

ON NORMS AND STANDARDS OF DIGITIZATION

Dr. Manish Kumar Singh,

*Information Scientist, Central Library,
Banaras Hindu University, Varanasi*

ABSTRACT

Libraries develop their collections over a long period with large proportion of it comprising of out of print and out of copyright books. These out of copyright books can be safely digitized, in full and provided to the users over the intranet or internet to benefit the user community. The copyrighted books cannot be digitized fully. There are various opportunities and issues related with the digitization work which is a wider concern than a single library. In this paper, various opportunities and issues are identified and some norms and standards are proposed which can be adopted by all libraries, only through library associations, for collaborative goal of digitization. Both copyrighted books and out of copyright books are considered for identification of opportunities and issues.

Keywords: *Digitization, Digitization Standards, Issues in Digitization.*

INTRODUCTION

The well-established libraries all over the world are embarking on their plan for digitization of their collection with a view to the twin purpose of preserving their collection and delivering the digitized material right on the screen of their users' digital device. A physical book, when digitized, can be reproduced and distributed in multiple copies with almost negligible cost as compared to the cost of repeated digitization. So, one physical document should be digitized only once and this is possible only with total cooperation amongst all the libraries. But, their efforts are not yet coordinated by a unified and collaborative effort. This effort is possible only with the association of libraries and more justifiably federation of these associations. Feather and Sturges [1] have discussed the definition of digitization at an early stage of application in libraries.

In this paper, we identify various opportunities with us to utilize the technology of converting a physical document into a digital document with digitization. Thereafter, various real life issues arising in the work of digitization are identified and analyzed. We will be presenting certain norms and standards useful for the work of digitization that can be adopted by library associations and more fruitfully by their federation.

OPPORTUNITIES IN DIGITIZATION

With the availability of technology of converting a physical document into a digital document, we can explore various opportunities for the benefit of library users. Some of the commonly, or not so commonly, identified opportunities are given below:

Preservation: With the conversion of a physical book into a digital book, it will be free of wear and

tear for the rest of life, or more fittingly, for the rest of its infinite life.

Copying and Distribution: There will be no limitation to the number of copies of it to be produced, that too without any cost, and also, to its distribution anywhere in the world with negligible cost. So, it may not be out of print any time.

Searching within documents: Optical Character Recognition (OCR) technology can be used after digitization of the document to make the text of the document, within a document, searchable. This searching capability can be utilized for searching the required document in addition to the use of Online Public Access Catalogue (OPAC).

Inter Library Loan (ILL): ILL makes the borrowing of documents by a library from another library possible. With physical books, this process is very cumbersome and time taking. But, with digitized books and a suitable arrangement between libraries, this can be a matter of few mouse clicks.

IDENTIFICATION OF ISSUES IN DIGITIZATION

Although looking attractive, the digitization entails a lot of problems and issues with the libraries embarking on the work. Some of them are discussed below:

Technical know-how: Various technical parameters of the scanning device, resulting image and file format need to be identified for suitability. In 2017, the International Standards Organization (ISO) issued the technical specification for an image science driven method of assessing the image quality of reflective original sources typically held by cultural heritage organizations [2][3]. Digital archives should be transcribed every ten to twenty years to ensure that they will not become technologically obsolete [4].

Cost: This is the prime factor inhibiting fast digitization. Costly machines require less manpower, but are difficult to afford. Cheaper machines require more manpower and the whole process is very slow,

resulting into greater cost for large libraries. According to Masakazi [5] and Ezeani [6], digitization projects are capital intensive requiring extensive resources in terms of software and hardware technology, highly skilled personnel and other necessary infrastructure requirements.

Copyright issues: Copyright of the book to be digitized and the relevant laws must be looked into before starting its digitization. For delivery purpose, a mechanism should be designed so as to completely prevent the copying by the user and consequent possible commercial use of it. Ownership identification before selection of archival materials for digitization should first be based on a clear understanding of copyright law and rights of ownership [7], [8].

Storage: Scanned documents with satisfactory resolution occupy a very large space on the storage device. Compression of individual pages will lead to deterioration in the resolution, hence readability of the document. Purpose of digitization may be considered while deciding about the storage technology. If preservation is the objective, cheap magnetic storage device, like tape device, may be opted and if the purpose is online delivery, it has to be electromagnetic or semi-conductor device. Apart from this, retrieval efficiency needs to be considered while deciding about the storage mechanism.

Delivery: If the purpose of digitization of the document is delivery to the users, a controlled mechanism needs to be devised for this purpose. Generally, as a policy matter or due to copyright issues, libraries avoid the mechanism of file transfer from the library server to the user's device. To achieve this objective, a mechanism should be in place to prevent the user from copying the document or a part of it. Although with existing technologies, the complete ban is not possible. But, with the use of some intelligent mechanism, it needs to be made difficult and cumbersome for the user to do so.

Repetition of effort: A universal problem associated with digitization is repetition of efforts. Two or more libraries having a common title of

document with them, digitize their holding separately, resulting into repetition of effort, hence multiplication of cost, whereas the technology permits that the same document be digitized only once and copy be distributed to all the libraries. This happens due to lack of coordination between the libraries.

