Digital Library – Economies of Scale

Shreshta Shyamsundar[1*], Kamalkumar Rathinasamy[2]
[1, 2]Infosys Limited, Mysore, India.

* Corresponding author. email: shreshta_shyamsundar@infosys.com

Manuscript submitted March 15, 2014; accepted June 11, 2014.
doi: 10.7763/ijeeee.2014.V4.344

Abstract: This document helps demystify the phrase “digital library” and lends it relevant context from its purpose in the real world. In the process, this artefact also attempts to shed light on

- Its importance,
- What it takes to setup a digital library
- Revenue model - Sustainability
- Future of this concept in digital library

Essentially in the broad sense of the phrase, a digital library system is built on common standards and methodologies therefore when seen from the perspective of this volume; digital library does appear as a seamless amalgam of the conventional document set important from the historical perspective, along with transient material with unending activity.

Key words: Digital library, digitizing, e-library, library economics, how-to e-books.

1. Introduction

Amongst major crises plaguing the society at large today is the issue of quality information being out of reach of the common man, particularly those locked away in books and other traditional media forms. If this information was available in easy reach, several changes for the better might come upon us i.e., making research that much easier for the scholars today, easing budget pressures on the libraries. This bit can be addressed by virtue of digitizing content and making it available to access over a personal computer.

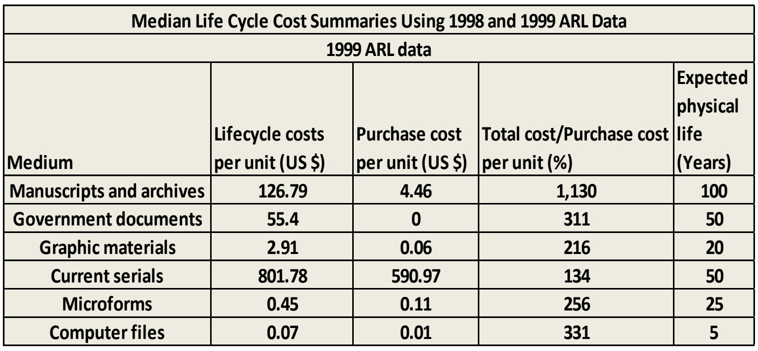
A digital library is an electronic collections of resources (reading material, research papers, journals et al) are stored in electronic media formats and accessible via computers amongst other electronic devices i.e., tablets, smartphones etc. The electronic content being referred here may be stored locally, or accessed over computer networks.

2. Why Digital Library

Given that universities libraries and district central libraries are finding it virtually impossible to maintain traditional book collecting practices within the budgets allocated, a digital library is indeed the need of the hour.

Between 1986 and 2004, journal expenditures of North American research libraries increased by a staggering 273%, with the average journal unit cost increasing by 188%. During this same period, the U.S. Consumer Price Index rose by 73%, meaning that journal costs have outstripped inflation by a factor of almost 4 [1].

Another problem faced with libraries operating in the current frame of things is being beset with increasing costs for buildings and storage. Some of the books have aged quicker than the others demanding better preservation techniques to keep them in the best shape possible. Below outlines some of the statistics based on research done by SR Lawrence on the topic of – *Life cycle costs of Library Collections* [2].



Median Life Cycle Cost Summaries Using 1998 and 1999 ARL Data

1999 ARL data

Medium	Lifecycle costs per unit (US $)	Purchase cost per unit (US $)	Total cost/Purchase cost per unit (%)	Expected physical life (Years)
Manuscripts and archives	126.79	4.46	1,130	100
Government documents	55.4	0	311	50
Graphic materials	2.91	0.06	216	20
Current serials	801.78	590.97	134	50
Microforms	0.45	0.11	256	25
Computer files	0.07	0.01	331	5

Fig. 1. Life cycle costs of library collections.

