1

Building Digital Collections: An Evaluation

M. P. Tapaswi*

Abstract The difference between 'digital collection' and 'digital library' has been explained. The evaluation parameters for digital collections are described. These are based on the basic principles of the digital collection. The descriptions are backed with experiences.

1 Introduction: Digital Library versus Digital Collection

Twenty-five years ago it was hard to believe that the desired information in the form of documents could be available to the user in the shortest possible time over networks. Developments in information and communication technology (ICT) have changed the life of citizens. Interestingly, these developments have not only benefited the developed world but also the developing countries - Friedman describes this and other related concepts with a phrase: "The world is flat". Today, a very large number of full-text documents are available for use over Internet - many free of cost. Many are being added every day. The scope of the content has even gone beyond text to images, sounds, etc. Individuals, organisations, user communities and/or any group of people come together to make such collections available for use. These processes have also given birth to new terminology. Some of these overlap in their concept/ meaning and hence at times are used synonymously. It is therefore

* Documentation Officer. National Institute of Oceanography, Goa

essential that the scope of these be understood to start with. In the present context, 'digital collection' and 'digital libraries' sound synonymous to many. But, they are not.

Let me first explain the difference. A digital collection is (a) an organised assembly of digital information objects in different formats like text documents, still or moving images, data, voice, etc., or a combination of all these and (b) systematically described in a structured format (metadata) using standard tools such as thesauri, controlled vocabularies (subject headings), taxonomies, etc. Some of the definitions of digital collections further extend this description of objects to the overall collection too as a higher level grouping in making the user aware of what it is about at a broader level. To understand them easily these groupings in some related software are described with terms like 'collections' and 'communities'.

So what is digital library' then? - Certainly not digital collection alone. Digital collections are, of course, the major components of the digital library. But thev are only means. There are value additions to the collection. The digital libraries add usability features to the digital collection like user-friendly search and browse interfaces; navigational tools; preview, print and download paraphernalia to make the user's life more comfortable. The human interface between the user and the collection as a system improvement and corrective measure also form part of the digital library. In short, these aspects are quite similar to the 'print collections' and conventional libraries where the collection brings together a set of books (as print documents) duly organised and catalogued using standard methods of cataloguing. The libraries then provide facilities for their borrowing or reference wherein the library staff act as an interface between the collection and users. This concept of digital libraries has been well described by many. [2,3,4]

The evaluation parameters for the digital collections could very well be set based on the NISO's framework that provide comprehensive principles.[5] I have used these in the following paragraphs with my input? to the same.

2 Collection Development Policy

Large collections available on Internet are a part of certain organisations. These collections normally get more focused on a particular topic/ aspect/ subject and match the goals of the organisation itself. There

are. of course, certain collections as 'public places' and their scope at times is unlimited. A policy on the content of a particular collection therefore becomes an evaluation parameter in assessing a collection.

2.1 Collections on What?

The advancements do provide very large scope for the development of the collections on anything that you think of. The scope therefore is needed to be defined at the initial stage itself by applying various parameters like items limited to an organisation, to a subject, type of material, etc. Let me describe these in the following sections.

2.1.1 *Collections Limited to the Organisation's Documents*

Usually, the collections limit their scope to the publications of the organisations which thev are attached to. It becomes easier then to define the scope because they then match with the goals of the organisation itself.

2.1.2 *Collections Limited to a Specific Subject/Discipline*

Some digital collections limit themselves to a particular subject. ar.Xiv (http://arxiv.org/) is one such example. It is an E-print service in the fields of physics, mathematics, non-linear science, computer science, quantitative biology, quantitative finance and statistics. Though it is owned and operated by Cornell University, it permits others to deposit their items and has become a standard source of information for researchers in physics. In the field of ocean sciences there are a couple of collections not limited to any organisation but for the discipline - oceanography. The Intergovernmental Oceanographic Commission of Unesco provides a platform (http://iodeweb1.vliz.be/odin/) for literature on the ocean sciences. Another service in the field of ocean sciences (http:// aquacomm.fcla.edu/) has similar objectives.

2.1.3 Collections by Type

Certain collections confine their scope to non-text materials. There are collections that archive photographs alone (e.g. NOAA's photo archive: http://www.photolib.noaa.gov/). Washington State Historical Society (http://digitum.washingtonhistory.org/index.php) provides a wide range of objects in their collection, e.g. photographs, artifacts, art, posters, music, maps, manuscripts, etc. The Internet Archive. (http://

www.archive.org/) is building a digital library of Internet sites and other cultural artifacts in digital form. Like a paper library, these collections provide free access to researchers, historians, scholars, and also to the general public to a wide ranging audio and video collection apart from print literature.

