: none"><li>• Bring together distinct topics</li><li>• Search Engines - too much data</li><li>• Portals</li><li>• What goal and your solution is to this.</li><li>• Restate the problem</li></ul>	See conclusion  Problem is lack of cooperation and coordination  Solution is primarily one of human cooperation and implementation of existing technologies
Chart I <ul style="list-style-type: none"><li>• Might recommend log scale</li><li>• Take off background shading</li></ul>	Tried Log Scale, but I thought it minimized the large differences between Google and Altavista – which was contrary to the point I was trying to make, i.e. that Altavista returned far too much information.

*Appendix C – Sponsor’s Comments, continued*

**Ron Morgan**

*e-mail:* [morganr@franklin.edu](mailto:morganr@franklin.edu)

Comment	Changes Made
"Beef up" Summary and Conclusion	These sections were not written yet when Ron read my paper (end of MIS 665).
Check for APA consistency	checked
Dislikes title "Web Portals: How to Edit the Web"	A difference of opinion. I liked the title and could not come up with one that was any better yet still as concise. One alternative was Web Portals: How to Edit the Web for Everyone? Possibly some mention of how the Web could offer a better searching experience tailored to the kind of user? Decided to stay with the original title.

*Appendix C – Sponsor's Comments (continued)*

*Mike Powers*

*e-mail:* [powersm@franklin.edu](mailto:powersm@franklin.edu)

Comment	Changes Made
First person usage -- you switch from using first person (referring to "I") early in the paper (first 3-4 pages) to third person, which is more formal and appropriate for research: I'd stay consistent with the latter through the whole paper.	Removed use of first person
Data mining. You may want to touch on the concept of data mining as a tool in your database descriptors, it may be tangentially related to your topic. Some excellent cases are available, particularly Amazon.	Decided that an in depth exploration of data mining was beyond the scope of this paper.

## *References*

- Barker, J. (2001). *Things To Know Before You Begin Searching*. UC Berkeley - Teaching Library Internet Workshops. Retrieved February 6, 2003 from <http://www.lib.berkeley.edu/TeachingLib/Guides/Internet/ThingsToKnow.html>
- Desire Consortium. (2000). About Us. Retrieved May 8, 2003 from <http://www.desire.org/html/aboutus/aboutus.html>
- Duffy, D. (2000). Know Thy Customer. Darwin (companion site of CIO.com) Retrieved May 7, 2003 from [http://www.darwinmag.com/read/100100/online\\_sidebar1.html](http://www.darwinmag.com/read/100100/online_sidebar1.html)
- Evans, M. et al. (1999). Strategies for content migration on the World Wide Web. *Internet Research*, 9(1), 25-34. Retrieved November 29, 2002 from ProQuest.
- Feighan, M. (2001). A healthy regard. *Pharmaceutical Executive, Digital Pharma Supplement*, 28-29. Retrieved November 29, 2002, from ProQuest.
- Greenspan, R. (2003). Internet Not For Everyone. *Internet.com*. Retrieved June 19, 2003, from [http://cyberatlas.internet.com/big\\_picture/demographics/article/0,,5901\\_2192\\_251,00.html](http://cyberatlas.internet.com/big_picture/demographics/article/0,,5901_2192_251,00.html)
- Hawkins, D. (1999). What is credible information? *Online*, 23(5), 86-89. Retrieved November 29, 2002, from ProQuest.
- Heery, R. (2000). Information gateways: collaboration on content. *Online Information Review*, 24(1), 40-45. Retrieved November 29, 2002, from ProQuest.
- HyLife, The Hybrid Library of the Future. (2002). Subject Gateways and Portals. Retrieved May 8, 2003 from <http://hylife.unn.ac.uk/toolkit/gateport.htm>
- The International DOI Foundation. (2003). FAQ. Retrieved May 8, 2003 from <http://www.doi.org/faq.html#1>
- Internet Healthcare Coalition (2000). eHealth Code of Ethics. Retrieved May 7, 2003 from <http://www.ihealthcoalition.org/ethics/ehealthcode0524.html>
- Jackson, M. (2002). The advent of portals. *Library Journal*, 127(15), 36-39. Retrieved November 23, 2002, from ProQuest.

- Kontzer, T. (2002). Plumtree Bets On Multiserver Tools. Retrieved June 26, 2003. from <http://www.informationweek.com/story/IWK20021018S0022>
- Library of Congress. (2003) Z39.50 Maintenance Agency Page. Retrieved May 8, 2003 from <http://www.loc.gov/z3950/agency/>
- Liedke, M. (2003). Napster Suit Unnerves Venture Capitalists. Associated Press. Retrieved May 7, 2003 from [http://news.findlaw.com/ap\\_stories/high\\_tech/1700/4-29-2003/20030429123006\\_23.html](http://news.findlaw.com/ap_stories/high_tech/1700/4-29-2003/20030429123006_23.html)
- Locutio.com (2002). Sénèque. Retrieved May 3, 2003 from <http://www.locutio.com/images/histoire/galeriesenequegrand.gif>
- McDonough, B. (2002). *Bringing a Much Bigger Internet to Light*. Retrieved February 6, 2003 from <http://www.crmdaily.com/perl/story/18625.html>
- McGovern, G. (2000). Managing information in the digital age: How the reader is king. *Irish Marketing Review*, 132(2), 55-60. Retrieved November 23, 2002, from ProQuest.
- Microsoft Word. (2002). Help: Grammar and Spell Check.
- Seneca, L. (n.d.). *On Tranquillity of Mind*. 9.4ff (trans. J.W. Basore). Retrieved February 6, 2003 from <http://ccat.sas.upenn.edu/jod/texts/seneca.english.html>
- Surmacz, J. (2003). Prime time for Portals. CIO.com. Retrieved June 26, 2003 from <http://www3.gartner.com/Terminate?src=timeout>
- Vogel, C. (2001). Untangle the Web. *Communications News*, 38(9), 82-3. Retrieved November 23, 2002, from ProQuest.
- WebMD. (2003). About WebMD. Retrieved May 7, 2003 from <http://my.webmd.com/content/article/60/67019.htm>
- Webopedia.com. (2003). *The Difference Between the Internet and the World Wide Web*. Retrieved February 6, 2003 from [http://www.webopedia.com/DidYouKnow/Internet/2002/Web\\_vs\\_Internet.asp](http://www.webopedia.com/DidYouKnow/Internet/2002/Web_vs_Internet.asp)
- Webopedia.com. (2003). *How Web Search Engines Work*. Retrieved February 6, 2003 from <http://www.webopedia.com/DidYouKnow/Internet/2003/HowWebSearchEnginesWork.asp>

Webopedia.com. (2003). *Web Portal*. Retrieved May 8, 2003 from  
[http://webopedia.internet.com/TERM/W/Web\\_portal.html](http://webopedia.internet.com/TERM/W/Web_portal.html)