Preserving Information for Posterity: Is 'Going Digital' the Answer?

Natasa Milic-Frayling

Microsoft Research Ltd
Roger Needham Building, 7 J J Thomson Avenue, Cambridge, UK
natasamf@microsoft.com

Abstract. Digitization has been adopted as a strategy for preserving content of deteriorating physical artefacts. At the same time, by removing the boundaries of physical containment, it provides new opportunities for sharing and exploiting information. Generally, the use of digital format has dramatically changed how we create, manage, and communicate information. However, ensuring long term preservation of digital media is a non-trivial matter. Failing to find an adequate solution threatens the survival of valuable information created in the digital era. In this paper we reflect on the many issues associated with digitization and preservation of the digital content. We describe the economic climate that sets the context for increased activities in this area.

Keywords: digital library, digital archive, metadata, standards, XML, preservation.

1. Introduction

Proliferation of information and communication technologies has resulted in a dramatic shift from recording information in paper manuscripts, photographs, and audio-visual tapes to creating content in the digital format. Our ability to produce, publish, and communicate digital content with ease and in abundance, has transformed the way we view and manage information. We recognize the benefits of having relevant information at the right time and continue to develop technologies that enable us to exploit information optimally.

Through various digitization techniques, we transform the traditional documents into the digital format and exploit it alongside the 'born digital' content. In this paper we reflect on the increased value of information that stems from the highly agile nature of the digital format and the risk of loosing digital information unless we take appropriate steps to ensure its long term preservation. We discuss the current business climate that drives digitization and possible economic models that can make digitization a self-sustainable effort.

2. From Physical Artefacts to Digital Media

Through generations, information has been recorded to pass on the acquired knowledge and experience. Nowadays, information is mostly created in the digital form, even when it is disseminated as paper books, newspapers, and similar. The digital content has the advantage of being amenable to automatic analysis and aggregation. We can more easily create new knowledge with the aid of software analysis tools and disseminate information using online communicate channels. Thus, it is not surprising that the content of books, sound recordings, and video material have all been converted to the digital form. Digitization techniques essentially *liberate the content from its containment within physical objects*. The digital format significantly increases *information agility*, which in turns has significant social, economical, political, and legal implications on societies.

Furthermore, digitization has been adopted as a strategy for preventing the loss of information from deteriorating paper artefacts, film repositories, and sound recordings. Technological requirements for this process have instigated research in optical character recognition, speech recognition, video and image processing, and related areas. At the same time, the infrastructure requirements for managing the digitized data are pushing the limits of typical IT architectures and dramatically changing the cost structure.

The cost of digitization is significant. It may cost up to £1 to digitize and apply OCR to a single page of a newspaper. The national, regional, and international newspaper collection at the British Library alone contains approximately 750 million

pages, some dated back to 18th Century. In order to make this effort economically feasible, we need to identify economic forces that can drive digitization of deteriorating information resources. Assuming that the public sector cannot absorb all the cost of this effort, what are the incentives for businesses to get involved in content digitization? While current or recent information is probably most valuable to businesses, most of it is copyright protected and thus needs to be handled appropriately. It is clear that the success of the effort depends on our ability to address many intricate issues.

In the following section we reflect on the recent activities in content digitization that have been influenced by a highly competitive on-line search market and the associated advertising business.

2.1 Business Climate and Digitization

Almost all major libraries have been engaging in digitization projects over the past decade. Such efforts are typically sponsored by national funding agencies or private donations. It is interesting that the recent boost to their digitization efforts came from investments by businesses involved in Web search.

Online information services gain revenue from advertisements related to search results and clicks on ads placed on Web pages. Typically, the 80-20 rule applies by which a large portion of search service revenue (80%) is generated by a small percentage of queries (20%) that correspond to most popular topics and Web content. Thus, the main business objective is to gain the market share of on-line queries. That is typically achieved by strategies to increase the users' loyalty though an improved on-line experience. In addition to the user interface enhancements and branding through the browser extensions, online services are expanding the spectrum and increasing the quality of the content they include in their search results. Instead of collecting and indexing only freely available Web data they are partnering with publishers to provide access to recently published premium content. With libraries and archives, they are exploring ways to digitize and bring on-line valuable content from more distant past.