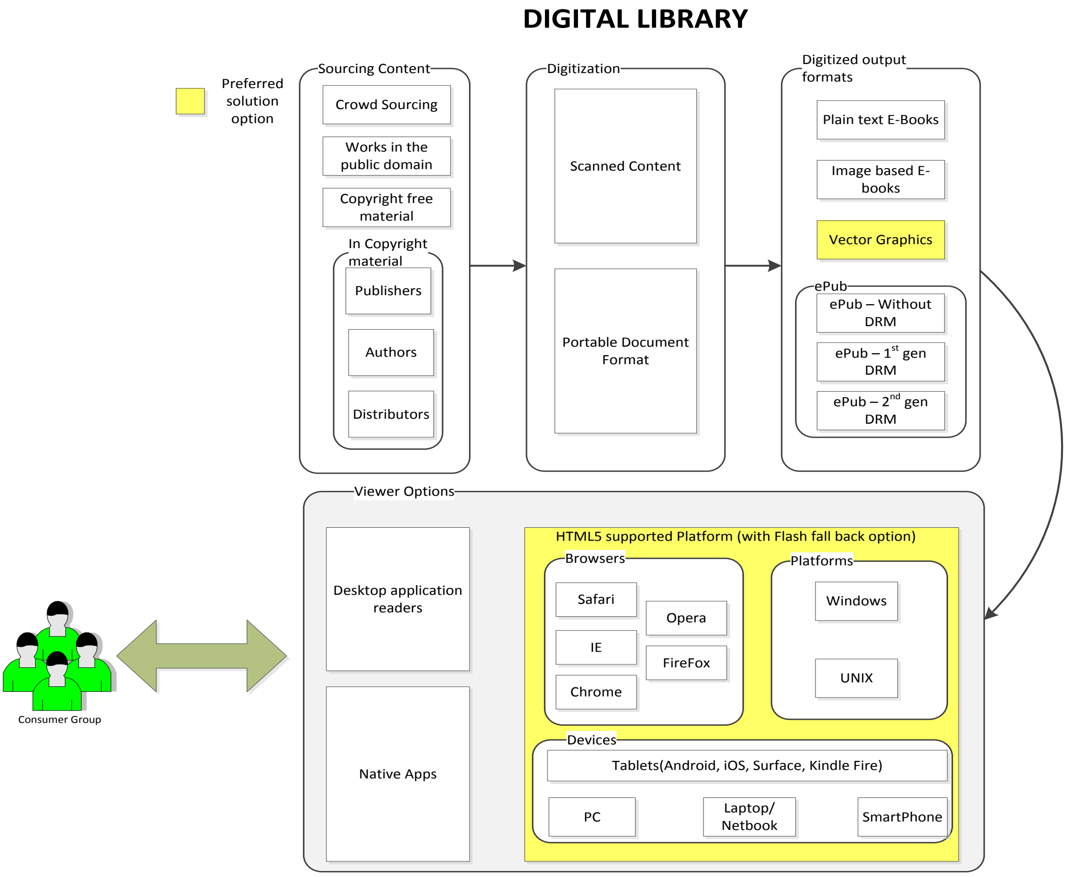


Fig. 2. Lifecycle of the digital library.

Essentially this underscores one point; the system in its current state is near collapse. However digital library in principle can address both factors of preservation and collection of material.

On an average an American university spends a third of its budget on purchases; of each dollar spent about three-tenths goes to the publisher/author with the rest feeding the chain. To summarize, only a tenth of what the university spends on its library is paying for the original creation of information.

If we imagine even for a second that a new distribution system be introduced which sent information directly to the students' desks, most of the money in the system is available to pay for it; need for the digital library seems justified.

3. Sourcing Content

It forms a visual representation of the sequence of steps that go into formulating a digital library at a high generic level in Fig. 2.

The foremost point about setting up a digital library as suggested in the visual representation above is towards sourcing content. Crowdsourcing is an act of outsourcing routine chores otherwise taken up by employees to a group/community where the benefits obtained from the deal are proportionally shared. As a technique it has proved to be a very effective symbiotic solution from both perspectives.

Another way of sourcing content is through getting content present on the public domain i.e., W*orks in the public domain are those whose intellectual property rights have expired, have been forfeited, or are inapplicable with examples including the works of Shakespeare. Books published before the 19th century is in public domain and the copyright status of books published since 19th century varies by country* [3].

