United States Educational Foundation in India,
DRTC/Indian Statistical Institute,
DLIS/University of Mysore
Joint Workshop on Digital Libraries
12th – 16th March, 2001

Paper: C

Digital Libraries: Selection of Materials for Scanning

H. K. Kaul

Director, DELNET-Developing Library Network,

New Delhi

Email: *hkkaul@delnet.ren.nic.in*

1 Introduction

Digitisation of documents is becoming a major activity in libraries and archives in the world. The libraries and archives that contain some of the valuable resources are either beginning to digitise their documents or trying to find resources to do so. The very activity of digitisation makes the job an activity that 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 broad terms the selection methods fall under three categories:

a. Nomination
b. Evaluation
c. Prioritisation

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

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

In the second phase when the publisher uses an extensive evaluation and editorial processes. 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.

2 The Basic Questions

Before we look into the actual processes of selection, it would be in the fitness of things to find what are the basic questions that are asked before the selection process actually begins. [1] They are:

2.1 What is the value of the document that is being selected for digitisation purposes?

Here the issues like the rareness or uniqueness of the document, relevance of the subject in the present context, the extent of coverage, usefulness and accuracy of content, usefulness of the text in the digital form etc. form part of the scope.

2.2 Will there be sufficient demand if the document is published in the digital form?

Here issues like the relevance of the document for current studies and interests to an institution or general public, demand from new users and global support or institutional support form part of the scope.

2.3 Is there another document in digital form on the same subject and covering the same scope?

2.4 Is the document forming part of series of documents? If so are other volumes being digitised by the institutions housing or owning them? If yes, is digitisation considered necessary to complete the document in digital form?

2.5 Is digitisation of the document considered to be part of a collaborative programme for the digitisation of relevant documents in a given subject?

2.6 Is the digitisation of a document meant to enhance its value through links to bibliographic records, provide better indexing facilities, provide hyperlinks or to disseminate further the content of the document?

2.7 Is the document in the public domain. If so, is it relevant today and will its presentation in the digital form increase its utility?

2.8 Is the document fragile and is it considered necessary to digitise it for preservation purposes?

3 Selection of Materials for Scanning

3.1 The Principles

The purpose of selecting documents for scanning should be based on the following principles [2]:

3.1.1 Digital Document on the Web is a Published Document

A document that will be digitised and made available to the users in the world will be open for use and comment by any scholar, professional and the general public. It is important to see that whatever has been digitised is a unique publication and it carries with it all the characteristics of a published document. This would imply that the publishing body digitising the document under the project should be able to give permission for copying, further publication, increased access etc.

3.1.2 Own Copyright of the Document before Digitising it

In order to execute the right of a publisher without violating copyright regulations, it is important that the publisher has licence to digitise the document or owns copyright of the document. For non-copyrighted materials it would be better to see that it has not been digitised before by any other Agency. If so, it may be necessary to discuss if access to the already digitised document would be cheaper than digitising the document again.

3.1.3 Arrange Financial Resources to Support Preservation of Digital Databases

Digital databases need to be accommodated in new formats and hardware using compatible software from time to time. The continuing financial support will have to be there to support the change over so that the users continue to have access to the document. It is therefore important that the project should have the means of revenue generation or the facility of getting a regular grant for the maintenance of the digital data.

3.1.4 To Reduce Costs on Scanning Select the Best

The initial costs on scanning may not be much but coupled with quality control checks, preparation of indexes, catalogues and metadata the whole job becomes expensive. It is therefore advisable to select the best documents at the initial stage.

3.1.5 For Each Document Create a Well Researched Documentation

Before scanning is started each document, part by part, should be processed for appropriate captions and completeness of the document. Accompanying material need to be written to give the necessary context to the document being digitised.

3.1.6 Don't Publish Sensitive Documents on the Web without Consulting the Concerned Officials or Organisations

It should be kept in mind that any document that is of sensitive nature for a group, society or country should not be digitised unless it is considered tht the propagation of such an information is part of the objectives of the project.

3.1.7 Undertake a Final Overall Quality Check

The Committee for Selecting Documents for Scanning should undertake final quality check in terms of the authenticity and accuracy of information both for textual or visual materials and the overall presentation of the digital data on the Web.

3.1.8 In addition, the following apply

The Committee of Experts for the Selection of Materials or its sub-committees will ensure before digitisation begins that:

a. the document conforms to the broader focus of the project;

b. 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;

c. 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.

