



Western Washington University
Western CEDAR

WWU Honors Program Senior Projects

WWU Graduate and Undergraduate Scholarship

Spring 2006

The Impact of Digitization on Research and Libraries' Roles

Kristin N. Kohles

Western Washington University

Follow this and additional works at: https://cedar.wwu.edu/wwu_honors



Part of the [Liberal Studies Commons](#)

Recommended Citation

Kohles, Kristin N., "The Impact of Digitization on Research and Libraries' Roles" (2006). *WWU Honors Program Senior Projects*. 240.

https://cedar.wwu.edu/wwu_honors/240

This Project is brought to you for free and open access by the WWU Graduate and Undergraduate Scholarship at Western CEDAR. It has been accepted for inclusion in WWU Honors Program Senior Projects by an authorized administrator of Western CEDAR. For more information, please contact westerncedar@wwu.edu.

Running head: IMPACT OF DIGITIZATION

The Impact of Digitization on Research and Libraries' Roles

Kristin N. Kohles


Western Washington University

In partial fulfillment of the requirement for the Honors Program

June 10, 2006

HONORS THESIS

In presenting this Honors paper in partial requirements for a bachelor's degree at Western Washington University, I agree that the Library shall make its copies freely available for inspection. I further agree that extensive copying of this thesis is allowable only for scholarly purposes. It is understood that any publication of this thesis for commercial purposes or for financial gain shall not be allowed without my written permission.

Signature Date 6/10/06

What is Digitization?

Digitization is a broad term used when describing turning written text, visual media, and audio into a format readable by a computer. It can mean anything from creating the original digitally or transferring a non-digital format into a digital one. This makes it a difficult topic because everyone has their own idea of what it means, and most of their definitions fall into the broad one. This paper will focus on written text being digitized in a variety of ways.

Reasons for Digitizing

According to Karen Coyle (2006), former librarian for the University of California, reasons for digitizing can be summed up into six categories. These categories are: digitization for preservation, discovery, delivery, reading, research, machine manipulation, and born digital.

When a book is digitized for preservation, making the created file available for the future and of the utmost quality are the two biggest goals. Images are taken of the book and stored in very large files that hold very high detail. These files contain so much detail that when viewed on computer monitors today, some details are lost. The file type chosen to store the file is of a known format so that computer programs can always be written that will be able to view the file. This makes sure that in the future the file type does not become obsolete and the data stored on it not viewable.

Digitizing for discovery is when a text is scanned using optical character recognition (OCR) so that the text can be searched. OCR scans the text and stores it into a machine-readable format. However, this does not store any formatting except for the occasional heading. OCR does not store any images including graphs or pictures. Examples of digitized for discovery texts can be found in various databases such as ProQuest. It should also be noted that when scanned, errors do happen. Similar letters can be exchanged, such as an O for a D for example. Errors such as these are easily noticed by people, but machines cannot understand the context and fix the errors themselves.

With the advent of the internet, letters can be sent in seconds through email rather than days through the post office. This has given birth to the digitized for delivery document. These documents are scanned so they can be sent via email to the intended recipient and usually end up being printed. The document is readable, but not of the best quality so the file can stay small, be sent quickly, and printed easily.

Digitization for reading is the area that needs the most work. E-books are a form of digitization for reading. This is the category for material that is meant to be read onscreen. The major qualification for this group of digitized materials is that it needs to be able to support sustainable reading. In order for material to be able to read for long periods of time on a screen it needs to be portable, able to be bookmarked, and easily navigated. As it is, these criteria have not been met as will be explained later.

Electronic reference tools such as dictionaries and encyclopedias are examples of digitizing for research. These types of materials have been scanned and can usually be searched by their entries. They are meant to only be read a few pages at a time so there

are not the problems associated with text that needs to be able to support sustained reading.

Information that has been digitized for machine manipulation is not usually meant to be read. This information is usually data that has been entered into a special format, such as tab delimited, so that it can then be used in various database or spreadsheet computer programs. Once it has been exported into these programs the information is then used further to create more information such as charts and graphs that show past and future trends.

