LIBER QUARTERLY, ISSN 1435-5205
© LIBER 2001. All rights reserved
K.G. Saur, Munich. Printed in Germany

Management of the Life Cycle of Digital Library Materials

by MARILYN DEEGAN

INTRODUCTION

This paper does not deal in any detail with digital preservation, rather it examines the development of digital content and digital libraries in the broader content of digital developments and library developments. Implicit throughout this is the need to create or acquire digital content for long-term access, which means that preservation needs must be foregrounded at all stages of the digital lifecycle[1].

DIGITAL MATERIALS AND THEIR ORIGINS

Digital materials are usually divided into two broad categories: „born digital“, materials created in digital form from the start which may or may not have an analogue output, and „reborn digital”, materials digitized from analogue originals which may or may not be retained. The distinction is in many ways not always that easy to make, and in preserving library objects it is vital to keep in mind the fact that print and digital intertwine as different instantiations of an information object, sometimes it is the analogue version satisfying the users need, sometimes the digital. Data may be searched and then printed out, the digital being a route to another analogue object. Sometimes digitization is used so that access to the analogue object can be restricted, sometimes materials are digitized so that the originals can be deaccessioned (though this is rarer now than was once predicted). So the relationship between the digital and the analogue remains complex, and both need to be factored into the lifecycle vision.

Almost all materials are now produced in digital form, and increasingly the digital is an equal or even primary product to the print, for instance in the case of e-journals, major reference works like the online OED or the Encyclopaedia Britannica, and the increasingly popular ebooks.

In a recent article, Robert M. Braude points out the difference between the „product that we manage in libraries, information, and the familiar container for that product, the codex book" (Braude, 1999, 85). These containers have influenced library architecture, but they do not themselves define what a library is. Braude suggests that as „we did not bother to qualify our libraries by calling them clay libraries or papyrus roll libraries, why now do we have to call them digital libraries?" (Braude, 1999, 86). While these comments are valid, and in terms of the intellectual management and understanding of data need to be taken seriously, there is a distinction to be made between traditional libraries and digital libraries (or digital collections within libraries). The physical containers for information are capable of direct access, and are managed physically—they are stored in environments best suited to their particular needs and delivered physically to the users for access. Digital data is ephemeral, fragile, can be detached from its „container" with (usually) minimal effort (which works inscribed in codices cannot), can be replicated indefinitely almost to infinity, can be altered without trace. While the differences between analogue and digital data may be of degree more than of substance, they are sufficiently large to need different approaches.

The relationship between container and information is, in fact, a complex one, where form and function intertwine to contribute to meaning. The book allows much more sophisticated organization of information than did the scroll; interestingly, the computer screen owes more to the scroll than to the book, and though hypertextual linking is complex, it has probably not yet achieved the sophistication of the printed page for information presentation. There are, of course, many advantages of digital data, but, interestingly, the terminology of the print world still rules: digital collections are digital libraries, new electronic texts are eBooks, and Whittaker's guide to the availability of CD-ROMs is called CD-ROMs in Print.

THE ADVANTAGES OF DIGITAL CONTENT

Whether materials are born or reborn digital, their management for the long-term is something that almost all libraries are now having to consider, and though in terms of the satisfaction of user need the role of the library is no different that it ever was—to bring information to users or users to information—digital content management brings into play a whole range of new factors and requires new skills of librarians. There are, however, sufficient advantages to the provision of digital data to make tackling the difficulties which they might pose for libraries worthwhile. Digital data can be accessed from anywhere there is a network connection and a terminal or computer (subject to the necessary authorizations). It can be searched and manipulated in ways impossible to manage in the analogue world, it can be downloaded, printed, annotated, shared, or exchanged. Importantly in a library environment, digital data has permeable boundaries—that is, digital media can be interrelated in ways that make it difficult to tell where one object begins and ends. In a digital version of an article, for instance, the footnotes can be live links to the referents, allowing instant access. If, in turn, that resource has live links to other referents, then the user can pass from article to book to article ad infinitum picking up information along the way. Given the underlying structures of digital data, these links can be made between many different formats—images link to text, text to maps, texts to video or sound, and so on. And the information objects do not need to reside in the same collection, or even in the same country or continent. As long as the network connection can be made, the link can be followed. From virtually anywhere, the user can access resources from virtually everywhere—digital data has the virtues of ubiquity and simultaneity.