3.2 Basic Selection Methods

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

3.2.1 Nomination

A meeting of experts, authors, library and information scientists, archivists etc. should be held in the discipline of concern to collect names of documents that need to be selected for digitisation purposes;

Handbook for Digital Projects [2] presents the following guidelines for nominating materials for digitisation of documents and the forms to be used for nomination purposes:

Guidelines:

a. "How much of the collection is well and accurately documented at the item level in reliable and complete indices and finding aids, and where are these well-documented items?

b. How much of the collection is in stable or good condition, and where are these stable materials;

c. What portion of the collection is standard and of consistently normal in size, with black-and-white and/or printed materials, and where do these materials fall? Note: Avoid oversized, unusual and varying format, long-tonal range, colour, and handwritten materials for new projects;

d. What materials are easy to provide to researchers because of their size, format, or viewing requirements and where are they available?

e. What percentage of the materials does the institution have the copyrights to or licenses for, and where are the public domain materials?

f. What percentage of the materials has no restrictions or sensitivities of any sort (such as privacy, publicity, defamation, obscenity and sensitivity, or donor restrictions), and where is this restricted and nonsensitive material?

g. What materials are of highest monetary value and well secured, and where are they in the collections?

h. What materials are judged to be at highest risk and why, and where are they located? Of these, which are stable enough to be scanned without damage or which have already been well photographed?

i. What materials are used most frequently, how are they used, and where are they located?

j. What materials are unique to the institution, and where are they located?"

In order to collect a variety of view points as given above, the use of the two forms presented in the *Handbook* [3] should be used. They are used for the selection and deselection of documents. See *Appendix I.*

3.2.2 Evaluation

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

The Selection Committee will evaluate the recommendations made in Form A (*Appendix I*) for selection of a document and Form B (*Appendix II*) for deselection of a document according to international practices set for this purpose. The following principles which are based on the recommendations given in the *Handbook* [4] need to be kept in mind while evaluating a document.

a. *Mission Statement:* Is the document to be digitized falling within the purview of the Project? If not, don't digitize.

b. *Scope of Collections Statement:* If a complete collection is digitized, confirm that the document to be digitized falls within the repository scope of the collection.

c. *Deselection Requests from the Supporters of the Project:* If the supporter of the Project recommends that the document should be digitized and if this is challenged by equally important sources , the document should not be digitized. However, if the Selection Committee finds deselection recommendations frivolous or insubstantial, then ignore them.

d. *Donor Restrictions:* If the donor of the document-to-be-digitized puts substantial or un-negotiable 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.

e. *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.

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

g. *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.

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

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

j. *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.

k. *Authenticity:* If the document is authentic and original in contribution it should be digitised.

l. *Visual Accuracy:* If the print/appearance of the document supports the creation of an accurate and sharp digital version then digitise it. If not, find alternate methods for doing so.

m. *Documentation:* If the document does not have appropriate captions and the budget does not permit to appoint staff to create them, then defer the digitisation of that particular document.

n. *Contextualisation:* If a document essentially needs substantial and expensive research inputs in terms of contextual support such as hypertext support for certain portions or viewing of document in relation with other documents simultaneously etc. it may be necessary to reconstruct the archaeological support in the Encoded Archival Description (EAD) format or another suitable format. If it is not possible to do so, it would not be advisable to digitise such a document.

o. *Added Value:* If the document has become available for the first time, also if it fulfils the necessary conditions laid out for the selection of a document and is considered necessary to make it available to a larger audience, then if funds permit , digitise the document in order to:

  i. make the unique document available to a larger audience;
  ii. create linkages to the document through HTML, SGML, XML coding;
  iii. make it part of the virtual collections on the same subject using different techniques, format and bringing together physically separated doocuments either on the Web or in CD form.
  iv. add new indexes and searching aids.

p. *Audience:* If the digital version and the printed version reaches the same audience, yet considering that the document is important, the digitisation of the document should be considered.

q. *Sipplementary Selection Criteria:* If the audience creates its own selection criteria, such recommendations should be taken into the evaluation process;

r. *Technology:* If the audience can not afford the expensive equipment for using the digital version, then avoid digitisation. However, now that Internet is becoming available to more and more users in India, this factor does not apply. However, the technology should be such that every Internet user can access the document easily.

s. *Condition:* If the condition of the document to be digitised is very bad and it is likely that in the digitisation process the document will get damaged, then do not digitise it.

t. *Control:* Make sure that rare materials are kept under security during the process of digitisation and are returned to the owner in the original condition.

u. *Duplication of Effort:* If the document has been digitised elsewhere, locate the source, and find the quality of the digital version. If we can get a copy for general use, then it is not worth digitising the document.

v. *Accessibility:* If the original document is inaccessible but it is available in microfilm or microfische form widely, it may not be ideal to digitise the document at that stage.

w. *Cumulation:* If the document is relevant as part of collection only, then the digitisation of it alone must be seriously questioned by the Selection Committee and other reasons obtained for its selection before taking a decision for digitisation of the document.

