Andrew J. M. Smith
Emporia State University

02:004.738.5
https://doi.org/10.18485/filkult.2016.1.ch1

THE ROLE OF INFORMAL DIGITAL LIBRARIES IN SUPPORTING E-PHILOLOGY

Сажетак

Although the emphasis in the library world is on large scale digital libraries run by publishers or for-profit corporations, the growth of digital libraries from small beginnings to significant resources supporting e-philology is really the story of grass roots organizations, of groups of interested amateurs, and of independent libraries and academic institutions, who experimented with formats and content and paved the way for the larger commercial operations. Academic libraries in particular can be so focused on promoting use of their databases and e-resources purchased at great expense, that they fail to promote other, freely available resources that provide access to an extensive range of digital materials. This paper examines the development of the digital library and highlights the wide variety of resources that are available free of charge to support learning and research.Tools and strategies for the discovery and promotion of these types of digital resources are also discussed, as well as barriers to access and potential operational problems.

Digital Libraries – opportunity

The digital library provides an excellent resource for the study of language and literature and more and more material is now available in digital format to encourage the study of e-philology. Every day, increasing amounts of data are being published by highly regulated and controlled digitization projects and by less regulated informal digital projects. On the one hand there are expanding opportunities and new discoveries. On the other, there are increasingly fragmented libraries and difficulty in finding and accessing resources. How did we reach this point? What are the problems we now face and what might the appropriate solutions be?

In the technology world, it is hard to remember the time when certain technologies were not available. Technological progress is now so quick and technology itself so ubiquitous that only those with longer memories remember the first commercial e-books and the difficulties involved in their use. However, the commercial publishers were not the first to understand the potential of digital content in the humanities and it was the efforts of individuals, small groups of interested amateurs, and academic experimentation that initially fostered the creation of digital resources for the humanities.

The claim to the first e-book and the first digital library is made by Project Gutenberg, which declares the first material digitized for free distribution was the United States Declaration of Independence in 1971 (Hart, 1992). From this small beginning over a period of 45 years, developed a digital library that now numbers over 50,000 volumes, published in multiple e-book formats. This library has also spawned a host of partners and affiliates in different countries, offering a further 100,000 volumes in different languages and formats (Project Gutenberg, 2016).

However, the path to this type of digital library was not smooth, and illustrates many of the problems that these early digital adopters faced. The first is the lack of standards or consistency. Project Gutenberg in particular adopted a policy of non-prescription for its supporters, so that the emphasis was placed on making works available, on experimenting, and on exploring new tools and methods, rather than on providing a cohesive, standardized library (Hart & Newby, 2004). This led to many of the works in the library being described as “works in progress” in that they may continue to be developed, improved, corrected and changed. While this does fit with the the idea of experimentation and development, it poses problems for the use of the texts in study. How authoritative are the texts? At what point are they declared finished? Are earlier versions available for comparison? What kind of citation is required to cope with the idea of a changing resource, and how useful is content “in progress” for academic study?

Even where there are standards, these are often not observed, as the standards may be aspirational rather than actual. One example from the recent past may serve to illustrate this point. Although the resolution standard recommended for digitizing materials by the United States Li-

brary of Congress at that time was 400 dpi, there were projects at the Library of Congress itself that were digitized at 300 dpi rather than the recommended 400 dpi. The reason given for this discrepancy was that the project had limited funding and that they simply could not afford the storage that would be required for items digitized at the standard of 400 dpi, so a lower resolution that did not require so much storage was adopted. The current Library of Congress standards have now adopted language that suggests the standard should be the highest resolution and bit depth available (Library of Congress, 2016. See also the Federal Agencies Digital Guidelines Initiative, 2017). This offers the dual benefit of keeping pace with technological advances, while at the same time recognizing that not every project may have the resources or funding necessary to meet the highest current technologically possible standards.

Another problem is the type of digitized material that is provided. Project Gutenberg has committed itself to the provision of completely reset text in machine readable form, at the lowest common technological denominator of basic text. This makes the item accessible by the largest number of people, it makes it searchable and analyzable, and it makes it transferable into other formats. It also uses the minimum of storage space. In contrast, other digitization projects have digitized text as pictures, resulting in items that are not searchable or analyzable, are not transferable into other formats, and which require large amounts of storage capacity.

Quality control can be an issue for all kinds of digital libraries. Issues range from missing pages to unreadable pages due to poor scanning technique or tightly bound gutters in the source material, rendering scanning problematic or impossible. Depending on the source material location and condition, these problems may not be solvable, even once the problem has been identified. Source material that is not paginated is particularly difficult to check, as only a detailed reading of the material will reveal any omissions, and even then they may not be obvious.

