THE SEARCH ENGINE MAP LANDSCAPE

Michael P. Peterson
University of Nebraska at Omaha (USA), mpeterson@unomaha.edu

Abstract: *A variety of experiments have been conducted beginning in 2001 to examine the state of maps and the Internet, particularly in reference to their availability through search engines. Search engines are viewed here as a window to the Internet, a way of understanding what types of maps are easily available to Internet users. In order to improve the Internet as a medium for cartography, we need to understand the current state of map availability. A series of high school and college freshman were asked to find map resources through the Internet. The time was measured for students to find different maps. It took only 16 seconds for some students to find a map of Africa in the JPEG format. Most of the maps that are indexed by search engines are simply static maps, often scanned from paper, not the highly-interactive hypermedia maps that were expected with this new medium. There are very few examples of highly interactive maps for general use that are currently available through the Internet, and those that do exist are very difficult to find. The continued development of Internet cartography is based on both the accessibility of of highly interactive maps that make exploit the potential of this new medium and a system of remuneration so that the authors of these time-consuming products can be compensated.*

1 INTRODUCTION

The Internet has become a major form of information delivery. According to the Computer Industry Almanac (2005), there are currently 935 million Internet users or nearly 16% of the world's population (see Table 1). This is up from 533 million Internet users worldwide at year-end 2001 which, at that time, represented only 8.7% of the world's population. There were only 200 million Internet users at year-end 1998 and 61 million in 1996. It is expected that the number of Internet users will reach 1 billion by mid-2005 and 1.46 billion by 2007.

With the growth of the Internet comes the expansion of map use through this medium. Internet map use was tracked at four major sites between 1997 and 2001. The growth in Internet map use was compared to the growth in the use of the Internet itself. It was found that while both growth rates are strongly exponential, the growth in the use of maps through the Internet exceeded the growth rate for the Internet itself (Peterson 2003). This change in map distribution happened over only a decade. Never in the history of cartography has there been such a dramatic shift in how maps are delivered to the map user.

The Internet has clearly changed the process of mapping and map use. The new medium has drastically increased the availability of maps for those who have access to the Internet and led to more interactive forms of mapping. The distribution of maps through the Internet is still relatively new and much work lies ahead in order to make the Internet an effective means of transmitting spatial information, particularly developing forms of mapping that use the potential of this new medium.

Year	Internet Users in Billions
2007	1.46
2005	1.00
2004	0.935
2001	0.533
2000	0.327
1996	0.061

Source: Computer Industry Almanac (2001, 2005)

Table 1: Number of Internet Users

Based on the concept of hypermedia (Nielsen 1990), the hypermedia map is a map-based interactive multimedia presentation that combines some mix of text, pictures, video, graphic images, sound and other forms of media. For example, a hypermedia map presentation on Africa might include links to regions, countries, music, climate, or population—all

linked together with maps that emphasize both location and thematic information. The hypermedia map has evolved since the mid-1980s as methods of map delivery have evolved from the CD-ROM to the Internet. With the growing interest in interactive forms of communication, the integration of media with maps has become a major area of research and development. The goal is to transcend the static and sequential nature of information presentation and ultimately create a greater and broader understanding of the world in which we live (Peterson, 1995, p. 127-128). The overall objective is nothing less than a revolution in how spatial information is communicated.

While we can envision and even implement highly interactive hypermedia maps that help guide users to a greater and broader understanding of the world, bringing such maps to a large audience through the Internet is an entirely different matter. If we are to be successful, the challenge is to not only conceive of useful hypermedia maps but also finding ways of marketing these maps, funding their creation, and making them available to people at a reasonable cost. The purpose here is to examine the current availability of Internet-delivered hypermedia maps through search engines, and ways of making these highly interactive maps more available. We first examine the search engine information landscape.

2 THE INFORMATION LANDSCAPE

Landscapes come in many shapes and sizes. There are tropical valleys and windswept desert plains. There are landscapes formed by methods of farming and urban landscapes develop as a result of human patterns of habitation. O'Day and Jeffries (1993) use the term "information landscape" to refer to the landscape of information that we traverse when we venture into a hypermedia environment. They study the specific process used by individuals in finding information through hypertext. Researchers also examine what patterns of behaviours are exhibited when users browse information landscapes, how performance varies across the two distinctive tasks and how different navigational aids influence patterns of use Toms (1996). It has even been suggested that maps can serve as a guide to the information landscape (Block 2002).