3.2.3 Prioritisation

The final list prepared for digitisation purposes will be ranked in the order of priority keeping in mind the relevance of each document in a historical perspective, its use in the present context and its physical condition. However, the processes of prioritisation involve the following three major criteria [5] to be taken into consideration:

a. Value
b. Use, and
c. Risk

3.2.3.1 Value

The Columbia University Libraries 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 added value criteria. The selection criteria for digital libraries as prescribed by Columbia University are given in *Appendix III.*

The Handbook for Digital Projects [6] divides the process of finding the value of a work into the following five categories:

.1 Informational Value
.2 Administrative Value
.3 Artifactual Value
.4 Associational Value, and
.5 Evidential Value

3.2.3.1.1 Information Value

The documents of topical content on issues like events, places, important people and projects are listed in this category.

3.2.3.1.2 Administrative Value

The documents that are considered useful in the functioning of an organisation form part of this category.

3.2.3.1.3 Artifactual Value

The primary documents of archival nature such as diaries, photos etc. are listed here.

3.2.3.1.4 Associational Value

The primary documents like correspondence, papers and related documents that have a bearing upon an important monument, city, country, person or a subject are listed in this category.

3.2.3.1.5 Evidential Value

The documents that serve as legal or historical proof of an event, activity or occupation are categorised here.

In all the above categories, the documents could be grouped under 'High Value', 'Moderate Value' and 'Low Value' documents in order to score the value for prioritisation purposes.

3.2.3.2 Risk

During the selection process the documents that are having a risk factor inherent in them due to legal and social implications, such documents are not selected. However, documents that are undergoing a chemical process due to age factors and are decaying as a result, such as negatives of films, or paper infected with insects and mold, damaged CD,s discs etc need to be classified in 'High Risk', 'Moderate Risk' and 'Low Risk' categories. The digitisation needs to be done to minimise the risk factors involved.

3.2.3.3 Use

Documents that are on demand (re-demand) regularly for reference purposes are of high value. Surveys need to be done on the basis of the present practice and expected use to prioritise a document on the basis of its use.

All the factors related with value, risk and use need to be considered before prioritising a document. In each category of value, risk and use can be sub-divided into three sub-categories with numerical values given to them such as high=6, moderate=3 and low=1, one can evaluate documents numerically and prioritise them.

4 Conclusion

In conclusion let me stress that the selection process of materials for scanning need to be meticulously undertaken keeping in mind the basic questions, the principles for selection and the guidelines for nomination, evaluation and prioritisation. If this exercise is undertaken in case of each document, we are sure to select the right and appropriate document for digitisation purposes.

5 References

1. AYRIS (Paul). Guidance for selecting materials for digitisation. Joint RLG and NPO Preservation Conference Guidelines for Digital Imaging. Pp.7-8.
2. SITTS (Maxime) [ed.]. Handbook for Digital Projects. Andover, Mass.: Northeast Document Conservation Centre, 2000. Pp. 36-38.
3. Ibid. pp. 51-54.
4. Ibid. pp. 55-59.
5. Ibid pp. 44-47.
6. Ibid. pp. 44-45.

Appendix I

Form A : Nomination Form for Selection

X Name of the Institution where the documents are located.

1. Materials Being Nominated for Digitization (Please indicate collection number, series, number, box number, folder number, item control number or equivalent and the creator; caption of the item or a bibliographic citation to the fullest extent possible.)

2. Reasons for Nomination (Describe why the fmaterials are important, who might want to use them in a digital form, and what usages are likely if they are digitized.)

3. Potential Assistance Sources (Please indicate if you have any special knowledge or skills that might be shared with the X repository during the selection process. For example, can you provide caption information, historical background, or are you aware of potential funding sources or digital projects that are covering similar materials to those you are nominating?)

4. Restrictions (Indicate if you are aware of any reason why the specified materials should not be digitized, such as legal, ethical, or cultural sensitivities. Please be as specific as possible citing a source, such as a law or culture group and a contact name if necessary.)