Several digital libraries make use of the distributed and asynchronous nature of the internet to employ an army of volunteers in their proofreading and correction efforts. Although not dedicated to the study of language and literature, both the Choral Wiki, founded in 1998 as the Choral Public Domain Library, and now hosting over 25,000 scores of cho-

ral and vocal works by over 2800 composers, and the International Music Score Library Project, also known as the Petrucci Music Library, started in 2006, and offering over 352,000 music scores of almost 107,000 works by over 14,000 composers, make extensive use of volunteers for various proofreading and editing tasks (ChoralWiki, 2015; IMSLP, n.d.). The Choral Wiki does actually support the study of texts set to music by its ability to search the library for authors or specific texts, as well as the expected search for composers and work titles.

Software development

The distributed nature of the development of digital libraries has meant that many people were involved in developing software for different applications within the digital library world, such as reading tools and collection management or access tools.

An interesting example of a reading tool is the software developed by the British Library, Turning the Pages™ (British Library, n.d.). Development began in-house in 1996 as a way to display rare books, manuscripts and single page documents, but from 2001 the Library has been in partnership with a commercial software developer (Turning the Pages™, 2017). The scope and versatility of this project is such that it is able to display items from multiple cultures. Western style books are able to be read by turning the pages from right to left. If appropriate, though, the software can turn the pages from left to right. Scrolls or parchments can be rolled from one side to the other, or from top to bottom, depending on the language and the reading direction. The software is also capable of displaying two versions of a text on a split screen either side by side or top and bottom. When the software was first promoted, it was possible to view the first and second folios of Shakespeare together, so that textual differences between the folios were more easily identified. (At the moment this particular opportunity does not appear to be available on open access from the British Library – ironically in this, the 400th anniversary of the playwright's death.)

The move from an in-house software project to a partnership model with commercial software developers also highlights the difficulties

brought about by rapid technological change. Libraries and individuals are not well equipped to keep up with the demands of ever-changing devices and software standards, so the technology that is supposed to provide open access becomes a barrier to access, and a new barrier that must be overcome every time there is a change in computing platforms and machines. Oliver and Knight (2015) note that problems can also arise because of the very different expectations and philosophies of librarians and archivists compared to those of information technology professionals.

Other software projects designed for open access and digital libraries are Greenstone, open-source software for building and distributing digital library collections, first developed by the New Zealand Digital Library Project in the year 2000 in cooperation with UNESCO and the Human Info NGO (New Zealand Digital Library Project, n.d.), and Omeka, a product of the Corporation for Digital Scholarship, at George Mason University in Virginia, begun in 2007, whose aim is more to display digital collections and manage digital exhibitions (Corporation for Digital Scholarship, 2015).

Copyright problems

An issue of great concern in any discussion of digital libraries for the humanities is that of copyright. There are several different concerns over copyright, including multiple changes in copyright law, the aggressive extension of copyright terms, and the inherent conflict of open access digital libraries complying with multiple jurisdictions of copyright regulation. When many of the original digital humanities libraries were started, the nature of what they were trying to achieve – often simply making available materials that had long been out of print and out of copyright – offered few opportunities to run foul of existing copyright laws. Although the libraries were thinking about open access, they also tended to be more regional in outlook, and there do not appear to have been many concerns with complying with copyright laws in multiple jurisdictions. Now, however, almost all the digital libraries include very detailed copyright notices verifying compliance with the laws of the country in which the hosting servers are located (for example, the Petrucci Music Library follows Cana-

dian copyright law (IMSLP, n.d.), while the Choral Wiki follows the copyright law of the United States.)

Multiple changes in copyright law have occurred in the past 40 years. The copyright laws of the Untied States, for example, have changed from a copyright period of 28 years, renewable for another 28 years, which was in force when the Gutenberg Project was launched, through automatic renewal of the copyright term, to life of the author plus fifty years, to the current life of the author plus 70 years (Association of Research Libraries, n.d.). Even this does not fully capture the complexity that must be faced, as different regulations apply to works for hire, anonymous works, works published with or without copyright notices and so on (Hirtle, 2017). At various points, works that were in the public domain became protected again under new copyright laws, and digital libraries discovered that works that they were legally able to distribute one day were no longer available to them the next and had to be removed.

There has also been an increasing impact on copyright protection by various trade treaties that have sought to override the copyright legislation of individual countries and push for adherence to the longest copyright terms of participating countries.

The whole idea of open access is that it is unrestricted (Berlin Declaration, 2003). The nature of a digital library is that it is not bound by space or time, so its collection is available 24 hours a day, seven days a week, to anyone who has the equipment to access the content. There is a direct conflict to the ideals of both open access and digital libraries if content is restricted based on geographical location and conflicting copyright terms. But this also raises the issue of how logical it is that something is public domain and freely accessible in some parts of the world, but not in others.