2.1.1 Investment in Content Digitization

Increased competition in the Web search marketplace has caused a flurry of activities in content digitization. In summer 2005, Google reached an agreement with three leading university libraries in the US, university of Stanford, Harvard, and Michigan, with the Public Library in New York City, and the Oxford University Library in the UK to scan and index selected material. The scanning involves creation of two copies, insuring that the material is fully indexed and searchable through the Google book search services (http://books.google.com/googlprint/library.html). For legal considerations, the selection of the material is restricted to the works that are out of the copyright. The digitization is carried out in the dedicated scanning centers established locally.

On October 3, 2005, the Internet Archive, Yahoo! Inc., Adobe Systems Inc., the European Archive, HP Labs, the National Archives (UK), O'Reilly Media Inc., Prelinger Archives, the University of California, and the University of Toronto

formed the Open Content Alliance (OCA) (http://opencontentalliance.org), a global consortium focused on providing open access to content while respecting the rights of copyright holders. The main remit of the OCA is to provide infrastructure and services to enable permanent storage and free downloads of material, including cultural, historical and technological digitized print and multimedia content from libraries, archives, and publishers. In October 2006, MSN joined the Open Content Alliance and reached an agreement with the British Library in the UK to digitize 100,000 selected books in partnership with the Open Content Alliance.

2.1.2 Coordination Effort by the European Union

Following the early signs of business initiatives to engage with academic libraries and invest in digitization, on April 28, 2005 six Member States from the EU put forward a request for an organized effort by the EU Commission to harmonize and coordinate national digitization efforts across Europe. On September 30, 2005 the EU Commission responded in favour by releasing a Communication document and a call for online public consultation "i2010: Digital Libraries", inviting feedback on important issues around preservation of the national heritage through digitization [2], [3]. The scope of the challenge faced by the EU member states in preserving the cultural and national heritage of European nations is best illustrated by the statistics quoted in the Communication document [2]:

> *"The total number of books and bound periodicals (volumes) in European libraries (EU 25) was 2,533,893,879 in 2001"*. (Ibid)

The concern about the deteriorating material particularly applies to the audiovisual documents since the analogue formats deteriorate with time and cause a loss of content:

> *"A survey of ten major broadcasting archives found 1 million hours of film, 1.6 million hours of video recordings, and 2 million hours of audio recordings. Total European holdings of broadcast material are probably 50 times larger. Most of the material is original and analogue. 70% of the material is at risk …"* (Survey by the IST Presto project, Oct 2002, http://presto.joanneum.ac.at/index.asp).

The request for consultation referred to specific questions on digitization and online accessibility of digitized content as well as the long term preservation of digital media (Table 1). In March 2006 the EU Commission published a report on the responses received from 225 contributors [4] and on March 27, 2006 formed the High Level Expert group to assist with defining the strategy for the EU Digital Library effort.

Table 1. 'i2010 digital libraries' Questions for online consultation, by the EU Commission. September 30, 2005 [3]

Digitisation and online accessibility	
1)	*What additional measures could be taken at national and European level to encourage digitisation and online accessibility of material in all European languages?*
2)	*What measures could be taken to promote private investments and new business models such as public-private partnerships for digitising and making historical collections accessible?*
3)	*What measures of a legislative, technical, organisational or other nature, could facilitate the digitisation and subsequent accessibility of copyrighted material, while respecting the legitimate interests of authors?*
4)	*Is the issue of orphan material economically important and relevant in practice? If yes, what technical, organisational and legal mechanisms could be used to facilitate wider use of this material?*
5)	*How could public domain material and other material available for general use (voluntary sharing) be made more transparent and widely known in order to facilitate its online availability for subsequent use?*
Preservation of digital content	
6)	*What priority measures – in particular of an organisational and legal nature-– should be taken at national and European level to optimise the preservation of digital content with the limited resources available?*
7)	*Is there a risk that national legal deposit schemes lead to a multiplication of requirements on internationally active companies? Would European legislation help avoiding this?*
8)	*How could research contribute to progress on the preservation front? Which axes of work should be addressed in priority by the forthcoming Specific Research Programmes as part of the 7th Framework Programme?*

The consultation responses cover a wide range of suggestions from various stakeholders, reflecting different interests and perspectives on the digitization and preservation issues. For details we refer the reader to the full report. Here we outline the issues that we believe are in the very core of the content digitization challenge.