The information landscape that most people experience is the one presented by search engines. It is the only effective tool at present to traverse the information landscape. These automated systems explore this ever growing landscape and collect keywords that are entered into a database. Keyword searches are subsequently made by the search engine user. Of course, there is more to this landscape than can be found by the search engine. Some estimate that search engines have only indexed a third of all web pages. It is useful to examine how these systems work.

Search engines are based on a program called a web robot or a spider that traverses web pages in an automated fashion. There are two steps in the traversing process: 1) words are entered into a special index; and 2) other links are found for later traversing. If a page has links to 100 pages, and each of these has links to another 100 pages, a web robot, in theory, will create an index for all of these of pages – and this can soon number into the millions. Web robots only need to be told to search the main page and they find all linked pages automatically. The index created by the web robot cannot be too large, no more than about 2 KB. If the index were larger, it could essentially be making a copy of the entire page.

After making a collection of these small index files of billions of pages, the web robot begins to prioritize the links by keyword. This is where the problems start because the process is inevitably arbitrary and based on inadequate information. Each search engine also prioritizes in a different manner as well. Companies do not provide the exact algorithms because this would lead to abuse by those people who trying to get their web pages to show up near the top. There is somewhat of a Catch-22 situation in which we need to know how the web robots work so that they provide better matches, but as soon as we know how the robots work, they get abused by those that want to manipulate them and they no longer provide good matches.

The ranking system used by Google, called PageRank, relies on a system that counts the number of times pages are linked from each other. It also incorporates a weighting factor that analyzes the "importance" of the page that is making the link. Important, high-quality sites receive a higher ranking, which Google remembers each time it conducts a search. Google states that it goes beyond the number of times a term appears on a page and examines all aspects of the page's content (and the content of the pages linking to it) to determine if it's a good match for a query.

Some basic search engine rules seem to be shared by all search engines. Smaller pages are given preference over longer pages with a similar set of keywords. The text in the title of the page is given a higher priority than text in the page. The META tag, listed near the top of the page, is another way of making the web robot give a page a high priority. Before the META tag was available, a high rank for a page was only possible if a phrase was used repeatedly throughout the text. Now, the phrase just has to appear in the META tag. Popular pages that have more links directed to them are listed higher. This tends to cause new pages to fall into an "anonymity trap" since they are not linked from other pages, and they can't be linked because they can't be found.

Search engine rankings are extremely competitive and the operators of websites are under pressure to make sure that their sites are listed near the top. Webmasters will spend a considerable amount of time making sure that their sites are highly placed on the major search engines. Most search engine companies now accept cash in return for a high placement. In general, the search engine remains a necessary if frustrating part of people's lives.

Sherman and Price (2001) identify five major problems with search engines:

1) Cost of crawling: Crawling the web is very expensive and time-consuming, requiring a major investment in computer and human capital.
2) Crawlers are dumb: Crawlers are simple programs that have little ability to determine the quality or appropriateness of a web page, or whether it is a page that changes frequently and should be re-crawled on a timely basis.
3) Poor user skills: Most searchers rarely take advantage of the advanced limiting and control functions that all search engines offer.
4) Quick and dirty results: "Internet Time" requires a fast if not always thorough, response. A slower, more deliberate, search engine would not gain user acceptance although it might lead to better results.
5) Bias towards text: Search engines are highly optimized to index text. For non-text pages, such as images, audio, or streaming media files, the search engine can do little more than record filename and location details.

2.1 Media Search Engines

Media search engines are designed to index images, sound, or video files, sometimes collectively referred to as "multimedia." Image searching is the most common and is offered by major search engines like Google and Yahoo. WAV, MIDI and MPEG sound files can be found through specialized search engines like MusicRobot.com. Music sharing programs like Kazaa and Limewire locate and download music and other types of media through a point-to-point protocol. Yahoo includes a search for video files on its search page. Media search engines represent an alternative to the text-based search engine and represent a more promising technology for finding hypermedia maps.

Image search engines generally index images in the GIF, JPEG, or PNG format and represent a major way that Internet users find maps. Google reports an index to more than 880 million images. Of these, approximately 3,750,000 are images of maps. These dominate in any type of place name search. For example, an image search for "Africa" results in 12 maps in the top 20 links, and 23 in the top 40 links. Of these 23 maps, 13 are in the JPEG format. The Yahoo image search finds 10 maps in the top 20 images and 14 maps in the top 40 image links. Google finds a total of 145,000 hits for "Africa map" while Yahoo has 30,400 hits. Maps, in the form of images, represent a major component of Internet image content.