The last of Coyle's categories is born digital. This describes all items that were created via a computer. Items that belong in this group include e-mails or Microsoft Word documents. These items were created via a program and the file type can usually only be read by a handful of programs that may become obsolete.

Benefits of Digitizing

Digitization has greatly affected the information world. It has allowed libraries to save on shelf space, allowed materials to be available to a wider audience, and protected the original documents.

When books go digital the paper copies can be moved off-site. This means that libraries do not have to be made larger to support growing collections and the money can be spent elsewhere. Information is easier to retrieve when one does not have to go into the stacks in the library to find a book. Digital materials do not have to be re-shelved as physical books do (Heterick 2003). Paper books fade or lose color (Gelernter, 2004).

They also “disintegrate. They go out of print. They’re expensive to produce, bulky to store, and back breaking to move” (Johnson, 2004, p. 44). Robert Lopresti (personal communication, May 11, 2006), librarian at Western Washington University, also pointed out that when a copy is made, not only does the original stay safe from the elements, but it stays safe from people who may wish to steal it.

Digitization has also allowed a greater availability of material. Users from anywhere can access information as long as they have access to an internet connection. Materials can be accessed at home when libraries are closed. Older, rare books that would otherwise be only available for use in the library due to their delicate nature can also be accessed from outside of the library (Hirtle, 2002). Not only does digitization provide convenience for some but is a necessity for others. People in remote areas or foreign countries would not be able to access these materials at all if they weren’t available through technical means (Kelly, 2006).

Drawbacks of Digitizing

Like most things, along with its benefits digitization also has its drawbacks. These include the commercialization of information, preservation of digital materials, costs, and as of yet poor usability for navigating digitized works.

With the advent of Google Book Search, a digitization project by the internet megalith Google Inc., major concerns have risen over the commercialization of information. Since Google, Inc. is a for-profit company traded on the stock market, profits are very important to the company. This means that they will do what they need

to in order to turn a profit, even if it means shutting the book search down and making all their materials inaccessible (Chillingsworth, 2005b). It also means they may have a bias towards making money rather than spreading information. This is not a concern only for Google's project either. Many digital book collections are controlled by the vendors that sell them (Falk, 2003). These vendors may change the formatting to their own preferences or even discontinue a text without notice. Works provided in a collection are up to the publisher and there is limited flexibility in choosing which works the subscriber wishes access to (Hunter, 2000). There is also a constant change in the journals available through already subscribed collections as those collections lose access themselves for various reasons.

If commercial enterprises are in charge of providing electronic materials, will they do the best job of preserving the materials? This is what has many concerned including R.M Sharma (2003), director of the libraries at West Virginia University, who is worried about the lack of long-term commitment provided by these companies. "We don't want to have to go through this again in ten years' time," states Robert Kiley, head of systems strategy at the medical research funding charity, Wellcome Trust (Chillingsworth, 2005b, p. 12). If the companies providing the collections do last, what if the library subscribing can no longer afford access? Then they risk losing any back issues to which they once had access (Falk, 2003).

There is also worry about the lack of preservation for the formats themselves. As stated earlier, only materials digitized for preservation attempt to address this issue. But what about the others? If, as some hope, digitized materials will replace paper copies, how will those digitized versions stand the test of time and not become obsolete?

DSpace, a digital archive, stores everything in the universal binary language in hopes that the information can always be transferred into new file formats (Falk, 2003). However, this has not been tested yet since DSpace is only a few years old. The only preservation techniques that have been time-tested are those for paper materials. Even digitization supporters such as Kevin Kelly, author of the article “Scan this Book” (2006) realize the stability and durability of paper books.