The permeability of boundaries in the digital world has some interesting consequences (and disadvantages): if we are unsure where the information objects begin and end, how can we know who owns them? What are the implications for the management of intellectual property rights and for payments to owners or rights holders? How are we to cite sources? How can the user trust that the object is what it purports to be? In short, who manages what parts of the digital lifecycle? Libraries have only been managing digital content on any scale for the last twenty years, and so there is still much to be learnt. However, librarians have coped with many new formats and the technologies to access them over a longer period than this, and are rising to the challenges well, and seeing and grasping the new opportunities offered.

Digital Collection Development

In libraries, digital collection development is, in many ways, no different from any other kind of collection development, as suggested above. But it may be different functionally as relevant content cannot be acquired if there is no means of delivering it to the user because of some technical barriers: huge satellite images when the bandwidth of the library network is inadequate, for instance. Digital content is being delivered in libraries that derives from many different sources and under a whole range of divergent financial arrangements or controls. Some of the key sources of digital data are:

- Library's own holdings which have been digitized
- Purchased datasets on CD-ROM
- Purchased datasets which are online
- Electronic publications which have a paper equivalent
- Electronic publications which have no paper equivalent
- Electronic reference works which increasingly have no paper equivalent
- Ebooks.

What is interesting here is that there is a shift in the notion of what constitutes a „holding" as many of these resources are accessed through the library, but come from a range of remote sources as well as from within the library itself, from other libraries, consortia, aggregators, publishers. As digital content can be accessed from anywhere (almost), traditional notions of supply and access no longer apply.

The digitization of resources opens up new modes of use, enables a much wider potential audience and gives a renewed means of viewing our cultural heritage. These advantages may outweigh the difficulties and disadvantages, provided the conversion process is well thought-out. Institutions large and small are therefore embarking upon programmes of digital conversion for a whole range of reasons. The advantages of digital surrogates include:

- Facilitating immediate access to high demand and frequently used items
- Easier access to individual components within items (e.g. articles within journals)
- Rapid access to materials held remotely
- Ability to reinstate out-of-print materials
- Potential to display materials which are in inaccessible formats, for instance, large volumes or maps
- „Virtual reunification"—allowing dispersed collections to be brought together
- The ability to enhance digital images in terms of size, sharpness, colour contrast, noise reduction, etc.
- The potential to conserve fragile/precious originals while presenting surrogates in more accessible forms

- The potential for integration into teaching materials
- Enhanced searchability, including full text
- Integration of different media (images, sounds, video etc)
- Satisfy requests for surrogates—such as photocopies, photographic prints, slides etc.
- Reducing the burden or cost of delivery
- The potential for presenting a critical mass of materials.

Whether digitizing library collections or acquiring digital materials from elsewhere it is important to apply a lifecycle approach. Digital collections will need to be managed in a much more active way than documentary collections if they are to give value to their creators and users, and to survive for the long term. As Jones and Beagrie (2000, 18) suggest:

> The major implications for lifecycle management of digital resources, whatever their form or function, is the need to actively manage the resource at each stage of its lifecycle and to recognise the interdependencies between each stage and commence preservation activities as early as practicable. This represents a major difference with traditional preservation, where management is largely passive until detailed conservation work is required, typically many years after creation and rarely, if ever, involving the creator. There is an active and interlinked lifecycle to digital resources which has prompted many to promote the term „continuum" to distinguish it from the more traditional and linear flow of the lifecycle for traditional analogue materials.