Some digital libraries have taken a very direct approach to this problem. The International Children's Digital Library, originally a project of the University of Maryland in the United States, has adopted a policy of digitizing both public domain children's books and then simply asking for permission from the copyright holders to digitize books they wish to have in their collection. Although they are not successful all the time, they have received permission in many cases, thus providing an interesting collection of old and new children's books available in multiple languages (International Children's Digital Library, n.d.).

Coming together and splitting apart

The digital library world is a strange place of both coming together and splitting apart. There are numerous instances of collaborative projects – the Greenstone open source software, for example, or Europeana, as a way of preserving cultural heritage (New Zealand Digital Library Project, n.d.) – but there is also a trend of splitting apart, of libraries cannibalizing each other, and of a plethora of digital repositories that in many ways prevent the easy sharing of information rather than promote it.

It is interesting to see some very old-fashioned library collection building going on in the digital library world. In an online environment, it should be sufficient for there to be an open access source of a document with a clear retrieval path. There should not be a need for multiple versions of the same document contained in different digital libraries, with different retrieval paths, unless the point of the digital library is not to provide access, but to increase the size of its collection. However, there are many instances (and many of the policies of many digital libraries seem to encourage this) of people downloading a document from one library only to add it to another. (I can actually trace a music score I scanned in the 1990s from its original digital home to three other digital libraries, over the course of 15 years. The original of the scan was a photocopy with several unique identifying marks, so tracing the score's journey was easy.)

Another aspect of this is the cooperative agreement, whereby digital libraries are contributing their digitized content either by linking through a central portal, or by allowing portals to mount their collections on mirror sites. The Hathi Trust digital library is a good example here, where a small number of large academic collections collaborated to provide free access to their digitized materials (in compliance with copyright restrictions). At its conception in 2008 it was limited to certain large American universities or university systems, but now includes many other universities and research institutes, both from the United States and beyond (Hathi Trust Digital Library, n.d.).

The emergence of the digital repository in academic institutions is both a positive and a negative in the digital library world. While the goal of providing open access to university–sponsored academic work is to be

supported, particularly in the face of increasingly expensive and restrictive publishing in the academic journal market, the fact that each institution is providing its own repository, running on different software, with different access rules, different finding aids, and different formats, makes the discovery of appropriate documents problematic. Online directories, such as the OpenDOAR (University of Nottingham, 2014), are useful as long as they are kept current, but it is not always easy to determine how current a particular directory is, for example, based simply on the date of the hosting webpage.

Search tools can go some way towards bridging the gap, but are dependent on the consistent use of metadata and formatting standards. OAIster, now run by OCLC, and available through the WorldCat interface, is an example that began as an attempt to provide a union catalog of the open access digital resources from research libraries, but has now greatly expanded through automatic harvesting of records and institution contributed records (OCLC, 2017). However, the results of any automatic harvesting are only as good as the original data. How well original cataloguers adhered to standards and how they interpreted certain rules, or applied them in their own situation can greatly affect the search results.

Nor are the discovery problems confined to academia. The Europeana project offers a slightly different example, this time of a web interface for a single project that again depended on a wide variety of participants. While its finding aids are useful, they are confined to collections that were defined as Europeana, so again they offer a portion of what is available, rather than a comprehensive offering (European Commission, 2015). Wikipedia offers an example of a third option, with a simple directory of open access resources around the world, where the user has to search each website in turn for the desired materials (Wikipedia, 2017). A note on the Wikipedia article says it is out of date, but of greater concern is the lack of any description of the method of defining or identifying digital libraries, and therefore of the reader knowing what may have been missed.

Where to does this leave us?

The difficulty is in finding a meta search engine that is smart enough to identify open access digital content, that is not bound by geograph-

ic, national or language boundaries and that can cope with copyright restrictions. In other words the problem, as so often it is in the information world, is not in the mechanics of providing the information itself in digital form. The process of scanning or keying in the data is well understood. Neither is it in providing accessibility, as again the procedures of mounting web pages and providing web interfaces are also well understood.

The problem is in the organization of information – in the cataloging and metadata. Cataloging has traditionally been about providing access paths for a defined group of users. Items may be catalogued differently in one situation than in another, depending on how the final users are most likely to search for information. Providing metadata that works for all users in all situations, across national and cultural boundaries, and in multiple languages, is a much more complex proposition, and not one that can be undertaken easily.

The development of new metadata standards, such as Dublin Core (DCMI, 2017), has laid the groundwork for this, but the challenge is how this is applied to informal digital libraries, to small collections that lack professional cataloguers, and not least how it is applied retroactively to existing collections.

It is not hard to give access to 30 million documents and we now have easy ways to do this. What is much harder, and the problem we must now address, is how to give open access to the much smaller number of documents that the researcher actually wants and needs, ensuring that we are offering everything that is available, and not only what is easily discovered. We need intelligent search systems that can match the skills of an experienced librarian.

