RISK FACTORS ASSOCIATED WITH MIGRATION IN DIGITAL LIBRARIES

By

M. Ramasamy Pillai*

ABSTRCT

Discusses the trend and the need for the transformation of traditional library collections from print into digital form. The concept of Hybrid Library is also briefly presented. Migration as preservation strategy is discussed. The influence of technological innovations on migration of digital collections is also discussed. Different risk factors involved in migrating the digital data from one media and format to another are highlighted.

---

* Documentation Officer, INSDOC Centre, CSIR Complex, Taramani, Chennai – 600 113

0. Introduction

The new realities of Internet publishing are that information is provided in a widely distributed manner, and it is up to the consumer to locate what is needed. Those who use the Internet regularly find what they want by employing search engines to seek out all electronic documents that contain certain combinations of words, names, or even acronyms. Through networked digital libraries we can access all the electronic resources available globally. We all stand to benefit greatly from networked digital libraries. While digital technologies are enabling information to be created, manipulated, disseminated, located and stored with increasing ease, preserving access to this information poses a significant challenge. A Digital Library (DL) is a collection of electronic information organized for use. To meet user needs, the founders of DL must accomplish two general tasks : establishing the repository of electronic materials, and implementing the tools to use it. The objective of digital libraries is to acquire information, organize it, make it available and preserve it in a digital environment. Most important, the ability of the digital library users to give serious weight to electronic information depends upon their trust in such information being dependably available, with authenticity and integrity maintained.

1. Hybrid library

The UK eLib Electronic Libraries program has coined the term "Hybrid Library" to cover services that unite the functions of the traditional library with those of electronic, digital or virtual library services. A hybrid library is envisaged as the bringing together of technologies from electronic, digital or virtual library as in the UK's eLib programme, plus the electronic products and services already exist in the libraries. A hybrid library environment can be described as one where an appropriate range of heterogeneous information services is presented to the user in a consistent and integrated way via a

single interface. It may include local and/or remote distributed services both print and electronic. For most libraries, the implementation of a web-accessible catalogue has been the catalyst for creating a local hybrid library.

This has enabled the provision of a web interface that allows the user to access:

* The books and other physical information resources in the library's collections;
* Digital copies of physical information resources in the library's collections.
* CD-ROMs and online information resources which the library is licensed to access on behalf of its users, including full-text databases, union catalogues, indexing and abstracting services; encyclopaedias and other reference tools.
* Information resources freely available on the Internet.

2. Migration and emulation

The steady growth of digital information as a component of major research collections has significant implications for many libraries. Many institutions have been creating or collecting digital information produced in a wide variety of standard and proprietary formats, including ASCII, word processing, spreadsheet, and database documents. Each of these formats continues to evolve, becoming more complex as revised software versions add new features or functionality. The most pressing problems confronting managers of digital collections are data format and software obsolescence. It is also important to decide what should be preserved, in what priority, and with what techniques. Currently, there are two radically different strategies for managing the life cycle of a digital collections : Migration and Emulation.

Migration is broadly defined, as "the periodic transfer of digital materials from one hardware/software configuration to another, or from one generation of computer technology to a subsequent generation". A more specific definition would indicate that migration changes the structure of the original data file. With the exception of files that are simple data streams, most files contain two basic components: structural elements and data elements. A file format represents the arrangement of the structural and data elements in a unique and specific manner. In this context, migration is the process of rearranging the original sequence of structural and data elements (the source format) to conform to another configuration (the target format). The migration of digital information is a continuous process in many libraries as a method of dealing with technological obsolescence. Migration to new operating environments often means that the copy is not exactly the same as the original piece of information. Decisions need to be made about what aspects of the material to be migrated (eg functionality, presentation) and to be preserved.