OCR Performance: The OCR scan of the document cannot be expected to give 100% correct recognition of all words. The performance of the OCR scans determines the use of content based search inside the document. Vijayarani and Sikala [9] has provided the overview of various Optical Character Recognition (OCR) tools and analyses of their performance by applying the two factors of OCR tool performance i.e. accuracy and error rate.

PROPOSED NORMS AND STANDARDS OF DIGITIZATION

A brief set of queries was put before ten senior professionals involved in digitization activities and a discussion was organized. As per the responses and general discussions, we have come out with following proposed norms and standards that may be implemented to solve several problems of digitization and achieve its objectives.

Registry of Digitized Documents: Primarily, with the objective of avoiding repetition of efforts, all the digitization efforts should be brought under one umbrella registry keeping record of digitized documents. Any document to be digitized may be looked-up in the registry before digitization. If the cost of digitization is found less than procuring the digital document from the holding organization, procurement may be done to save cost and time. Apart from avoiding repetition of efforts, the registry is expected to induce a proliferation of standards and practices among all institutions participating in digitization informally.

Enabling Integrated Content Based Search of the Documents: The use of digitized document can be enhanced manifold if extensive metadata is

associated with it and its content is searchable with the help of OCR scanning. The digitized document can be made a part of a set or universe of documents in the form of digital library or an openly accessible repository compliant with OAI-PMH enabling harvesting of its metadata by other libraries. Harvested metadata can be stored in different databases to quickly know about the document and its availability.

Devising Secure Delivery Mechanism: Delivering a digitized document through a digital library poses a chance of violation of possible copyright violation or other misuse of the document. In case a digitized document warrants its protection for the sake of copy protection or access limitation, it is necessary that appropriate software mechanism be applied to prevent such use, including copy protection. Online readers integrated into a digital library do provide such mechanisms.

Federated Storage: A digitized document must reside with the rightful owner of the digitized document to ensure full protection and control of the rightful owner of the digital document. Therefore, there may not be an effort to centralize the storage of the document. Instead, small digital libraries may be created under the full control of the owner of documents and these may be integrated into a network of libraries sharing their metadata by way of copy catalogue or harvesting.

Archival: Several digital documents are not in frequent demand. It is very much possible that a document not requested even once for ten preceding years may be demanded very eagerly. The ethics of library science will like to server that user. For such cases, documents are digitized for archival purposes, as well. For a large volume of digitized documents that are likely to be requested very rarely, costly storage space may be saved by storing them on storage devices fit for archival storage, e.g., tape drives.

CONCLUSION

The large libraries, all over the world, have their plan for digitization of their collection with a view to the twin purpose of preservation of their collection and delivering the digitized material conveniently to their user. It was observed that there is a need to integrate the individual efforts of digitization into a collaborative effort following a set of well defined norms and standards. It was seen that OAI-PMH compliant repositories are largely standardized repositories and fulfilling the aspirations of the users. It was a general agreement that library associations and federation of associations have a great role in standards adoption and disseminating know-how.

REFERENCES

- 1) J. Feather and R. P. Sturges, International Encyclopedia of Information and Library Science. 2d ed. London: Routledge, 2003, pp. 138.
- 2) Photography — Archiving systems — Image quality analysis — Part 1: Reflective originals, ISO/TS 19264-1:2017, International Standards Organization, Geneva, Switzerland, 2017.
- 3) D. Wueller and U.B. Kejser, "Standardization of image quality analysis – ISO 19264," Proc. IS&T Archiving 2016, (Washington, DC), (19-22 Apr.), pp. 111-116, 2016.
- 4) G. O. Alegbeleye, "Avoiding technological quicksand: Coming to grips with the preservation of digital information in Nigeria," presented at the 47th National Conference and Annual General Meeting of the Nigerian Library Association, July 26-31, 2009.
- 5) Masakazi, N. (2009), "What is the value of digitizing South African arts, culture and heritage to its citizens?", E-Mzazi Information Society, available at: www.pnc.gov.za/termzanzi-issue.1-07.
- 6) Ezeani, C.N. (2009), "Digitizing projects in developing countries: the case of the University of Nigeria", Library Hi Tech News, Vol. 26 Nos 5/6, pp. 14-15.
- 7) P. De Stefano, "Selection for digital conversion," Moving Theory into Practice. Mountain View, A. Kenney and O. Reiger, Ed., CA: Research Libraries Group, 2000
- 8) R. Tennant, "Selecting collections to digitize," Library Journal, vol. 125, no. 19, pp. 26, 2000.
- 9) S.Vijayarani¹ and Ms. A.Sakila, "Performance Comparison of OCR Tools", International Journal of UbiComp (IJU), Vol.6, No.3, July 2015.

Copyright © 2017, Dr. Manish Kumar Singh. This is an open access refereed article distributed under the creative common attribution license which permits unrestricted use, distribution and reproduction in any medium, provided the original work is properly cited.