One very popular single source for public domain books is the website, *Gutenberg.org* where there are 42,000+ free e-books available for download and reuse. Below illustrates the growth of the publications at the Gutenberg project between 1994 and 2008 [4].

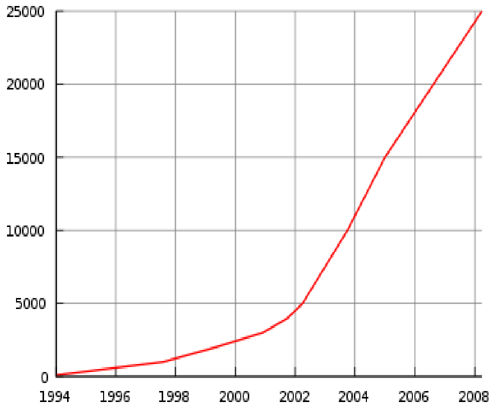


Fig. 3. Growth of publications at the gutenberg project.

A simple yet effective way towards sourcing content is to source copyright free books i.e., all books licensed under *CC BY*, *CC BY-SA*, *and CC BY-ND* can be included in the catalog. *Pratham* is a popular 'copyright free' children e-books publisher and they are redistributed from many e-book resellers or digital library

There are websites where articles are classified under creative commons (CC). If their CC licenses permits us to *copy, remix, transform, build upon the material and redistribute the material in any medium or format for any purpose, even commercially*, then the contents can be compiled in to e-books.

Importantly, the 'Copyright status' of e-book has to be checked before adding them to the digital library catalog. Given that adding huge volumes of e-books might not necessarily add value to the digital library; the goal should be to select and add appropriate books for the target readers.

The in-copyright e-books that might be of interest to our target group can be obtained by approaching their respective copyright owners for their permissions to redistribute their e-books in our digital library.

1) Authors / Publishers: The copyrights of the e-books, depending on the contracts, might be with either the author or the publisher or both. Since e-books are considered as an extension to paperbacks, majority of the time, publisher would own the rights but they might contact the respective authors for their permissions. It is time-effective to approach publisher as they, in general, would own more books than an author. Either the rights for resale can be bought for a specific period of time or the copyright

owners can be included as business partners and so they get a share of digital library subscription revenue.

2) Distributors: e-books rights can be bought from distributors/wholesalers for fixed period of time. One such popular e-books wholesaler is the *Gardner books*. There are options to become e-book retailer but that would not suit the digital library platform where revenue is generated from the subscription of the library package and not from individual e-books.

Another concept towards sourcing content would be a direct interfacing with an e-book provider akin to the integration maintained between Houston Public Digital Library with that of *Overdrive*. Any offline integration which requires our digital library customer to login to a third party website (instead of auto redirect and backend/server-to-server authentication) to access/read e-books might not be user friendly.

4. Digitizing Process – Criterion Considered

Digitization is a process of creating a digital representation of a document, image etc. i.e. converting the currently analog component into a digital one or outlining a document with respect to its content and orientation into a digital representation.

Steps to digitize: While the process would vary depending on the choice of technology used in building the Digital library platform (website); the steps (guidelines) below are generic to be considered regardless of technology used.

1) Processing Source Format
   - If the source format is a pdf, go to step (c.b).
   - If the source format is a scanned document, check the file size of the images. If it is more than your threshold value (say 200kb), reduce the size of the image to an extent beyond which it loses readability.
   - The DPI i.e., dots per inch settings can also be adjusted to get required results.
   - The image dimensions can be reduced without affecting its aspect ratio for better results.

2) Content Analysis
   - Post the actual digitizing process of recording content into a digital form; the output is sent for analysis to make sense of the data collected.
   - One of the most effective analysis techniques being employed today is the *OCR i.e., Optical* Character *Recognition*; essentially this process attempts to convert what is perceived to be an image into basically, text. The rationale behind this step is to enable the user to search through the content for relevant information. The process of OCR isn't foolproof unfortunately as certain characters belonging to certain languages may be too cryptic to process. reCaptcha from Google used by Google books to essentially help convert some hard to detect words/phrases best decoded by the human eye.
   - Further analysis is conducted towards formatting the document towards identifying the subsections of the document i.e., numbering, anchor sections, support sections etc.