Until this happens, we appear to be using technology to replicate the problems of the analog world – unless we know to go to a particular library, we will miss out on its riches.

References

Association of Research Libraries (n.d.). *Copyright timeline: A history of copyright in the United States.* Retrieved on June 5, 2016 from http://www.arl.org/focus-areas/copyright-ip/2486-copyright-timeline

Berlin Declaration (2003). *Berlin declaration on to knowledge in the sciences and hu-*

*manities.*British Library (n.d.). *Virtual Books.* Retrieved May 7, 2017 from Choral-Wiki (2015). *Welcome to ChoralWiki, home of the Choral Public Domain Library.* Retrived April 30, 2017 from http://www2.cpdl.org/wiki/index.php/Main_Page

Corporation for Digital Scholarship (2015). *Omeka.net.* Retrieved May 6, 2017 from DCMI (2017). *Dublin Core Metadata Initiative*. Retrieved from http://dublincore.org/about-us/

European Commission (2015). : A European digital library for all. Retrieved from

Federal Agencies Digital Guidelines Initiative (2017). *FADGI guidelines.* Retrieved on April 30, 2017 from http://www.digitizationguidelines.gov/

Hart, M. (1992). *The history and philosophy of Project Gutenberg.* Retrieved from https://www.gutenberg.org/wiki/Gutenberg:The_History_and_Philosophy_of_Project_Gutenberg_by_Michael_Hart

Hart, M., & Newby, G. B. (2004). *Project Gutenberg principle of minimal regulation.* Retrieved from https://www.gutenberg.org/wiki/Gutenberg:Project_Gutenberg_Principle_of_Minimal_Regulation_/_Administration_by_Michael_Hart_and_Greg_Newby

Hathi Trust Digital Library (n.d.). *Our partnership.* Retrieved May7, 2017 from https://www.hathitrust.org/partnership

77IMSLP (n.d.). *IMSLP Petrucci Music Library*. Retrieved April 30, 2017 from http://imslp.org/

International Children's Digital Library (n.d.). *International Children's Digital Library: A library for the world's children.* Retrieved April 30, 2017 from http://en.childrenslibrary.org/index.shtml

Library of Congress (2016). *Library of Congress recommended formats statement 2016-2017*. Retrieved from https://www.loc.gov/preservation/resources/rfs/RFS 2016-2017.pdf

New Zealand Digital Library Project (n.d.). *Greenstone digital library software.* Retrieved, May 7, 2017 from OCLC (2017). *The OAIster database.* Retrieved May 1, 2017 from http://www.oclc.org/en/oaister.html Oliver, G., & Knight, S. (2015). *Storage is a strategic issue: Digital preservation in the cloud.* D-Lib Magazine, 21(3/4). DOI: 10.1045/march2015-oliver

Project Gutenberg. (2016). *Free ebooks by Project Gutenberg.* Retrieved on June 4 2016 from https://www.gutenberg.org/

Turning the Pages™ (2017). *Turning the Pages: The leading digital facsimile software for rare books.* Retrieved May 7, 2017 from http://ttp.onlineculture.co.uk/

University of Nottingham (2014). *The Directory of– OpenDOAR.* Retrieved April 30, 2017 from Wikipedia (2017). *List of digital library projects.* Retrieved May 7, 2017 from https://en.wikipedia.org/wiki/List_of_digital_library_projects

Ендрју Смит
Школа за библиотекарство
и управљање информацијама
Емпориа Стејт Универзитет

ОД СКРОМНИХ ПОЧЕТАКА: УЛОГА НЕФОРМАЛНИХ ДИГИТАЛНИХ БИБЛИОТЕКА КАО ПОДРШКА Е-ФИЛОЛОГИЈИ

Сажетак

Иако је у библиотечком свету акценат стављен на дигиталне библиотеке великих размера покренутих од стране издавача или непрофитних корпорација, развој дигиталних библиотека од скромних почетака до значајних дигиталних ресурса који подржавају е-филологију је заправо прича о самониклим организацијама, групама заинтересованих аматера, независним библиотекама и академским институцијама, које су експериментисале са форматима и садржајима и тиме поплочале пут већим комерцијалним подухватима. Академске библиотеке могу нарочито бити усредсређене на такав начин да подржавају употребу база података и е-ресурса који су прибављени уз велике трошкове, да не успевају да промовишу друге, бесплатне ресурсе који омогућавају приступ широком спектру дигиталних садржаја. Рад разматра развој дигиталних библиотека и наглашава разноврсност бесплатно доступних ресурса који представљају подршку учењу и истраживању. Алати и стратегије за откривање и промовисање ове врсте дигиталних извора ће такође бити узети у разматрање.