Organising digital collections: problems and issues

Dr. H.K. Kaul
Director, Delnet
J.N.U Campus, Nelson Mandela Road, Vasant Kunj
New Delhi - 110070
Telephone : 32471001, 32471002, 32471010
Fax : 91-11-24619325
E-mail: hkkaul@delnet.ren.nic.in

Introduction

The importance of digitization has been growing over the past two decades along with the growth of Internet. Every one who is able to afford or manage the Internet technology has been arranging the hosting of the content on the Web. The 21st century began with a new world view about the relevance of digitization and digital collections. Each individual, institution, and country has been desirous of using digital resources and contributing digital resources to the Web.

The relevance of digital resources grew as digital documents are a dynamic resource which can be used by millions of people around the world simultaneously. One can also improve upon them from time to time as considered necessary. Such content can reach people very fast. Access to digital resources from various search strategies is also possible and that makes them useful and meaningful.

During the recent past the desire to provide knowledge and information in digital form free of charge to improve community facilities and standards of living is becoming a reality. We have many examples of such activities around the world. [1, 2, 3, 4] As a result the relevance of digital resources is increasing among public as it also offers government information, technical information and research based information. It can be given to people free or at cost recovery basis. Thus it is becoming necessary day by day to invest in digital technology especially in order to make knowledge accessible fast to public. As the digital technology is getting applied in converting knowledge and information available in print and MS forms into digital forms the framing of policies and development of infrastructures is also becoming necessary in order to help the present and future societies around the globe to manage and organize digital collections in user-friendly ways.

At present mostly the access to knowledge and information based publications is through catalogues, whether online or printed, national or local, broad subject–based or in-depth metadata based. It is generally not through properly indexed online publications. Thus information and communication technologies have a role to play in the implementation of this technology, the constraints of which exist at different levels. These constraints include the efforts involved in framing of policies, training of staff, training of public, providing infrastructure, content development, dissemination of knowledge, ascertaining of the requirement of the public etc.

Types of Content to be digitised

Any type of document can be digitized. The examples include:

- Printed Documents
- Manuscripts
- Video-recordings,
- Sound-recordings, such as oral history collections and music;
- Traditional heritage, culture, medicines etc.
- Etc.

The catalogues act as useful metadata if they are in conformity with the International standards and are available online. For otherwise new metadata using international standards has to be developed for each document and for each type of document. Each country, each institution and each individual interested in digitizing content will have to prioritize first the type of documents that need to be digitized and then have criteria for the selection of documents for the purposes of digitization.

Criteria for Selection of Content for Scanning

Digitisation

The digitisation of documents is a major activity in libraries and archives around the world. The libraries and archives that contain some of the valuable resources are either beginning to digitise their documents or are trying to find resources to do so. The very activity of digitisation has both local as well as global implications.

Digitisation is a form of printing and making the digital document available through the Web is yet another form of publishing. A reputable publisher interested in the publication of good works makes use of various selection methods before the publisher selects the MS for publication. In digitization also the selection methods fall under three categories:

- Nomination
- Evaluation
- Prioritisation

In the first instance the publisher selects experts who in broad terms:

- name topics on which books need to be published;
- name experts who are already working on such themes; and
- name experts who could be commissioned to write on such themes besides what MSS the publisher receives directly.

In the second phase the publisher uses an extensive evaluation and editorial process. In the third phase the publisher prioritises a MS in relation with the other MSS for the purposes of publication. The same three processes are pursued in a more scientific way in the selection of MSS for digitisation purposes. Well developed selection mechanisms have been evolved for this purpose during the last two decades.

The Basic Questions

Before we look into the actual processes of selection, it would be in the fitness of things to find out:

- The value of the document;
- Public demand for the document;
- Whether already a similar document in the digital form is available;
- Is the document forming part of series of documents?
- Is the document already digitised by another institution/Agency?
- Are we providing a better digitised version with proper indexes;
- Physical fitness of the document; etc.

Selection of Materials for Scanning

The Principles

The purpose of selecting documents for scanning should be based on the following principles [5]:

- Digital document on the Web is a published document;
- Own copyright of the document before digitising it.;
- Arrange financial resources to support preservation of digital databases;
- To reduce costs on scanning select the best technology;
- For each document create a well researched documentation;
- Don't publish sensitive documents on the Web without consulting the concerned officials or organizations;
- Undertake a final overall quality check;
- In addition the Committee of Experts for the selection of materials or its sub-committees should ensure before digitisation begins that:
    - The document conforms to the broader focus of the project;
    - The document is in perfect physical condition for digitisation; if the condition is not good, it is verified that no another copy of the document is available for digitisation purposes anywhere else;
    - The document is not available for general use because it is tiny / oversize; its physical condition is bad; it is housed in a storage vault or it is available on a format like glass or birch bark which is not open for general use.