5. Your Name
6. Your Address
7. Tel. Fax:
8. E-Mail

Note: The Selection Committee will make all final decisions on what will or will not be included in the digital project. If you have any special information you would like to share with the committee, please write it below.

Appendix II

Form B: Nomination Form for Deselection

Name of the Institution

1. Identify the Materials That Shouldn't be Digitized (Please indicate collection number, series, number, box number, folder number, item control number or equivalent, and the creator, or caption, of the item to the fullest extent possible.)

2. Reason for Deselection (Describe why the materials shouldn't be digitized or shared electronically. Identify problems or concerns hat would arise, including legal, cultural, social or ethical concerns. Identify who might be affected if the materials are available electronically.)

3. Specific Restrictions (Indicate if you are aware of any reason why the fmatgerials should not be digitized by citing specific laws, policies, or equivalent documentation. Please be as specific as possible citing a source, such as a law or culture group, and a contact name if necessary.)

4. Your Name
5. Your Address
6. Tel. Fax:
7. E-Mail:

Note: The Selection Committee will make all final decisions on what will or will not be included in the digital project. If you have any special information you would like to share with the committee, please write it below.

Appendix III

Selection Methods used by Columbia University

A. **Purpose and Program Description**

The Libraries digital information collection development objective is to select those materials, which will support the current curricular and research needs of the Columbia community.

Members of the Columbia Community access digital information using personal computers in their offices, dorms, homes, and by public terminals located in libraries and other public buildings on campus. The sophistication and knowledge of the Libraries' digital information users varies significantly. The Libraries have many patrons whose only contact with computers has been word processing. Others routinely locate and exchange information with colleagues the world over using the Internet. The digital Library research skills of most of Columbia's undergraduates, graduate students, faculty, and post-doctoral researchers lie somewhere along the continuum between these two extremes. The breadth and depth of information sources needed by them also differ. Consequently, the Libraries provide access to a wide variety of materials using an equally wide variety of access mediums to meet patrons needs.

B. **General Selection Guidelines (See classed analysis for further details)**

Overall, the Libraries collect at the research level.

C. **Specific Delimitations**

1. *Access mediums employed:* The Libraries employ remote Web access to digital sources of information on an extensive basis and networked CD-ROM and single CD-ROM workstations selectively. The Libraries favor Internet and WWW products because they offer ease of use, wider access, more rapid updating, and the cost savings over local maintenance and storage. It is in the process of migrating away from all Telnet applications in favor of Web access.

2. *Archiving:* The Libraries share with other research and educational institutions the responsibility to determine the most effective methods for the long-term preservation of the digital materials accessed by Columbia but not stored locally in its collection. It has a special preservation responsibility for digital materials unique to Columbia. Items for which the specific archiving responsibility has not been established may be purchased. In the future, however, the lack of a fixed archival responsibility may become increasingly important selection criteria.

3. *Consortial purchasing:* The Libraries participate in the Northeast Research Libraries Consortium and the New York Consortia of Consortia in order to take advantage of aggregated purchasing agreements. It seeks consortial licensing opportunities whenever possible.

4. *Coordination and promotion:* Networked Resources Coordinators (NRC's) are assigned for each networked title. NRC's are responsible for promoting the use of each new tool, of overseeing librarian and user training, of communicating to members of the Columbia community the strengths and limitations of the digital collections, and of working with vendors and publishers when problems cannot be easily resolved by Library Systems Office personnel.

5. *Digital conversion:* As funds are made available, new technologies absorbed, and current copyright barriers overcome, materials previously purchased in print form will selectively be converted to digital form.

6. *Document types collected:* Reference tools, e.g., indexes, abstracts, directories news services, etc., and full-text e-journals are collected very extensively; relevant Web pages extensively; Zines, machine readable data files, and digital monographs selectively; and multimedia courseware are collected very selectively.

7. *Duplication:* The Libraries generally purchase duplicate copies of the same content for only high-use titles. However, until more is known about the level of patron acceptance of digital sources of information, the Libraries will acquire both print and digital versions of the same material. The intent, however, is to gravitate toward the digital format a soon as possible.