While Jones and Beagrie are referring to the lifecycle approach in terms of digital preservation, the concept, by its very nature, has to be built into all stages of the creation, accessioning and management of digital resources. The UK-based New Opportunities Fund, which in 2000 made available some £50 million for digitization projects[2], suggests in its technical guidelines that the stages of the lifecycle are: creation, management (for the long term), collection development (that is, aggregating the digitized materials in cohesive sets), access and repackaging (or ensuring reusability). Digitization and the management of digital resources are costly activities, and it is vital that the value of the resources created can be realized over the longest possible period: creating sustainable resources is the best way to ensure this.

The most important reason for this lifecycle approach is to create a sustainable resource right from the start. For many analogue originals, there will only be one chance to digitize it and so it is crucial that the digital files are „future-proofed" as far as that is possible. The goal (desirable but not always possible) is one-time capture for all future uses, so planning for the long-term is essential. This is no easy matter, as the long-term issues and consequences of creating and managing digital objects cannot be known, and in particular it is difficult to estimate on-going costs. But the better a digitization project is planned, the more likely it is that the digital materials will have a long and useful life.

WHAT DOES A DIGITIZATION PROJECT INVOLVE?

Whilst there is significant variation in the original materials and the potential methods of digitization there are some common themes to every digitization project. Firstly, it is essential to assess the original materials to identify the unique characteristics of the collection. These unique characteristics will drive the digitization mechanisms and help define the required access routes to the digital version. Additionally, whether the end product is one data file or thousands they will have to be organized, given file names and placed in some logical structure. Having a clear vision of the information goals to be achieved from the original materials and the means of delivery are essential to a successful digitization project.

It must also be remembered that digital capture (generally thought of as the major activity in creating digital resources) is only one of the many processes involved in the highly complex chain of activities which are attendant upon the creation, management, use and preservation of digital objects for the long term. Capture is likely to incur only a relatively small proportion of the total project costs. Any digitization project is likely to involve some or all of the following activities:

- Assessment and selection
- Grant writing and fundraising
- Feasibility testing, costing, and piloting
- Copyright clearance and rights management
- Preparation of materials
- Benchmarking
- Digital capture
- Quality assessment

- Metadata design and creation
- Delivery
- Workflow processes
- Project management
- Long-term preservation.

Without careful planning for all these elements, projects are unlikely to succeed. Costs will rise, deadlines slip and acceptable quality may not be achieved. Detail about the diverse practical activities involved in carrying out digitization projects is outside the scope of this paper, but there are many excellent guides available in both print and on the web. See Hazen, Horrell, and Merrill-Oldham (1998); Kenney and Rieger (2000); the Arts and Humanities Data Service[3]; the Higher Education Digitization Service[4]; RLG DigiNews[5] and D-Lib Magazine[6] are essential reading for anyone planning a digitization project.

LONG-TERM PRESERVATION FOR ACCESS: SOME KEY ISSUES

Digital data is at risk of loss because it is recorded on a transient medium, in a specified file format, and it needs a transient coding scheme (a programming language) to interpret it. The basic unit of data, the „bit" (BInary digiT), is represented by one of only 2 states: „1" or „0", linked together in a „bit stream" consisting of many millions of bits or electrical impulses. The kind of digital data which concerns us here is complex, and meaning derived from data can depend as much on how individual data objects are linked as on what those objects are. Of course, written documents are also highly complex objects, but their structure does not need to be comprehended for their preservation, only for their interpretation. Over time, knowledge of how to interpret documents can be lost, but this can usually be recreated, as their textual and physical characteristics are explicit. Their decipherment generally needs only human faculties. As Rotherberg points out, with physical documents „saving the physical carriers saves all those attributes of the original that it is possible to save" (2000, 16). With digital data, a machine needs to be interposed between it and its human interpreter, which adds another layer of complication.