Basic Selection Methods

The selection of documents should be done in the following three phases:

Nomination
A meeting of experts, authors, library and information scientists, archivist's etc. specialising in a discipline of concern be requested to give names of documents that need to be selected

for digitisation purposes. *Handbook for Digital Projects* presents the guidelines for nominating materials for digitisation of documents. [6]

Evaluation
The Selection Committee or the Sub-Committee in a particular discipline should examine the suggestion made and decide about which items to be included and which to be deleted.

The Selection Committee should evaluate the recommendations for deselection of a document according to international practices set for this purpose, keeping in view the :

- Value
- Use, and
- Risk involved.

Prioritisation
The Columbia University Libraries [7] include many factors such as intellectual content, historic value and physical value in order to ascertain the value of a document for prioritisation purposes. These characteristics are further subdivided covering various attributes including rareness of the work, coverage of the subject area, usefulness and accuracy of the content, the demand from the users, non-availability of a similar work and other value added criteria.

The Handbook for Digital Projects[8] divides the process of finding the value of a work into the following five categories:

- Informational Value
- Administrative Value
- Artifactual Value
- Associatiional Value, and
- Evidential Value

The above principles and practices help in deciding which document should be digitized and which should not be digitized.

Copyright/IPR Issues

Legal Issues Concerning Digitsation

Copyright laws in each country restrict the unfair copying of documents under the Berne Convention or the Universal Copyright Convention etc. [9]. However, the international practices which are also important *Handbook* [10] need to be kept in mind while evaluating a document.

- *Donor Restrictions:* If the donor of the document-to-be-digitized puts substantial or nonnegotiable restrictions which prevent the users to use the document according to the policy defined for the project, then don't digitize the document. If the document is important and no where else available try to re-negotiate the terms with the donor.

- *Copyrights:* Don't digitize any document unless you are sure that it is in the public domain or you have obtained copyrights or licenses/ permissions.

- *Privacy Rights:* If a document contains images/pictures of living persons obtain permissions from them before digitizing the text?

- *Publicity Rights:* If the document includes images or recordings of famous persons such as motion picture or recording stars, scientists, artists or authors obtain permissions from the persons or their estates before digitizing the text.

- *IT Regulations:* Don't digitize the document which is not permitted under the law or the Information Technology Act.

- *Sensitivity:* If the document contains sensitive information on subjects such as defense, religion etc. or is unbalanced in its point of view the Selection Committee should get the advice of experts before taking a decision on digitisation.

- *Evidential Value:* If the document contains material that is evidential in nature or supports events with legal and historical proofs and/or interests a key audience as it has substantial information, then the document should be digitised.

Electronic Legal Deposit

The delivery of books to the State or the legal deposit of printed publications has been in vogue in most of the countries. With the onset of digitization every publisher generally has a digital copy of the MS with him. The law of legal deposit has to take into consideration the deposit of electronic versions of each published work with the Government. [11] If the publisher submitted the final MS copy in digital form to the Government as the MS was getting published this should be termed as an electronic legal deposit. It would reduce the cost on digitization of new publications considerably. And the Government agency, such as the National Archives for Digital Resources (which could be established) would provide content in all such cases where the publisher had given permission to do so free of charge or at a given fee, for dissemination to other national agencies like the National Library. Keeping in tact the interests of the publishers different formulae could be worked out in each case, in each country and at the international level. Public access to digital resources should be made possible, either at no charge, or at a reasonable charge to the users. For example the Council of Australian University Libraries (CAUL) has signed agreements with vendors in order to provide access to licensed content. [12] This way licensed content can become accessible to public.

Institutional Repositories

Each institution, including educational institutions such as universities, research organizations, Government departments are now creating MSS in digital form. If there are national and state level agencies to receive such digital data for archiving and dissemination, the archiving problems at local levels would diminish and centralized archiving facilities using international standards would emerge. However all such institutions which have capacities to develop their own archiving facilities would anyhow maintain institutional repositories for local use.

It is necessary to develop national and state level archiving facilities for digital data which could also guide local institutions, agencies, individuals on software, standards, hardware, communication issues and dissemination of information to public.

Direct access or access through Consortiums

Access to commercial digital content like e-Journals, e-Books etc. can always be made directly through vendors if one can afford to pay for it, for otherwise consortium approach has been tested in many countries and it helps in getting access to commercially available content at a low price. The issue of archiving licensed content for which one has paid or the issue of archiving content for which the permissions are not available are the issues that also need to be resolved. The pricing models should be worked out in such a way so that each institution has access to licensed content once the subscriptions expire and the institutions do not have the finances to continue the subscriptions.