8. *Electronic journal aggregation preferences:* The Libraries prefer a subject interface that permits the use of both a controlled vocabulary and key words to search the full-text contents of electronic journals without giving up the ability to browse the contents of individual issues. It favors the following models of access; (In order of preference)

   a. Web-based indexing and abstracting services linked to full-text electronic journals, e.g., ISI's ability in the near future to link between SCI/SSCI/AHCI and the contents of full-text journals.

   b. Aggregated databases composed of a variety of media, e.g., newspapers, newsletters, government reports, monographs, reference works, etc., e.g., Academic Universe, ProQuest Direct.

c. Aggregated collections of journals published by a variety of publishers but sharing a common searching a common searching interface, e.g., J-Stor.

d. Aggregated collections of journals published by a single publisher but sharing a common searching interface, e.g., Ideal, Project Muse, etc.

e. If none of these aggregated forms of access are possible, individual titles will nonetheless be collected.

9. *Electronic selection criteria:* Like print materials, digital titles added to the collection need to match the needs of our clientele; be of appropriate scope, content, depth, and quality; be affordable; the content must be timely; be bibliographically accessible; and in the appropriate language, etc. It is also presumed that there are no technical reasons why the Libraries cannot provide access, e.g., doesn't use a proprietary browser, permits printing, etc., and that their use by library patrons and librarians will not require an inordinate amount of training.

10. *Funds used:* Regular library materials funds are used to purchase digital forms of information with separate budgets for networked digital resources. Aggregated packages of electronic journals are also purchased with these separate digital resource funds (Humanities, Science, Social Science, and General, Interdisciplinary and Undergraduate). CD-Roms used on individual workstations are normally purchased with individual monographic or periodicals funds. Individual electronic journal subscriptions are also purchased with individual subject periodical funds.

11. *Language and place of publication:* The Libraries collect largely English-language commercial/electronic forms of information but which are produced worldwide. Non-commercial Web page links are collected without regard to language or origin.

12. *Licensing:* Licenses should provide the Libraries with permanent rights to the content that has been paid for; should not require the Libraries to police the use of or hold it liable for the use of the information; should require only "reasonable effort" on the part of the Libraries to address misuses by Libraries patrons when discovered by the publisher or vendor; allow use by all of Columbia's faculty, staff, and students as well as casual walk-in patrons; should permit "fair use" of the information, understood to mean to include the same sorts of curricular

and research purposes that have been pursued with print materials; should allow the Libraries to enhance the use of the data to make it more visible or convenient as needed, e.g., on-line reserves; to respect the confidentiality of information about individual users and their use of the information; should protect the Libraries right to the information as advertised by the vendor or publisher; should allow for reciprocal rights to terminate the license agreement; should clearly identify what information is confidential; and to provide use data to facilitate internal needs and service analysis.

13. *Multiple electronic formats:* When more than one digital format is available for the same title, decisions about which to acquire are based upon the alternative costs for each medium and the breadth of access needed. In general, the larger the user group needing access, the more likely the Libraries would prefer that the title be available on ClioPlus or LWeb so that there is broad access across campus and from dorm and home computers. The smaller the user group, the more likely that the Libraries would prefer to provide access from a CD-ROM LAN or single workstation. If the cost of each alternative is the same, the Libraries prefer ClioPlus or LWeb access.

14. *Non-commercial forms of information:* In addition to the purchase of digital forms of information, because of the nature of the Internet, the libraries actively acquire "free" forms of information to a much higher degree than in the past. These are selected and added to the individual library and subject home pages maintained by the library materials selectors and reference librarians.

15. *Security:* Columbia currently utilizes Kerberos, Cheesewiz, and WebScript for campus authorization and authentication and vendor-based IP authentication to prevent the unauthorized use of services licensed by Columbia. The Libraries is also developing a proxy server approach to allow off-campus patrons, employing a commercial Internet service provider, to dial into the campus network using a modem. The Libraries is encouraging vendors to use a WebScript-like login system to facilitate the use of digital resources from off campus.

16. *Selection responsibilities:* Digital library materials will be selected by each of the regular subject specialist library materials selectors. These selectors have the responsibility of assessing the needs of the users in the subjects assigned to them, monitoring the scope of commercial and non-commercial materials that become available, selecting and ordering employing appropriate channels needed materials, and seeking out opportunities to cooperate with other librarians to better meet the needs of Columbia's faculty, students and staff.

17. *Technical considerations:* Selectors need to be mindful of a variety of technical considerations when evaluating new digital tools: Technical compatibility with existing hardware, software including Internet browsers, the availability of technical support, response time and reliability of telecommunications, servers, etc.; and the significant staff and digital storage costs associated with local tape.