3) Converting to Target Format
   - Merge all the scanned images in to a single pdf. Any pdf writer, like nitropro, can do this operation.
   - There are commercial tools, like 3dissue, which provides the html5 book reader, flash book reader and the associated tools for conversion and configuration. Choose a tool, which provides options to block downloading and, provides options to pre-fetch contents.
   - Convert the book pdf into individual page swfs and list them in the same order in the metadata file in the html5 book reader and in the flash book reader.
   - Configure the options in the reader so that download is not allowed and pre-fetch is allowed.

The books in the digital library has to be in a format where the following criterion is met: *the books should not be downloadable on to user's machine outside the scope of digital rights management a.k.a. DRM; the contents of the books must not be subject to duplication; however at the same time the books should be accessible and readable in all popular platforms even at low bandwidth.* Digitizing is a process to convert source books to desirable format which meets above mentioned criteria at low cost in quick time.

Below is an attempt to go over the effectiveness of some of the listed options which are available in the market;

Book reader: is an application which allows user to read books. A book reader application can be made more effective if it can utilize the power of the underlying hardware and software i.e., the application will have to run on the user machine in some form post being downloaded and installed.

The issues with the downloadable installers, in general, is as discussed below

1) Multiple versions of the same book reader application have to be built to support variety of software platforms like windows / Unix and hardware platforms like desktops / tablets / smartphones. There would be no end to the amount of versions that needs to be supported. Majority of the book reader application developers would cater the needs of target group's specific platforms or would build applications only for popular platforms. Essentially these form additional steps that is capable of slowing down the time to market.
2) It is a common deduction from our daily life that *usability of a feature is inversely proportional to the effort required to use the same*. Hence, downloading and installing the book reader might not be preferred among the library subscription members.

An alternate to desktop book reader application is browser based solution which does not require end user to download/install. Prior to HTML5, browser's native language was not suitable enough to develop a user friendly book reader feature. The other option considered was to have it built using plugins supported by the browsers i.e., Flash/Applet/Silverlight.

Installing browser plugins may also be construed as an action reserved for advanced users as do the process of installing the book reader desktop application.

However with the advent of HTML5, the need for the plugins has diminished to a large extent. A book reader can be developed just in plain HTML5 without the need for browser plugins.

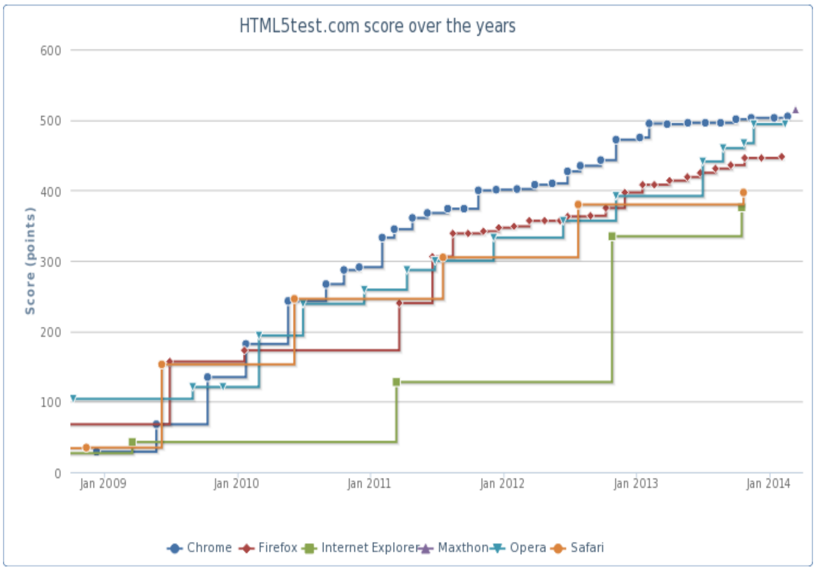


Fig. 4. Rise in acceptance of HTML5 with popular browsers.