Development of digital collection

The development of digital collections can take various routes. Various models can be adopted. [Graphical model - 13, Open access model -14; Business model - 15] etc. We find texts that are keyed in in highly structured form and there are page images where in metadata is added. There are digital collections that are born as digital collections and these include nowadays newspapers, e-journals, documents in offices, MSS, publications, e-mails etc And also, there are digital documents which are converted from print to digital versions. Such digital documents are getting developed regularly in libraries and under other national and international programmes.

Born Digital Publications

Digital resources that are born in digital form initially could be obtained from the owners either as gift, on payment or under the provisions of a legal deposit. [16] Even if the provisions of the legal deposit are not finalized any agency can ask authors of digital works to present/donate them to the library or give them on long term lease. The general public could be intimated about the availability of such a facility on specific terms. This facility can make the content available to public for a wider use.

Licensed Content Subscribed Through Vendors

It is important to simplify licensing arrangements so that content becomes available to researchers/public under simple conditions. It is found always advisable if each library buys core journals and the remaining are shared digitally/ physically under arrangements with vendors so that users get direct access to the content. The consortia functioning around the world have been negotiating the terms for such a use.

Harvesting/ Downloaded Web Resources

There are harvesters that collect metadata from digital repositories and forward the same to those who maintain them. DRTC in Bangalore offers SDL which is an OAI compliant search facility among digital libraries. Many of the open access digital repositories are not OAI compliant. [17] Another case study is of Prototyping Digital Libraries Handling Heterogeneous Data Sources – The ETANA-DL Case Study in which the harvester works at given periods of time regularly and collects new data. [18] The National Library of Australia presents statistical evidence on increasing web usage by using Open Archives Initiative Protocol for metadata harvesting. [19]

Collections Created by Digitisation Processes

A number of digital library software has become available for digitizing publications in the libraries and /or for other purposes. These are either available as open source or as commercial products. The following are a few examples:

DSpace
DSpace open source digital library software is jointly developed by MIT Libraries and HP Labs and it is used effectively for building digital collections with uploading facility through the Web. It gives persistent identifiers to each digital collection and conforms to OAI-MPH v.2.0 and Dublin Core standards. As it conforms to UNICODE standard it can be used for multilingual digital collections also. [20]

Greenstone
Another open source software for building digital libraries is Greenstone. It is also used for developing and distributing digital library collections. Digital collections developed on the software are published on the Internet or on CD. This software is developed at the University of Waikato, in New Zealand. UNESCO is also supporting the project. [21]

Management Software for Digital Libraries
This is a scientific literature digital library software which is being used among other libraries by the University of Singapore. [22]

There are a number of commercial products available globally and I have Not included them in this paper.

Digital Preservation and Archiving

The quality of digital preservation will always be defined by the quality of the software that is used for preservation purposes. However, such software are evaluated keeping in view at least the minimum guidelines that are given below: [23]

- open-source software,
- data documentation,
- rights management,
- archival authenticity,
- controlled access, and an
- Institutional commitment to preservation.

Metadata

The availability of appropriate metadata with a digitized document makes access to the document user-friendly. Also, proper access to the document enhances the utility of the document. But in order to develop appropriate metadata one should take into account:

- Terminology which should be globally comprehensible and standard; and
- How access to content becomes possible through a variety of search engines.

Metadata would generally include author, file name, creation software and version, creation date, modification date, subject, size, and any additional pertinent information. Once the preservation process is on the addition of metadata becomes necessary. Various methods are being used but the most common is the **Dublin Core Metadata Initiative** which gives guidelines for creating interoperable online metadata standards. [24]

Replication and Reproduction of Data

In the preservation of data the processes should take into account the following issues:

- Reproduction of data should be faithful in terms of quality, long-term access, interoperability, including its appearance in comparison to the original page, etc. There should also be version identification possible for each version of a document;
- The first digital copy of a document, even though in a poor form should be maintained for future, reference, or recovery of data.
- It has been noticed that as soon as the digitization is completed the original physical document is destroyed in the name of saving space. This should not be normally done.
- Migration of files from the existing hardware and software to the
- New hardware and software should be possible. Migration is important and it should be achieved from one system to another. To retain the data only in CD form alone may not be the right thing to do as the life of CD's is limited.

- Preservation of data has to be done in a standard format. If the format is not standard the migration my not be possible. The standards such as Tag Image File Formats (TIFF) or Portable Network Graphics (PNG) files are recommended.

The system should allow data to be preserved without compressing or encoding. Compressing and encoding could act as barriers in case an error gets generated in the system. As a result the archivists also would not be able to read the data.

Preservation process is an ongoing work which has to be undertaken with great precaution. Non-experts should not be allowed to handle its metadata recording, file conversions, file migrations etc. Digital preservation issues such as inter-operability among digital archives, and harmonization of metadata are still important and research is going on these issues.