There are two key issues for data preservation, which surprisingly have little to do with preserving the original bit stream:

- Preserving the physical media on which the bit stream is recorded
- Preserving the means of interpreting, reading and utilizing the bit stream.

Given that the bit stream is merely a very long series of binary codes, the preservation of the physical media should maintain its integrity over time. However, being able to read, use or interpret that bit stream may become increasingly difficult as systems evolve, adapt and eventually become redundant, so presenting a fog through which the bit stream becomes unusable.

Digital data is in danger, not because it is inherently fragile or flawed, but because there is a continually accelerating rate of replication, adaptation and redundancy of hardware, software and data formats and standards which may mean that the bit stream may not be readable, interpretable or usable long into the future. All data is stored as a code and therefore requires an element of decoding before it is recognizable and usable in a computing environment. We take this automatic decoding for granted until we try to read a word processing file from 10 years ago and find that none of our current systems or software have any idea what the bit stream means without significant coaching or expert help. The longer the data is left unattended, its data coding unrecorded, systems will become obsolete and the expertise to recognize and decode that specific type of bit stream will become unavailable. Data could be lost forever, unrecoverable without effort that will probably not be cost-effective.

Digital data needs much more active, interventionist methods of preservation from a much earlier stage in its lifecycle than analogue, which is why such stress is laid on lifecycle management. Digital preservation is receiving urgent attention in the international library community because of the exponential developments in computer technologies, software and data storage methods. Digital data does not have a long enough natural lifetime for us to wait for better media to come along but, as yet data storage has not found its stability equivalent of paper or microfilm, though the evolution of the technology may be around the next corner. The storage of data started with punched cards only some 50 years ago and has transitioned through paper tape, magnetic tape, to magnetic disk, optical disk and portable memory such a flash memory cards to the present day. It is now extremely difficult to find card or tape readers if old archives of originals come to light, even after as short a time as 30 years.

Much discussion is currently taking place on just how digital data is to be preserved for the long term, and a mixture of methods is likely to be adopted by libraries including migration of data to new storage media; reformatting data so that it can run on new hardware and software platforms; and the emulation of former platforms in the case of highly complex data. It is almost impossible to know what the costs for data preservation are likely to be over the long term, which is a strategic nightmare for library managers. In the future, management of data from early in its lifecycle should hopefully make the issues and costs more knowable, but the present situation is that there is much valuable data from the past and present that is in severe danger of loss.

Conclusion

This is the briefest possible overview of some of the keys issues and strategies in digital lifecycle management for long-term access to valuable content, but it highlights some of the main points that it is necessary to consider. Those needing more detailed approach are referred to Deegan and Tanner (2001), and also to Jones and Beagrie (2001).

References

Braude, R.M.: Virtual or actual: the term library is enough. In Bulletin of the Medical Librarians Association, 87 (1) 1999, pp. 85-87.

Hazen, D., Horrell, J., and Merrill-Oldham, J.: Selecting research collections for digitization, CLIR, 1998, available at <http://www.clir.org/pubs/reports/hazen/pub74.html>.

Jones, M. and Beagrie, N.: Preservation management of digital materials workbook, Arts and Humanities Data Service; JISC Preservation Focus. Pre-publication draft. 2000.

Kenney, A. R. and Rieger, O. Y. (eds): Moving theory into practice: digital imaging for libraries and archives, Research Libraries Group. 2000.

Rothenberg, J.: Using emulation to preserve digital documents, Koninklijke Bibliotheek, 2000.

[1] The concepts discussed in this paper are dealt with in more detail in Marilyn Deegan and Simon Tanner, *Digital Futures: Strategies for the Information Age*, Library Association Publishing, 2001 (forthcoming).

[2] <http://www.nof.org.uk/tempdigit/index.htm>

[3] <http://www.ahds.ac.uk/>

[4] <http://heds.herts.ac.uk/>

[5] <http://www.rlg.org/preserv/diginews/>

[6] <http://www.dlib.org/>