It is getting harder and harder for institutions to justify paying for both paper and electronic versions of materials (King, 2004). This means that the preservation issue needs to be solved soon. As it is, many places are very close to destroying their paper copies with the assumption that electronic copies will last. A historic example of the damage destroying materials can wreak without the proper preservation of the newer formats can be seen with the invention of microfilm. When newspapers were converted to microfilm, some pages were skipped. This meant that when the newspapers were destroyed, the skipped pages were lost (Maxymuk, 2005).

If an item can be preserved digitally, it will have to be constantly updated in order to work with whatever programs are being used at the time (Maxymuk, 2005). This can mean a large amount of time, energy, and money whenever a large overhaul needs to be done. Since technical advances are happening so quickly, this means a lot of money in a short amount of time and libraries, and other institutions, just do not have the budget to handle these costs (Sharma, 2003) or the costs of specialized training to staff (Middleton, 2005). This is not the only cost of digitizing; there are many others.

The time and cost of actually doing the initial digitizing itself is very costly. This is one of the reasons that big companies, such as Google, Inc., are doing the digitizing.

The University of Michigan is partnering with Google to have its works digitized. If Google had not stepped in, it would have cost the university \$600 million over a span of 1,600 years to complete their holdings (Chillingsworth, 2005a).

Costs are also associated with any subscription service to a database for full-text scholarly material. In order to keep all issues of a journal, a subscription fee needs to be paid on a regular basis or all access to back issues is lost (Sharma, 2003). These costs, along with the others, make it hard for poorer libraries to gain access to newer material. This is causing an inequity between libraries. To combat this, some libraries are merging their resources both internally and externally. Different libraries are combining various resources so that their patrons maintain access to a wide variety of materials. Internally, departments are merging as technical departments merge with various other parts of the library (Sharma, 2003).

Besides saving money, technical departments are merging more with other library functions to work on the problem of usability. Because of the prominence of electronic sources, all departments have to be able to understand how they work. User interfaces are constantly changing and there is no standardization between them (M. Mautino, personal communication, May 12, 2006). Electronic books do not usually have easy ways to categorize them and put them into a library's catalog and thus have to be inputted by hand which takes time and energy (Coleman, 2004).

As mentioned earlier, there has been no method found yet to allow a user to read an electronic text for extended periods of time. Reading from a screen for a long length of time is hard on the eyes. It is not easy to have a portable reader, either, because of size and quality issues. If it is too small it is hard to read but if the screen has been enlarged

for easy reading, then the reader is too large to easily carry and things are easily reflected off of the screen (Hillesund, 2001). It is also hard to browse easily because only a few pages can be viewed at a time (EPIC, 2003).

Impact on Research: Benefits

Besides allowing information to be more available to a wider audience, as discussed earlier, digitization has had many other positive effects on research. When someone wants to know about current events, they want to know about what just happened, not what happened yesterday. With the computerized transfer of information, this is possible. Digitized newspapers can be read quickly after they are released, even if they were released across the country. In the past, it would have taken at least a day for the newspaper to be mailed across country. When many materials can be viewed at once, it also allows for easier comparison (Kelly, 2006). Johns Hopkins has made different translations of old works available side by side so that the translations can be compared. In Kentucky, old manuscripts of *Beowulf* have been scanned and computers used to enhance previously unreadable text (Hirtle, 2002). These new research methods are only some of the newer features allowed by digitizing.

Digitized texts can be linked. This means that a footnote, bibliography entry, word, or phrase can be clicked by a user and have either the source cited or word definition appear in seconds. This opens a whole new avenue of research as readers can look at the sources immediately and question the author's interpretation of the material

(Chartier). With all of this interlinking of materials, value will be added to research as “no book will be an island” (Kelly, 2006, p. 45).

Being able to search a wide variety of texts at a single time allows more books to be discovered that may have not been discovered in the past. This includes out of print titles as well as ones that publishers felt would not bring in a profit. When this happens, these otherwise unknown authors can begin to gain a reputation and avoid obscurity. This may also give the author a reason to write more which, in the end, will create even more information available to the public.