The complexity of the migration process will depend on the nature of the digital resource which may vary from simple text to an interactive multimedia object. At its simplest, migration may entail copying digital information to a more stable non-digital medium, such as paper or microfilm. Transfer to a more stable digital medium like CD-R from floppy disk offers a short- to medium-term strategy for preserving access, but still

requires the CD-R to be migrated when the technology changes. Other factors which can influence migration are time, cost, frequency at which digital information will need to be migrated in the future as technology evolves. While adherence to standards will assist in preserving access to digital information, it must be recognised that technological standards themselves are evolving rapidly. In practice, migration is prone to generating obvious and subtle errors. An obvious error occurs when the set of structural elements in the source format does not fully match the structural elements of the target format. A subtle error might occur if the data themselves do not convert properly. In other situations, migration might preserve the content of the file but lose the internal relationships or context of the information. New media for storing digital information rapidly replace older media as reading devices for these older media become no longer available. Consequently, digital data which relies on obsolete technologies becomes inaccessible. Currently, it seems that the lifespan of digital storage media generally exceeds the technology that supports it.

An alternative preservation approach, emulation, is concerned with preserving the original software environment. Emulators are programs that mimic computer hardware. Emulation is essentially a way of preserving the functionality of and access to digital information which might otherwise be lost due to technological obsolescence. Strategies adopting this approach store copies of the initial software and descriptions of how to emulate the initial hardware to run the software along with the data files. Newer versions of software constantly render older versions obsolete and the hardware required by this software also changes over time. One of the benefits of the emulation strategy compared with migration is that the original data need not be altered in any way. It is the emulation of the computer environment that will change with time.

3. Migration and risk factors

Following are some major categories of risk identified while considering migration as a digital preservation strategy option:

3.1 Infrastructure:

The presence or lack of institutional support, funding, system hardware and software, and the staff to manage the digital collections are some of the important issues. For example, digital media is considered as a cost effective alternative to microfilm preservation. It will continue to remain so for sometime until newer storage media are evolved. The concept "technology refreshing" calls for periodic migration to new formats and new storage media as technology develops. Legal and policy issues associated with digital information will introduce additional risks. The collection, and the library users who use the collection, will be affected to some degree by the migration of data.

3.2 Data file format:

These include the internal structural elements of the file that are subject to modification during migration. Migration or the conversion of data from one format to another, has

measurable risk, and the quantity of risk will vary, sometimes significantly, given the context of the migration project. One form of risk depends on the nature of the source and target formats. Adoption to standards can assist by facilitating the transfer of information between hardware and software platforms as technologies evolve. Use of standards can also help ensure best practice in the management of digital information. Since basic file structure concepts are common to many file formats, experience with one format can be used to understand other formats. Tagged Image File Format (TIFF) is one of the master storage formats for scanned images developed by Aldus and Microsoft, and the specification was owned by Aldus, which in turn merged with Adobe Systems, Incorporated. It is a highly flexible and platform-independent file format. It is supported by numerous image-processing applications. Each file contains the digital data from the scanned page and a header that describes the characteristics of the image file.

Format migration affects watermark, digital stamp, or other cryptographic techniques. Because of different hardware and software dependencies, reading and processing the new file format require a new configuration. Linkages to other files are altered during migration. New file format may reduce the file size due to newly adopted compression technology and causes denser storage and potential directory-structuring problems if one tries to consolidate files to use extra storage space. File extensions change because of file format upgrade. A number of problems associated with using standards for preserving access to digital information have been noted.

The eLib Standards Guidelines (October 1998) lists some of the main problems as being:

* Several versions of a standard may be in use, with earlier versions no longer being compatible
* Suppliers may offer "value added" versions of standards in their implementations
* A standard that is not well specified may be differently implemented in software or some standards may have more features than are likely to be used, resulting in different subsets being used in different implementations.

3.3 Conversion software :

The conversion software may or may not produce the intended result. Conversion errors may be gross or subtle. By and large the following risk factors are identified.
Image quality includes resolution, colour spaces etc may get affected by alterations to the bit configuration. Conversion process results in cross examining a file before and after the migration process. A test file can be passed through the conversion software, migrating from source to target format. If the fields and field values of the original source file are properly transferred into the target file, the risks incurred in migration are significantly reduced. The risk of migration is significantly increased when the original data is not reproduced in the target file. When the field tags and values in the test file are known, data changes associated with file conversion can be independently verified . Risk assessment tools can be evolved with the proper testing of locally or commercially developed data migration software.