2.2 Periodicity

As a collection development policy, it is essential to decide how old the objects in the collection can be. At times, the new objects are added to the collection but the old items of archival significance are not considered. The importance of time depends on the subject the collection is dealing with. If it deals with say electronics, not many old items are of great value to the users. However, if it pertains to the history of science, the old items are of great value and the current items that get added would certainly increase their value over a passage of time.

While evaluating any digital collection, it is therefore essential to check the scope of the collection and comprehensiveness depending on the policy it defines.

2.3 Digitisation Process

Another important parameter for the evaluation of the collection is the process that is being used for digitisation. This is mostly applicable when the collectors decide to digitise the documents usually available in print and published without using computers. The word 'digital' is synonymous with the 'computer readable' format.[6] Thus the evaluation can be based on two parameters: the adherence to the digitization policy and the quality of the digitisation.

2.3.1 *Digitisation Policy*

The organisers of the collection should decide whether they would like to launch a programme by which a set of identified documents are digitised in a given period of time or the process of digitisation be demand-based. If the volume of digitisation is known the outcome of this activity can be evaluated on quality and quantity of this process. In the second case, the evaluation can be based on the time required to meet the demand. At times, the process is mixed. In the vear 2004. at the National Institute of Oceanography in Goa, it was decided that the 'reprints' of the contributions of the institute would be provided to the requester in E-

form by digitising them as and when the request is received. More than 60% of the objects of the collection (published literature) was digitised bv this approach and archived. The rest of the collection was digitised within a time-frame as a one task job.

2.3.2 Digitisation of Documents

Digitisation of documents is a two-stage process: Scanning and OCR (Optical Character Recognition). The process of scanning usually ends with the conversion of the print literature into 'computer readable images', The format in which these are to be stored, the resolution, the mode in which scanning is done, etc., are aspects of policy decisions and therefore evaluation parameters for a collection.

The computer readable images once available as a part of scanning processes can further be subjected to another process called OCR which converts computer readable text to 'computer editable text'. A lot of software is available for OCRing the documents in Roman script. Sporadic efforts seem to be in progress for Indian scripts (http://www.mail-archive.com/accessindia@accessindia.org.in/msg17879.html) too. Once the OCRing is done, the computer editable text page can be saved as file. This is a comparatively smaller size than the image file and also clean in look but may contain unrecognised / wrongly recognised characters - needing time- consuming manual check and cleaning. The process may also lose the originality of the document in its font, setting, etc. Of course, some software permit to maintain the originality of the document and the recognised text is kept in the back-end for indexing and searching desired information, etc. '

2.3.3 Formats

As indicated above, the ICT has provided verv large scope for the digitisation of non-documentary sources of information existing in audio, video formats. A sizeable number of formats are available for every type of ob|ect be it a text, audio or video. Objects stored in formats that require commercial software to read should normally be avoided as the agencies developing collections do not know the user's ability to use such software and object. Several formats can be read by the software available in open source platforms which are normally non-commercial. A collection that provides objects in the user friendly formats certainly is better than otherwise.

Formats are normally recognised by the extension of the file name. For the text documents, the .pdf (portable document format) has become an international standard. There are some collections storing text files in .docx formats which require specific software to open and read the content. This needs to be avoided. There are several formats for the image files. TIF or PNC are considered for unquestionable best quality and archiving. For the smallest file size, these could be converted to the JPG format. The worst image format is .GIF in terms of file size and colour quality. Among the audio file formats, WAV for example is an uncompressed audio format most commonly available on the Windows platform and the quality of the audio is very good. MP3 is another format with compression with which the file size is reduced to a great extent without much loss on the sound quality.

2.4 Deposition Process

The live digital collection is one that grows. A number of repositories are initiated without a well thought policy on the method of adding new objects to the collection and their upkeep. The objects are obviously deposited by the users themselves if the collection is not organisation centred. However, it the collection is organisation centred, there is no single thumb rule or policy that leads to success. It depends on the type of organisation and internal work flow, etc.