While people discover more and more material, they may start making their own collections of what they feel is of most interest and share this with others. These can be equated to playlists in a music collection (Kelly, 2006). Not only can users personalize content, but they can make it easier for them to understand by either changing the way the text is displayed or adding a speech program so they can listen instead of read (Hillesund, 2001).

Impact on Research: Drawbacks

Digitization has also affected research in some negative ways. When a researcher can go directly to where their search term is located in a document, they lose the concept of context. This not only includes the material, but the way in which it is presented. Digitization has also made some users think that they do not have to put as much effort out as before in locating information because they can settle for what they have received from their full-text searching. However, if they do find something relevant in their search

results, the ability to tell if it is a reliable source is much more limited than with printed works.

Where a piece of information is placed in a work has a lot to do with how a reader perceives that information. For example, if an article in a newspaper is placed above the fold on the front page where it can be seen when placed in the newspaper stand, it is given much more importance by the reader than something tucked away in the bottom corner of the last section of newspaper. Even if it is a subconscious act, the reader picks up on that location and rates the information's importance on it. What the information is placed by also makes a difference. If an article is placed by others of a similar topic, it is assumed that the article being read matches the topic. If a reader is looking for information on the subject, they will go to the section that holds it. However, if the reader has no interest, he or she will not pay that article much attention. All of these contextual relationships are lost if material has been digitized and organized in a different way such as alphabetical or by date. It is not only the location, but the font, punctuation, layout, and the way the text is divided that impacts the reader's response to the material (Chartier, 2004). Even the material used for a book can tell a lot about the text. For example, Robert Darnton, professor of history at Princeton, points out that the book's binding and paper quality can tell you a lot about the intended audience and when it was published (Lee, 2004).

Many college students do not realize the importance of context and they have found digitized texts a very easy way to research. The problem here is that they have not used the electronic formats to supplement their traditional research methods, but to replace them. The Electronic Publishing Initiative at Columbia (EPIC) conducted a

survey of college students' research habits (2003) and found that a third of students do not go any farther than electronic materials. Some even settle for what they find on-line, knowing that there is better information available at the library. There is worry in the information field that students will depend more and more on these on-line sources, even if the information is not the most accurate.

Searching is a skill that must be learned. Many do not realize this and when they do not get relevant results, the searcher may decide that the information that he or she is looking for is not available. Having digitized text available for searching enforces this belief (Fister, 2005). Going directly to the searched for keyword may also have the searcher skip over the relevant material which would have been easier to spot if the searcher had to look through the material themselves. This hands-on searching also allows for side trips when a different passage than the one containing the search term is noticed. These discoveries can add a whole new angle to the research (Baron, 2005). However, these discoveries are missing when only a small portion of the text is available to view in the digitized versions.

When people put all their faith in electronic sources, they can also lose sight of where the information is coming from. Source authenticity and reliability is harder to determine in electronic texts. Specialized training is needed in order to know how to determine where the text is coming from. According to the EPIC study (2003) three-fourths of students tried to evaluate their electronic sources but were unable. It is also easier to change a digitized version than a print version with no trace left of the changes (Chartier, 2004).

Impact on Libraries' Roles

With digitization creating such a major change in the information world, libraries have had to adapt. Their roles as providers of neutral, authentic sources and givers of guidance are more important than ever. Libraries as a place may change as they are no longer needed as much for book storage, but they will never completely disappear. Digital materials will also draw users to what original copies the library does keep.

Libraries already do their best to provide neutral and reliable sources. This will not change as the digitization of materials becomes more prevalent. It will in fact become more important. As stated earlier, digitized sources are harder to authenticate than print. This means users who need reliable sources will come to depend more and more on libraries for those sources. Libraries also have access to databases that cannot be accessed by the ordinary user at home (J. Sodt, personal communication, May 17, 2006).