Because of the following factors file migration technique can also be considered

* The routine refreshing of digital files;
* Varying changes in digital formats when files are converted from one application to another;
* Radical changes in digital formats, such as the conversion of numeric files from proprietary formats to ASCII

Conversion software should meet basic performance requirements :
Based on the review of conversion programs, Cornell determined that migration software should perform the following functions:

* Read the source file and analyze the differences between it and the target format
* Identify and report the degree of risk if a mismatch occurs
* Accurately convert the source file(s) to target specifications
* Work on single files and large collections
* Provide a record of its conversions for inclusion in the migration project documentation

Using software which is 'backwards compatible', that is, a later version of the software can decode files created on an earlier version, will simplify migration. Interoperability of systems will also facilitate migration by obviating the need to run a specific program in order to be able to access a digital resource which was created using it.

Considering the cost of writing conversion software for a wide range of file formats, a commercially developed solution for migration software will ultimately be cheaper and more flexible than locally developed conversion software.

3.4 Cost:

Long-term costs associated with migration are unpredictable because each migration cycle may involve different procedures, depending on the nature of the migration. Costs may be un-scalable unless there is a standard architecture (e.g., centralized storage, metadata standards, file format/compression standards) that encompasses the image collections so that the same migration strategy can be easily implemented for other similar collections.

3.5 Staff:

Staff turnover and lack of continuity in migration decisions can hurt long-term planning, especially if the migration path is not well documented. Decisions must be made whether to hire full-time, permanent staff or use temporary workers for completing the tasks. Staff may have insufficient technical expertise.
The unpredictability of migration cycles makes it difficult to plan for staffing requirements (e.g., skills, time, funding).

3.6 Functionality:

Features introduced by the new file format may affect other functions, such as printing. If the master copy is also used for access, changes may cause decreased or increased functionality and require interface modifications (e.g., static vs. multi-resolution image, inability of the Web to support the new format). Unique features that are not supported in other file formats may be lost (e.g., the progressive display functionality when Graphics Interchange Format [GIF] files are migrated to another format).

3.7 Legal:

Factors such us Copyright regulations may limit the use of new digital outputs that can be created from the new format (e.g., the institution is allowed to provide images only at a certain resolution so as not to compete with the original).

4. Conclusion

There appears to be a trend for many publishers of physical format publications to move towards e-publishing using the internet as the new medium for their publications. There may be scope for the provision of customised electronic copies of publications for libraries, which will be suitable for preservation purposes. It is clear that preservation and access are inextricably linked. Long term access cannot be provided unless preservation is achieved. It is recognised that migration and emulation are not a guarantee of preservation but appear to be necessary steps to increasing the duration of accessibility of information. The feasibility and value of migration across operating systems and other major technology changes will remain unclear without further testing . But some libraries examine the possibility of emulation as an alternative to migration in preserving the digital collections. Alternatively, service bureaus can also be entrusted with the job of migration on contract basis to develop standards and procedures for the creation, storage and use of digital collections largely due to economic and risks considerations.

5. References

1. Gregory W. Lawrence, Risk Management of Digital Information : A File Format Investigation, June 2000

2. Library of Congress. Preservation Issues , July 1999

3. Bennett, John C , A Framework of Data Types and Formats, and Issues Affecting the Long Term Preservation of Digital Material. British Library Research and Innovation Report,1997

4. www.clir.org

5. Judith Pearce and Warwick Cathro, The Challenge of Integrated Access: The Hybrid Library System of the Future. 10th VALA Biennial Conference and Exhibition, Melbourne,2000

6. eLib Project Summary: Hybrid Libraries, Joint Information Systems Committee (JISC), 1999.

7. Rothenberg, Jeff , Avoiding technological quicksand: finding a Technical foundation for digital preservation. CLIR Reports,1998

8. Arms, Caroline R. Risk Management of Digital Information: A File Format Investigation RLG Digi News,2000

9. Woodyard, Deborah, Farewell my Floppy : strategy for migration of digital information,1998

10. Hedstrom, Margaret , Digital Preservation: a time bomb for Digital Libraries,1995