In many organisations, where the digital collection is organisation centred, the information professionals take care of the entire cycle of deposition process. In such a situation, the responsibility lies with the individual (or a group designated this activity) to identify an object (e.g. document) that has to be added to the collection and catalogued. This individual/group is also normally responsible for the functions that pertain to the digital library (like search and retrieval, system performance and feedback). The advantages of the individual/ group monitoring input are:

- Designated responsibility
- Consistency in the input quality
- Responsible for the post-deposition cycle

Another option is distributed input. Every author (or the intellectual property rights holder) of the collection uploads the object. This has normally been mandated in many organisations where the author is

supposed to deposit his/her works to the collection. Once the item is uploaded, the repository administrator checks the validity of metadata entry, copyright issues for making the item available to Internet users and their upkeep.

In situations where there is no mandate on the authors to deposit their works to the collection, the responsibility of the repository administrator increases multifold. This is because as of todav. at least in India, there is not much awareness of the use of deposition. If the col lection is to be maintained comprehensively, the repository manager has to set a mechanism to obtain the document from the author.

The collection development process should suit the workflow of the object creation in an organisation. The depositors should not consider this as an additional task' but a part of routine process. Most of the metadata should automatically be updated from other sources of information in the workflow. There are better chances of keeping the collection up-to-date rather than having an additional task as 'deposition processes' in the organisation.

Whatever the deposition process is, this becomes the most important criteria for the evaluation of the repository as it is a decisive factor in deciding the comprehensiveness of the collection.

3 Currency

Usually the collections are not comprehensive unless the currency is maintained, i.e. the new / latest objects are added periodically. The deposition process may be by any of the methods described above. However, the collection should be up-to-date to maintain its standard. It is easy to evaluate the currency of the collections that are limited to institutional objects. However, with collections that are devoted to type, subject, etc.. there may be difficulties in setting up standards for this parameter.

4 Metadata

Metadata is the description of the object to understand scope,' restrictions, authenticity, integrity of the data presented in the object. In the case of the digital collections, metadata is the most important entity as the chances of retrieval of the relevant object is based on this information alone. It has therefore to be nothing but perfect.

The chances of entering metadata in standard formats are better in case of centralised deposition process by information professionals. Because then it becomes a person/group centred activity.

In case of distributed deposition process the metadata is described by the 'owner' (author in case of publications) of the object. The chances of information entered in such a situation is not likely to follow the standards since the author is unfamiliar with data entry and the information retrieval process. It is therefore the collection manager's role in rewriting the same, at times, becomes important. Normally, in order to maintain the quality of the data so entered, evaluate relevance of the object proposed for the collection and intellectual property issues (copyrights, etc), managers of standard collections check the validity of the entry before making the object/data open to the public over Internet. Wherever this is not done, the responsibility lies with the owner of the object depositing (and describing) it.

There may be a contradicting view to the above argument. The 'owner' of the object/document can describe data much better than anyone else! It is therefore irrespective who entered it, the completeness and accuracy become important parameters for the evaluation.

5 Respecting Intellectual Property

The digital world has made it simple to share any information objects in close groups or even among the general public. The copyright restrictions therefore are easy to violate. It is the responsibility of the collection managers to ensure that the objects made available in the collection do not violate such rights. It all depends on how stringent are the collection managers on such issues and therefore becomes a parameter for evaluation of a collection.

It would not be out of place to record here that it is also the responsibility of the publishers to tune their copyrights obtained from the authos to the changing scenario. A common platform of this information at http://www.sherpa.ac.uk/rorneo/ indicates that many publishers are now permitting the authors to deposit their manuscripts of published papers to the institutional repositories or personal Web pages.

6 Curation and Preservation

Day has explained problems and prospects of curation of the data.[8]

Curation adds value to the trusted body of information for current and future use. It is seen as an ongoing process over time. This is especially important in big science disciplines where the chances of obtaining the data once again are impossible (e.g. oceanographic observations for a given date and time in a particular area, rainfall, astronomical data, etc.) or expensive. Curation is becoming extremely difficult because of the following factors:

- Ever growing size of the digital data
- The social status of the staff engaged in curation is always secondary to that of the users of curated data, whereas the expertise required for data use and curation is similar

Parameters such as how frequently the metadata records were revisited and how often the new information about the information objects were added to make the resource useful for current and future use thus becomes extremely important. Because of the factors explained above, curation practice is hardly followed in maintaining digital collections ever-green.

It tempts me to give one practical example of curation of data. A library attached to an oceanographic institute maintained metadata about the articles published in various sources of information regularly. One day a need arose to find out how many (and which) articles were published using data collected on a particular ship. The library staff had to revisit all published literature to add this parameter in the existing metadata.