While most browsers today have started the support of HTML5, this is an ongoing exercise therefore this will take some time for them to fully support HTML5. This brings up another point which is that, not all the digital library members would have upgraded their browsers to latest version. Hence the most feasible solution one sees is to develop the book reader using standard browser plugin software and also in HTML5. The default option is to display the book reader in HTML5 and if the book reader detects that HTML5 is not supported in client browser then it should fall back to Flash based solution.

It is a snapshot in Fig. 5 showing the rate of the popular browsers have adopted HTML5 [5].

The only downfall with this approach is that if user does not have the latest version of the browser and the flash plugin, then the book reader cannot function unless he/she upgrades or installs the required software.

These book readers need not be developed from scratch as there are many similar third party book reader software viz., 3dissue. All these software provide the options to download the book and hence they need to be configured such that download options are not provided to the user.

Ebooks traditionally come in the following forms:

1) Plain *text*: The book contents can be delivered as plain text. The advantage of plain text is that it is light weight and it can be crawled and indexed by search engines. Hence, even if any search text in search engines match any section of indexed book contents then the book's page in digital library website will be displayed in search output. The issue with the plain text is that it can be copied and so it can be reprinted/reused easily.
2) Images: An alternate book format is to deliver it as images (almost like the scanned book pages). The advantage of images is that the text cannot be copied and reprinted. The disadvantage is that scanned images are not vector images. Hence, unless the book reader displays the page in the same area as the image dimensions, the text in the image will not be readable. The book reader page dimensions are dependent on target machine and so it cannot fixed either. Another disadvantage of image format is its file size. Images are always heavier than plain texts.
3) *Vector graphics*: If the book pages can be delivered as PDF(s) or SWF(s) then the content is as secure as it was in images as they cannot be copied/reprinted/reproduced in any way. Generation of PDF(s)/SWF(s) can be tuned to restrict their output file sizes. Since these are vector graphics formats, regardless of the book reader dimensions or target platform's viewport area, the book contents will be redrawn to fit the view port accordingly.
4) *ePub*: is a free open ebook standard designed for reflowable content i.e. type of content that has the capability of adjusting its presentation based on the viewport it's being rendered upon. Hence, the ePub can adjust its texts to match the target layout. The disadvantage of this format is that an ePub reader is required at user machine to process ePubs. The popular browser Mozilla Firefox has an extension to process ePubs which means that user has to install the specific extension in lieu of a desktop application. Also, ePubs would be downloaded to user machine and hence it is can be easily shared with others.

Digital Rights Management (DRM) is a framework that is frequently employed by publishers, and individuals who hold copyrights, hardware manufacturers with the intention towards controlling the use of their digital content. First generation DRM had the agenda of controlling any possible duplication/copying; however the 2nd-generation DRM possessed the agenda of extending the control from duplication to that of executing, copying, printing, viewing, and altering of works. A section of the demography considers the presence of DRM violates private property rights and restricts normal and legal activities of the user.

A DRM component would control a device a user owns by restricting how it may act with regards to certain content, overriding some of the user's wishes [6].

Source content format: The traditional paperback publishing is completely digital now which means that the publisher would have the pdfs of the latest books. Books published before 15 years might not be available in pdf format.

1) *Source pdfs*: Books available in pdfs are easier to process. The entire process of digitizing can be automated if the source books contents are provided as pdfs. They can be either converted to individual pages pdf or can be converted to individual pages swf, depending on the book reader application's choice.
2) *Scanned images*: Books available as set of scanned images cannot be converted to any other formats. To overcome the issue of loss in clarity when images are stretched or squeezed to fit the view port, multiple versions of images can be generated for different dimensions. Most common dimensions would be thumbnail size, tablet display size, desktop size (small) and desktop size (large). Appropriate sized images has to be downloaded to match the target machine (book reader) display size. There are attempts to read the text from scanned pages using OCRs. These attempts along with manual verification has achieved limited success for texts in popular language like English. Otherwise, it is not worth the effort to read text from scanned documents in the digital library project. However there are intelligent schemes on the market that work with this setup working out a performance angle to it i.e., when user is reading a page, the next set of pages can be preloaded so that there is less or no delay when user navigates to next page.