The primary problem with image search engines is that the indexes are created from words associated with the images, and not the images themselves. This is because the image itself cannot be searched. Rather, the indexing of the image is done by examining the title of the image. It might also take into account any accompanying 'ALT' picture tags coded into the HTML page or look for clues from the image's context – for example, the words or phrases that are close to the image, or the 'META' tags found at the top of the HTML coding. The nature of the Web site and its provider may also be taken into account. The image search indexing process can easily be fooled with a non-descriptive filename or associated text that does not relate to the image. For example, a search engine would likely be fooled by a page on South America with a mis-labelled link to a map of Africa.

An alternative to the automated approach are collection-based search engines that index a single or small number of image collections. Commercial stock photo collections, like Corbis, that offer images for sale, will implement their own search engine-like procedure to find images from within their collections. These search engines are augmented with "human indexers" that assign keywords for each picture.

All of these types of image search engines are text-based in that their indexes are created from words associated with the images. There have been attempts to create content-based search engines that 'index' visual characteristics of an image, such as its shape and colour. However, these attempts are still largely experimental and are often limited to single image collections (TASI 2004).

Yahoo is the only major search engine that provides a video search option. The search engine returns links to movies in five different formats (AVI, MPEG, Quicktime, Windows Media, and RealPlayer) but few movies of these movies incorporate maps. For example, "Africa map" returns only one movie that includes a map in the 96 movies that it returns for this search. The 7,800 video files that are listed for "Africa" contain very few map examples. Even a search for "animated map" that should be represented in a video database leads to only 36 hits. Clearly, the current video search option is not an effective way to search for hypermedia maps.

2.2 Map Search Engine

A search engine that is designed specifically for maps would serve many purposes. Most of the maps that are currently available through the Internet are classified as pictures. As a result, a place name image search results in a confusing mix of maps and pictures. The image search engine would benefit from a clear distinction between pictures and maps, something that could be determined in an automated way by looking at the level of pixel color variability in a file. Presumably, the map would have large areas that have the same color or shading and this would not be true of ordinary pictures that are characterized by gradations of pixel values.

Determining map content in other types of multimedia files would be more difficult. These files have a more complicated structure and performing similar automated inspections of files may not be sufficient to separate map and non-map content. An author could self-define the map content of a hypermedia product and insert this into an associated header or metadata file but enforcing such a self-classification of content would invariably lead to problems. A workable, automated map search would provide a better alternative and could categorize content by both file type and media content.

3 EXPERIMENTS IN FINDING ONLINE MAPS

A variety of experiments have been conducted beginning in 2001 to examine the state of maps and the Internet, particularly in reference to their availability through search engines. In an experiment conducted in 2001, a series of high school and college freshman were asked to find map resources through the Internet (Peterson 2001a). Over 100 students were tested in five groups in a computer user room with 25 Pentium III 500Mhz computers. Every student had their own computer with Microsoft Internet Explorer and a relatively fast Internet connection. They were instructed to find or make three maps: 1) a map of Africa by country; 2) a map of Africa by country in PDF format; and 3) a map that places a star on a street map showing where they lived.

Students could find these maps in remarkable speed. For example, it took only 16 seconds for some students to find a map of Africa. Almost all of the students found such a map within 45 seconds but some students required slightly more than a minute. The search strategies differed slightly between the high school and college students. The college students had learned to simply use the search field in the toolbar area of Explorer that linked to Microsoft's MSN search engine. The high school students first went to Google to start the search and this slowed their search times.

It is interesting to note which map of Africa was most commonly found by students. It was not an interactive map. Rather, it was a scanned paper map in JPEG format. The map exhibits the typical JPEG compression artefacts that make this file format less than ideal for use with maps. But, that didn't seem to matter to the students. It was the map that students were expecting to find of Africa and, although the map is barely legible and could not be zoomed or queried, they were quite happy with what they had found.