E. Yegin Chen, director, senior analyst, and author of the Eduventures.com report sums up digital libraries as

guarantors of quality...you know the people who have designed it have searched all this digital information and found this material—all of which is copyrighted and copyright cleared (and) has been proven to be useful—and have put it in an organized format...[You] can trust the material in an e-library because it's gone through a rigorous editing process. (Goodspeed, 2001, p. 35)

Librarians will also be needed more as guides to accessing information.(Hafner, 2004) as well as understanding the technology and user interfaces (M. Mautino, personal

communication, May 12, 2006). With more material comes a larger audience seeking that material. This means there is a wider gap between skill levels of users and this leads to a greater need for librarians (Jassin, 2005). With some users who do not know how to create good search queries, librarians are needed to help sort through the large amount of results (Lee, 2004). Librarians do quite a bit of “hand-holding” (Jassin, 2005, p. 24) and this will only increase as newer forms of materials come out and library patrons need assistance understand them.

One of the reasons librarians are needed is because people need human contact. Because of this, libraries will never truly disappear as people seek them out as places to interact and share ideas with others (Jassin, 2005). People who come to libraries are from all walks of life and the library is a safe place for the diverse views to converge in a positive way (Falk, 2003). Above all, “the ability to browse and to experience the combined effect of the many information resources available at the library has been an important aid to inquiry” (Falk, 2003). This includes print, electronic, and human resources.

At the moment only one of every twenty books has been digitized and that is an optimistic estimate (Kelly, 2006). Because of this, many books that are referenced in electronic resources are not available in digital format and must be viewed at a library (Chillingsworth, 2005b). In this regard digitization will actually guide more people to libraries. People may also wish to view original works that are only housed at a library.

Digitization has brought a lot of attention to copyright issues as various digitization projects make materials available to a wide audience and the line at where the authorized user is defined becomes blurry (Hunter, 2000). This is a very controversial topic and there is not enough room to address it in this paper. However, it is important to note that it does exist and it will have a large impact on libraries and the way in which research is done. The conclusion will impact what will be available digitally and may even impact what will be available physically. As it stands right now, a group of publishers have a law suit out against Google, Inc's book digitization project. It is unclear how this will end up as Google does have some backing and so is about even with support as the publishers. This issue has also put some libraries in a bad light as some groups against digitization have even begun calling libraries "havens of content theft" (Dames, 2006, ¶ 4).

Digitization Projects

The following are a few examples of digitization projects and what they are attempting to do.

Google Book Search (formerly Google Print)

This project is broken down into two parts: Google Books Partner Program and Google Books Library Project. Google, Inc. (n.d.) explains that Google Books Partner Program is "an online book marketing program designed to help publishers and authors promote their books by showing you a limited number of sample pages" (n.p.). The Google Books Library Project, on the other hand, is a "project to include collections of a

few libraries in Google Book Search and, like a card catalog, show you information about the book plus a few snippets – a few sentences of your search term in context – or to show the full book when out of copyright” (n.p.).

Project Gutenberg

At the Project Gutenberg website (n.d.), it states “Project Gutenberg is the first and largest single collection of free electronic books, or eBooks. Michael Hart, founder of Project Gutenberg, invented eBooks in 1971 and continues to inspire the creation of eBooks and related technologies today” (n.p.).

The Million Book Project

Carnegie Mellon Libraries (n.d.) is the group responsible for the Million Books Project (MBP) which “aims to digitize at least one million books and offer them free-to-read on the Internet. In addition to providing the world's largest collection of e-books, the project will provide a rich testbed for many different areas of research and innovation” (n.p.). They hope to complete their goal by 2007.

References

Baron, N.S. (2005, December 4). Why read when you can Google? *The Gazette*, p. A15.

Retrieved February 6, 2006, from ProQuest database.

Carnegie Mellon University Libraries. (n.d.) *Libraries Projects*. Retrieved June 9, 2006,

from <http://www.library.cmu.edu/Libraries/LibProj.html>

Chartier, R. (2004). Languages, books, and reading from the printed word to the digital

text. *Critical Inquiry* 31(1), 133-152. Retrieved April 8, 2006, from Thomson

Gale Expanded Academic ASAP database.