5. Sustaining the Concept of Digital Library – Revenue Model

A variety of revenue strategies are in the works to keep this concept in the works covering partnership with publishers/resellers, subscription with varying entitlements including trial models.

Scholarly Publishing Business Models and Online Corollaries [7].

Print Model	Revenue Stream	Market Features	Online Corollary
Controlled circulation	Advertisers or external funds	Requires high market coverage; Seldom used for refereed journals	Free, ad-supported site; or free with print and registration
Personal or Member subscriptions	Individuals	Refereed; tends to large audience; may be more general	Free or small fee with print; may allow online only
Institutional subscriptions	Libraries	Small, specialist audience; tend to higher prices	Site-wide for fee or free with print
Mix	Members and libraries (+ads?)	Complex interplay of markets	Unsettled, but usually fee + print required

Fig. 5. Popular business models with their attributes.

As indicated above, the revenue model for the controlled magazines is via advertising by virtue of putting up their content for being accessed for free. Personal subscriptions for Journals are usually the way monetization is achieved. Needless to say, some of these publishing houses may use a complicated mix of revenue options to monetize their offerings.

Overall it would be safe to say that if the circulation of the magazine/generally the material is in the four figures it is probably library-subscription driven. Also it could be the fact that these journals might be serving a very small demography therefore is sustainable only through the subscription model.

Other than the ideas discussed above, few publishers also toy with other ideas viz.

- Allowing customers to view a few sections to get them interested and then into subscribing or buying the material outright.
- Restricting the material only to certain geographies.
- Restricting the number of views based on payments i.e. in a pay-per-view model.
- Some even incentivize the public to move to an online mode of access from the print by offering the same content at discounted rates as its cost effective for the publisher as well.

6. Popularity and Acceptance of the Digital Library

Over the holiday season, a survey found that number of American adults who had read an e-book rose in the last year grew from 17% all the way to 21%. Even the e-book ownership (Kindle/Nook) jumped from 10% to 19% over the same time period. In fact e-book readers have communicated that they have read more books in all formats. Statistics are described in the graph below [8].

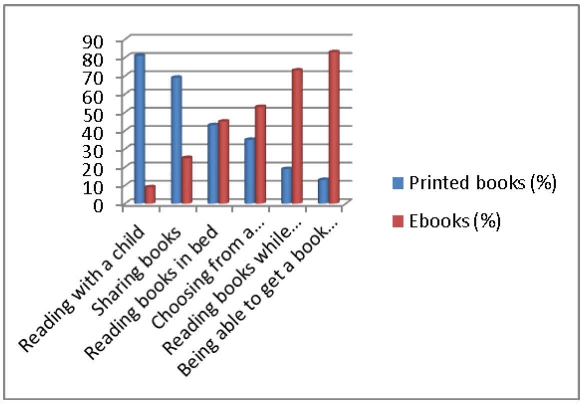


Fig. 6. Popularity of e-books across demographic.

7. Conclusion and Where We Go from Here

Except for areas like intelligence and defense, the US government budget constraints and shifting priorities make it tricky to establish research initiatives in the field of the digital library.

Currently the community of the digital library in the US is primarily supported by investments from the national science foundation amongst other sources.

Data curation essentially is a term to indicate management or preservation of data keeping in mind the purpose of reuse, extraction and conversion of scholarly articles by experts into a more manageable electronic format i.e., database. According to the University of Illinois' Graduate School of Library and Information Science, "*Data curation is the active and on-going management of data through its lifecycle of interest and usefulness to scholarship, science, and education; curation activities enable data discovery and retrieval, maintain quality, add value, and provide for re-use over time*"[9].

Needless to say, the community has a lot of work cut out in this aspect and having said that, it combines the need for high performance computing and research into complex systems to achieve this to the extent necessary.