To find the PDF map of Africa, the Google search engine helped the high school students find this type of map faster than their college counterparts. Their fastest response time was 28 seconds while the fastest college student required 40 seconds. Loading the Acrobat plug-in took a good portion of this time. Some students required over a minute and were then provided help. Few could explain the advantage of the PDF file type and only a couple students knew that zooming was possible. They did know that PDF files took longer to load because the plug-in needed to start first, and they found this to be annoying and mentioned that they would often cancel the process when this happened with other documents. While they seemed impressed that one could zoom into the PDF map, in the end, many liked the JPEG map better because it loaded faster, although it was essentially illegible.

Making a map of home proved more difficult because the search engine was less useful for this task. Students tried to search for "map of home" but this did not yield links to any of the interactive street mapping programs. Some remembered MapQuest.com from a previous visit and went directly to that site. For the few students using the Yahoo search engine, they were provided a link directly to MapQuest.com. The main Yahoo page has a link to Yahoo's interactive mapping site but students did not look for this indicating that these students preferred using a search engine to a directory list. At that time, about 1/3 of the students had never used an online mapping site to find a particular location. In the end, the fastest map of home was produced in 30 seconds but most students could not complete the task in one minute and some had to ask for help. Once the map was made, students complained that the star indicating their home was not in the right location. This is a common complaint and results from the address matching process in which locations of addresses are estimated along street segments. This process often produces an incorrect location.

The most interesting observation from these experiments was that static maps are much easier to find than interactive maps. Search engines are oriented toward static pages because this is all that they can index, and these pages usually have links to static files. Another observation was that students didn't understand how the interactive map was made. One student said it was "magic." Others said that the interactive street map was made through satellite imagery or used GPS. Although these may be involved, the process of bringing such a map to map user is much more complicated.

3.1 Finding Maps – A 2005 Update

The same experiment was conducted in early 2005, nearly four years after the first study in 2001. This time students in a third year college course were asked to find the same three maps but with faster computers (Pentium 4, 2.4GHz with 512 MB of memory). The results remained remarkably similar. Almost everyone found a similar map of Africa in less than 30 seconds. Some required less than 10 seconds. Finding the PDF map took slightly longer but most were still able to locate the map in under 40 seconds.

The most interesting finding was that all of these students had been previously exposed to interactive street mapping sites and everyone completed this map in less than 45 seconds. About half completed this task in less than 30 seconds.

3.2 Finding Hypermedia Maps

To determine the level of difficulty in finding online hypermedia maps, an experiment was first conducted by the author with the Google search engine. The purpose of the experiment was to determine what type of hypermedia sites could be found within the top 20 links provided by Google's first two pages. Six different text strings, such as "Maps Hypermedia Africa," were entered into the search engine. Each of the top 20 links were examined and potential links were followed through two additional links to see if a desired page could be found. Further links may have been followed if a promising lead was found in the first two links.

A major advantage of the Internet is the ability to include pictures, sound and video with maps. Try finding maps that include one of these elements using the following text in the search engine. If you can't find a map with pictures, sound or video in the top 20 links provided by the search engine, enter "Not found." Do not proceed down more than 2 links in any page provided by the search engine.

Search Engine Text	Time in Seconds
1. Maps Hypermedia Africa	_____
2. Maps Multimedia Africa	_____
3. Maps Pictures Africa	_____
4. Maps Sound Africa	_____
5. Maps Movies Africa	_____
6. Maps Flash Africa	_____

Figure 1 A portion of a questionnaire given to students in an upper-division geography course. Many students could not find examples of the desired maps on the Internet. Those that could find examples sometimes required over 8 minutes of search time. The search for Flash files produced the most results but required the knowledge that hypermedia maps would be presented using the Flash file format.

The results of the search are presented in Table 2. As can be seen, very few hypermedia or multimedia maps could be found. No true hypermedia maps were found using a typical search string that most people would use. Many links were to commercial sites (.com), particularly to books on hypermedia. Other links were to libraries that would list library audio-visual collections but there were no links to them. The only hypermedia maps that were found during the entire search resulted from a search on "Africa Flash." In this case, Flash is referring to an online multimedia authoring program from Macromedia. Most people would not know to perform a search using the name of the proprietary file format.

Google Search Text*	Sites with Maps in top 20	Sites with Pictures of Africa in top 20	Sites other Media of Africa in Top 20	Hypermedia Maps
Maps Hypermedia Africa	I	I		
Maps Multimedia Africa	~~IIII~~ IIII	III	II	
Maps Pictures Africa	II	II	I	
Maps Sound Africa	II	IIII	I**	
Maps Movies Africa	~~IIII~~		II	
Flash Africa	~~IIII~~	III	III	II

*Resource found within 2 clicks. Additional clicks may have been needed to activate resource. **Only site with sound presented bird songs.