Chillingsworth, M. (2005, January). Top libraries embrace 'industrial scale' Google Print

revolution. *Information World Review*, 209, 7.

Chillingsworth, M. (2005, April). Librarians demand digitisation answers. *Learned*

Information, 212, 12.

Coleman, Gordon. (2004). E-books and academics: an ongoing experiment. *Feliciter*,

50(4). 124-125.

Coyle, K. (2006). One word: digital. *The Journal of Academic Librarianship*, 32(2), 205-

207. Retrieved June 3, 2006, from the Science Direct database.

Dames, K. M. (2006). Library organizations should support Google book search. *Online*

30(2), 18-19. Retrieved March 2, 2006, from ProQuest database.

Electronic Publishing Initiative at Columbia: Online Use and Cost Evaluation Program.

(2003). *The EPIC online survey of college students: Executive Summary*.

Retrieved February 14, 2006, from Columbia University, Electronic Publishing

Initiative at Columbia web site:

<http://www.epic.columbia.edu/eval/find09/find09.html>

Falk, H. (2003). Developing digital libraries. *The Electronic Library*, 21(3), 258-261.

Retrieved March 17, 2006, from ProQuest Research Library database.

Fister, B. (2005). Google's digitization project – what difference will it make? *Library Issues* 25(4), 1-4.

Gelernter, J. (2004). The digital library edge. *Information Edge*, 8(8), 16-18. Retrieved April 8, 2006, from the Thomson Gale Military and Intelligence database.

Goodspeed, L. (2001). A reading room of one's own. *Boston Business Journal*, 21(5), 35. Retrieved April 7, 2006 from the Thomson Gale Business and Company ASAP database.

Google, Inc. (n.d.). *About Google Book Search*. Retrieved June 9, 2006 from <http://books.google.com/intl/en/googlebooks/about.html>

Hafner, K. (2004, June 21). Old search engine, the library, tries to fit into a Google world. *The New York Times*, p. A1. Retrieved November 28, 2005, from ProQuest database.

Heterick, B. (2003). The impact of being digitized on becoming digital. *Against the Grain*, 15(4), 44-48.

Hillesund, T. (2001, October). Will e-books change the world? *First Monday*, 6(10). Retrieved February 22, 2006, from http://www.firstmonday.org/Issues/Issue6_10/hillesund/index.html

Hirtle, P. B. (2002). The impact of digitization on special collections in libraries. *Libraries & Culture*, 37(1), 42-52.

- Hunter, K. A. (2000). Concerns carried into the third millennium. *Against the Grain*, 12(1), 22-24.
- Jassin, M. (2005). The flat track to new career options for information professionals. *Online* 29(5), 22-24. Retrieved March 16, 2006 from ABI/INFORM Global database.
- Johnson, D. (2004). Turning the page. *School Library Journal*, 50(11), 44-46. Retrieved February 22, 2006, from ProQuest Research Library database.
- Kelly, K. (2006, May 14). Scan this book. *The New York Times Magazine*, 42-49, 64, 71.
- King, R. J. (2004). The future of the special library. *Information Outlook*, 8(9), 10-16. Retrieved April 7, 2006, from Thomson Gale Military and Intelligence database.
- Lee, F.R. (2004, December 18). Questions and praise for Google web library [Electronic version]. *The New York Times*.
- Maxymuk, J. (2005). 13 ways of looking at Googlebrary. *The Bottom Line: Managing Library Finances*, 18(4), 197-201. Retrieved January 4, 2006, from Emerald Fulltext database.
- Middleton, K. (2005). Collaborative digitization programs: a multifaceted approach to sustainability. *Library Hi Tech* 23(2), 143-150. Retrieved March 17, 2006 from Emerald Fulltext database.
- Project Gutenberg. (n.d.). *About Project Gutenberg*. Retrieved June 9, 2006, from <http://www.gutenberg.org/about/>
- Sharma, R. N. (2003). The impact of digital collections on the humanities: a librarian's view. *Against the Grain*, 15(4), 48-52.