While it is important to preserve the works done by great scholars and researchers at large, it is of equal importance to ensure adequate protection to their work. The concept of DRM i.e., digital rights management is a set of technologies used by publishers and content owners/creators with an intent to control the use of the digital content and associated artefacts. At the moment the view on this topic is indeed divided over its use and purpose articulating its inconvenience to legitimate customers and the other faction arguing its capability to ensure the revenue stream to the content owners.

Given the popularity of the digital library and the usage benefit, large internet corporations in Google, Yahoo draw heavily on using this concept to power their search result technologies.

In a real sense, then, one can view digital libraries as offering a set of engineering techniques that are being harnessed by large tech companies to better concepts we use today and make life a lot simpler while addressing important issues of preservation and data curation.

Also the digital library also offers a platform for Universities and learning institutes to offer the Online learning programs. Distance learning was almost always considered a hype in early 1970's however it's more what the reality describes. Looking at some of the figures from the University of Massachusetts i.e. *the earnings doubled from US$ 7 million in 2001-02 to US$ 14 million in 2003-04. In the United States as a whole in 2006; 914,000 students enrolled (up from 315,000 2 years ago), increasing revenues from US$ 1.58 billion to US$ 4.75 billion* [11].

The overarching theme here is what the internet has achieved in its short span of existence connecting us to the digital information present and scattered.

The next ten years for digital libraries may well be characterized most profoundly by the transition from technologies and prototypes to the ubiquitous, immersive, and pervasive deployment of digital library technologies and services in the broader information and information technology landscape [10].

References

[1] Newman, K. (2009, 6 November). *The Cost of Journals*. The University Library, University of Illinois at Urbana Champaign. Retrieved March 14, 2014, from http://www.library.illinois.edu/scholcomm/journalcosts.html.

[2] Lawrence, S. R., Connaway, L. S., & Brigham, K. H. (2001, November). Life cycle costs of library collections: creation of effective performance and cost metrics for library resources. Retrieved March 14, 2014, from http://crl.acrl.org/content/62/6/541.full.pdf

[3] Boyle, J. (2008). *The Public Domain: Enclosing the Commons of the Mind*. CSPD. pp. 38.

[4] Book Milestones. (2008). *Graph Showing Total Number of Books.* Retrieved March 14, 2014, from http://www.gutenbergnews.org/statistics/

[5] Leenheer, N. HTML5test - How well does your browser support HTML5? Retrieved on March 14, 2014, from http://html5test.com/results/desktop.html

[6] Federal Trade Commission. *FTC Town Hall to Address Digital Rights Management Technologies*. Retrieved on 14 March 2014, from https://public.commentworks.com/ftc-DRMtechnologies/

[7] Scholarly monograph series. *Economics and Usage of Digital Libraries: Byting the Bullet*. from http://quod.lib.umich.edu/s/spobooks/5621225.0001.001?rgn=main;view=fulltext

[8] Pew Internet libraries. (2012). The rise of e-readin. Retrieved March 14, 2014, from http://libraries.pewinternet.org/2012/04/04/the-rise-of-e-reading/

[9] Cragin, M., Heidorn, P. B., Palmer, C. L., & Smith, L. C. (2007). An educational program on data curation, *Proceedings of ALA Science & Technology Section Conference*.

[10] Lynch, C. (July/August 2005). *Where do We Go from Here? The Next Decade for Digital Libraries.* Retrieved March 14, 2014, from http://www.dlib.org/dlib/july05/lynch/07lynch.html

[11] Lesk. (Spring 2006). *Digital Libraries: The Future*. Retrieved April 16, 2014, from http://comminfo.rutgers.edu/~lesk/spring06/lis553/L-futures.pdf

Shreshta Shyamsundar was born in Dubai, U.A.E on the 19th of August 1982 and received a bachelor of engineering in the discipline of information science & engineering from VTU, Belgaum, India in 2004. He has over 10 years of industry experience and is currently a technology architect with Infosys Limited, Mysore. The Author also holds a US patent #US20120254857 in the field of software asset management.

Kamalkumar Rathinasamy was born in Coimbatore, India on the 3rd of October 1977 and received a master of computer applications from Anna University, Tamilnadu, India in 2001. He has around 13 years of industry experience and is currently a senior technology architect with Infosys Limited, Mysore.