The archiving of digital data for a long term demands that there is an ongoing maintenance undertaken for the byte stream and that access to content remains possible irrespective of the changes taking place in the technologies of digitization and preservation. In order to achieve this, research into the archiving of scanned digital images and born digital images is still on. The long-term retention of digital data is an open issue for research and so is the identification of optimum metadata schema for complex image files. The digitization of pictures, images, music and maps do take a much larger digital space. So does the archiving of Web resources. The digital archiving needs more research so that it becomes robust in its character. Therefore all those who are contributing data for archiving purposes to a central agency should define their responsibilities and goals clearly. Also, the legal issues concerning archiving of data from various sources need to be looked into well in advance.

Distributed Archiving Records Management

Records management is very important when digital records get distributed. Stocktaking of records/ files to enhance record keeping practices is undertaken from time to time. Software selected for the purpose should undertake these jobs.

The database platform of Records Management's recordkeeping software, TRIM allows for reporting, and offers increased functionality. [25]

Accessing Digital Collections

Users need to access information from a variety of resources through a single search. Therefore search mechanisms should be such that several related databases get searched online. There should be institutional arrangements finalized before technical solutions are applied. A distributed search model should be so effective that it is accessible by a non-expert easily.

Accessing Digital Collections in Other Languages

The access facilities need to be such that can allow access to non-English digital collections and arrange mechanical translations as far as possible. This issue is equally important.

One of the issues that concerns librarians and scholars is the changes that take place in the citations to the digital resources. There is a need to have "persistent identifiers" [26] so that irrespective of the change in the locations identifiers can give information about digital resources referred to. It goes without saying that ongoing access to digital resources involves more expenses. This is why archiving of digital resources needs to be taken up at national and state levels on a priority basis.

Skills of Library Staff

One of the issues in organizing digital collections is the lack of proper trained manpower. The library staff needs to be equipped with new skills for handling new technology and accessing knowledge from various digital collections. These skills include training in 'metadata creation', 'cataloguing of digital collections', 'digital preservation and archiving', 'subject requirements of individual scholars', 'standards for digitisation', 'digital scanning', 'use of multimedia', etc. Being a specialized job younger staff generally acquires new skills fast and whoever is ready should be encouraged to adopt new skills.

There is also the need to train the users, wherever possible, in accessing digital resources online. This needs to be done at various levels and could be done in collaboration with the help of institutions and NGO's that are involved in information and knowledge dissemination.

References

1 http://www.ourplanet.com/imgversn/121/swamin.html
2 http://www.slq.qld.gov.au/about/visitors/ikc?SQ_DESIGN_NAME=text_only&SQ_ACTION=set_design_name
3 http://72.14.235.104/search?q=cache:VU1_1QTjn4kJ:www.public-libraries.net/html/x_media/pdf/poustie_engl.pdf+%22Knowledge+Centre%22+AND+%22for+public%22&hl=en&gl=in&ct=clnk&cd=8
4 http://www.gosnells.wa.gov.au/scripts/viewarticle.asp?NID=9413
5 Sitts, Maxime, ed. Handbook for Digital Projects. Andover, Mass.: Northeast Document Conservation Centre, 2000. Pp. 36-38.
6 Ibid. pp. 51-54
7 http://www.library.cornell.edu/colldev/digitalselection.html
8 Op. cit. Sitts, pp. 44-45.
9 http://www.unesco.org/culture/copyright/html_eng/convention.shtml
10 Op. cit. Sitts, pp. 55-59
11 http://www.natlib.govt.nz/en/services/5legaldeposit.html
12 www.**caul**.edu.au
13 http://www.rlg.org/en/page.php?Page_ID=20952&Printable=1&Article_ID=1835
14 www.valaconf.org/vala**2006**/papers**2006**/58_Coleman_Final.pdf
15 www.surf.nl/download/DRIVER_final_Annex1_v9.pdf
16 www.ala.org/ala/GODORT/communications/letters2004/CLRfinal.htm
17 http://www.fz-juelich.de/zb/datapool/page/768/Prasad_Abstract.pdf
18 http://feathers.dlib.vt.edu/~etana/Publications/ECDL2004Etana-DL.pdf
19 http://www.nla.gov.au/nla/staffpaper/2005/boston2.html

20 http://www.hpl.hp.com/news/2003/july_sept/dspace.html
21 http://www.greenstone.org/cgi-bin/library
22 citeseer.ist.psu.edu/672583.html
23 bancroft.berkeley.edu/info/guidelines.html
24 http://dublincore.org/documents/dcmi-terms/
25 http://www.infonet.unsw.edu.au/ras/trim.htm
26 http://www.nla.gov.au/initiatives/persistence.html