Table 2: Summary of Google search for hypermedia maps of Africa. Although many maps and photographs of Africa were found, few sites integrated maps with other media. Those that did could only be found using the "Flash" file type keyword.

A similar experiment was conducted with the same upper-division college students. In this experiment, the students were instructed to time themselves as they attempted to find a variety of maps of Africa that included pictures, sound, or video. An electronic stop watch implemented through a web page was used as the timer. Students were given the instructions as shown in Figure 1.

The results largely paralleled what was found by the author. In many cases, students could not find any maps of Africa that included the desired media. Many students spent over five minutes trying to find a suitable map before giving up. Unless they were hightly motivated, this is far longer than most people would normally devote to such a task.

Pictures were the most common media element found with maps. Maps with pictures were found with many of the search strings that were used. Even searching for "Maps Sound Africa" produced hits to maps of Africa that included pictures. "Flash Africa" led to the most sites with hypermedia maps. Searching for this type of content using Flash as a keyword is problematic because it requires that one be aware of a program that is used to create hypermedia content. Most people would not have this information.

CONCLUSION

Interactive and multimedia maps offer the potential of a more engaging form of mapping, and the Internet is a functional method of delivery. While examples of highly interactive maps are available through the Internet, they are difficult to find with current search engines – even with the newer media search engines that have recently appeared. Online multimedia maps are not well represented on the Internet. As a result, people don't see the advantages of combining maps with other types of media.

Search engines are based on creating indexes of words. The specific words that are used in labelling material is extremely important in the indexing process. Not only does the word need to be unique so that it is distinctive from other resources but it needs to have a meaning that is generally understood by most people. While multimedia and hypermedia are recognizable concepts, people do not use them in reference to interactive maps. While such interactive maps would have a broad appeal, at present, there are few online examples available through the Internet, and these can only be found with great difficulty. In order for this type of map to be more mainstream, academic terms like multimedia and hypermedia will either need to become more broadly understood, or new terms will need to be introduced that more effectively describe this type of map to the public. But, the real challenge for online mutlimedia maps will be to create a demand and market that would fairly compensate cartographers for their efforts.

REFERENCES

1. Block, Marylaine (2002) Mapping the Information Landscape. Searcher, Vol. 10, No. 4. [http://www.infotoday.com/searcher/apr02/block.htm]
2. Computer Industry Almanac (2005) Worldwide Internet Users will Top 1 Billion in 2005, http://www.c-i-a.com/200010iuc.htm
3. Computer Industry Almanac (2001) 15 Leading Countries in Internet Users Per Capita, http://www.c-i-a.com/ pr0904.htm
4. Libicki, Martin, Schneider, James, Frelinger, David R., and Slomovic, Anna (2000) Scaffolding the New Web: Standards and Standards Policy for the Digital Economy. Santa Monica, CA: Rand Corporation.
5. Nielsen, J. (1990) Hypertext and Hypermedia. Boston: Academic Press.
6. Peterson, M. P. ed. (2003) Maps and the Internet. Amsterdam: Elsevier Press.
7. Peterson, M. P. (2001a) Finding Maps through the Internet: An investigation of High School and College Freshman. Presented at the Annual Meeting of the Association of American Geographers, Los Angeles, CA, April 2001.
8. Peterson, M. P. (1995) Interactive and Animated Cartography. Englewood Cliffs, NJ: Prentice-Hall.
9. O'Day, Vicki L. and Jeffries, Robin (1993) Orienteering in an information landscape: how information seekers get from here to there, Proceedings of the SIGCHI conference on Human factors in computing systems, p.438-445, April 24-29, 1993, Amsterdam, The Netherlands.
10. Sherman, C. and Price, Gary (2001) The Invisible Web: Uncovering Information Sources Search Engines Can't See. Medford, NJ: Cyberage Books.
11. TASI - Technical Advisory Service for Images (2004) A Review of Image Search Engines. October 2004 [http://www.tasi.ac.uk/resources/searchengines.html]
12. Toms, E. (1996) Exploring the Information Landscape. Proceedings of CHI '96, April 1996, 63-64