While curation of metadata is one important aspect of digital collection in easing search and retrieval of the object of information, preservation is another similar challenge and difficulty. Preservation is also seen as a set of processes and activities that ensure continued access to information from all kinds of objects stored in different formats in the digital collection. With the advancements in ICT, the storage formats are changing very fast. There is also a change in the hardware and suitable software. Migration from one environment to the other currently acceptable environment is a big challenge in terms of money, materials and manpower.

Both these aspects, though expensive in terms of returns, become important as the parameters for the evaluation of the digital collections.

7 Interoperability

Popular search engines like Google and Yahoo use Web crawlers and robots.txt protocol to search different websites on Internet and index the content. However, there are some website settings that prevent these search engines from indexing their content. Besides search engines, there are some 'harvesters' that harvest the content from various collections and index to direct the users to the right destination.

If the metadata descriptions are written in standard formats, it becomes easy for the harvesters and search engines to collect such information for indexing. Searching for right information on Internet is not an easy task. If one wants the collection to be used to the maximum, the developers of the collection must use the standard metadata formats.

OAIster, for example, is a union catalogue of millions of records representing open archive resources that was built by harvesting from open archive collections worldwide using the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH).

The open source software such as DSpace, EPrints, etc., use OAI-PMH making the life of collection developers easier for the interoperability.

8 Sustainability Over Time

This is one of the very important evaluation parameters though often considered obvious. Digital collections normally contain resources of long-term value and therefore have to be maintained forever. The commitment of an organisation to maintain the collection is of paramount importance. This is backed up by financial and technical support.

Organisations give way to dynamic individuals to give birth to such collections. At times, they are not related to the mission of the organisation which supports it. In such a situation, as the individual leaves the job/ contract/ assignment, the collection so built dies or fades awav. A wonderful collection of Indian biodiversity information building about 80,000 records disappeared the moment the individual left the organisation.

Normally, projects start with financial support from external agencies. However, such collections demand financial support forever and the agencies do not sustain this. It is the responsibility of the host institution to ensure that the collection projects are to be sustained by internal financial support once the external resources dry out. Financial support is essential for manpower as well as hardware - upgrading the servers, etc., from time to time.

9 Uniqueness

I would end with the last quality of a digital collection that can become a parameter of evaluation. An answer to "how unique is the collection?" would provide a great value to the collection. We can safely conclude that there is nothing unique in the object type today. Several collections exist of different types. However, the property uniqueness' applies to the content of the collection. A collection with rare content and no other alternatives is certainly unique and ot high value.

Collection building is not unique to the group of people in the library and information profession. We have maintained print collections for our libraries since long past. The evaluation process in building a collection is a continuous process and needs to be applied to the collection periodically to maintain the standard quality of the collection and offer the user groups something that pleases them. There are some unique differences, of course, between maintaining a print collection and the e-collection as indicated earlier and those different aspects need attention.

References

1 Friedman, T. 2005. *The World is Flat. A Brief History of the Globalised World in the 21st Century.* Penguin, pp. 488.

2 Noerr, P. 2000. *Digital Library Tool Kit.* 2nd Edition; Sun Microsystems. 186pp. (http://daminfo.wgbh.org/digital_librarv_toolkit.pdf

3 Cole, T.W. 2002. Building Good Digital Library Collections: A Dynamic Framework. *Educause Review.* 2002 Nov/Dec 2002; 12-13 (http:// net.educause.edu/ir/librarv/pdf/erm0269.pdf).

4 Xie, H. 2006. Evaluation of Digital Libraries: Criteria and Problems from Users' Perspectives. *Library & Information Science Research* 28: 433-452

5 NISO Framework Working Group. 2007. *Framework of Guidance for Building Good Digital Collections.* 3rd edition. Baltimore: National Information Standards Organisation (NISO). 2007; pp. 95 (http://framework.niso.org/)

6. Noerr, op.cit.

7 Naik. P. 2008. *White Paper: OCR Softwares for Indian Languages,* (http:// www.mail-archive.com/accessindia@accessindia.org.in/msg17879.html). Mumbai: Daisy Forum of India meeting April 11-12, 2008

8 Day, M. 2008. Current and Emerging Scientific Data Curation Practices. *Tirrenia, Italy: 4th Summer School on Preservation in Digital Libraries,* June 12, 2008. 50 slides (http://www.slideshare.net/michaelday/research-data).