The shear scope of the digitization effort calls for a long term commitment and systematic approach to the key issues of storage, access, and preservation. We expect that some coordination of the effort can help, such as establishment of reporting and information services that hold information about all the content that has been already digitized, or at least about the content that incurs high digitization cost and thus duplication of effort should be avoided as far as possible. However, we believe that it is most important to establish a rich ecosystem around information services industry that will drive the digitization process in an economically viable and self sustainable manner. Let's reflect on a couple of key issues.

Guiding Principles for Content Selection

Defining a principled way of prioritizing material for digitization is essential. That is rather difficult considering the number and inter-dependencies of factors that need to be taken into consideration. Such are the condition and deterioration rate of physical artefacts, the cost, the relevance of the digitized content, and the legal, social, and technological constraints on the access to the digital content.

The value of information is not absolute; it depends on the context, in particular, its relation to the current events and needs. *Thus, the value of past information is maximized when optimally aggregated with the contemporary information.* Starting with this premise, we realize that prioritizing material solely on how relevant we expect it to be in the future is difficult – it is equivalent to predicting the future itself. A correct inference is feasible, however, for materials tied to recurring events. Second, it implies that the selection strategy should be tied to the *content exploitation models*, in fact, the model that identifies the demand for a particular type of information and adds value through integration with a related archived content.

One example is the aggregation of historical data with educational material. Augmentation of text books with digitized content of related archived documents provides a clear add-on value that can be captured through the supplementary cost of educational material and re-invested into digitization. The key is a clear connection with and full integration with the educational curriculum. Furthermore, at the national level, teaching history, language, geography, and literature is primarily done in the native language and focused on the national aspects of the shared history. Thus, education scenarios are particularly amenable to boosting digitization of materials in the native language.

Preserving Value Distribution through Copyright and Digital Rights Management (DRM)

In order to ensure that publishers and authors can recover the value of the published work, it is absolutely necessary to have two pieces of technology in place: a DRM and a micro-payment technology. Once the information stakeholders can control the revenue, they will be open to providing information online. A successful service will respect legal requirements and ensure that the interests of publishers of copyrighted information are protected and the protection of authors' rights is enforceable.

2.2 Metadata and Search

It is interesting to draw an analogy between online search and catalogue based retrieval in libraries. On-line search engines do not deliver the content of the live pages but rather provide a list of Universal Reference Locators (URLs) pointing to the Web servers that host the content of the result pages. The results are obtained on the basis of content features that the search engine automatically extracts from the crawled pages and hyperlink structure. Although the term metadata is typically defined as 'the structured data about the data', in a broad sense, we can say that the search engine extracts various types of metadata and uses it for indexing and search

Traditional libraries use carefully structured metadata for referencing physical objects and the quality of library catalogues is absolutely essential for accessing archived material. The first step towards 'digitization' of library services involved creation of electronic versions of catalogues and search over bibliographic data and abstracts. Searching for an item results in a bibliographic record with an indicator of the location where the physical item can be found in the library. The British Library currently stores 26 million catalogue entries in their Integrated Library System, comprising of the subject, title, and author information, with another 25 million still

to be entered. Back in 1970's, search over library metadata, initiated a flurry of new activities and brought to existence an exciting area of research - online information retrieval.

We also note the emergence of Web archives. Organizations such as the San Francisco based Internet Archive (www.archive.org) collect and store Web data for future reuse. Brewster Kahle, Digital Librarian, Director, and Co-Founder of the Internet Archive has set an ambitious goal:

> *"The Internet Archive is building a digital library of Internet sites and other cultural artefacts in digital form. Like a paper library, we provide free access to researchers, historians, scholars, and the general public."*

Through its simple search facility, Wayback Machine, the Internet Archive provides access to 55 billion pages stored since 1996. The user can type in the web address and browse through the stored pages by date, with an easy access to other pages that have been collected around the same time. Keyword searching is not currently supported.

2.2.1 Evolving Issues

Many benefits of digital archives and libraries stem from the services that aggregate information from distributed repositories. For that reason, they have been investigating frameworks for interoperability, faithfully using metadata standards, and collaborating on various metadata initiatives.

As new types of information services were introduced via the Web, the community tried to comprehend the implications on their metadata creation practices. The paper by Duval et al. [6] provides good insights in the concerns and the breath of issues that have drawn their attention. Through a joint effort of the Dublin Core Metadata Initiative (DCMI) and the Institute for Electrical and Electronics Engineers (IEEE) Learning Object Metadata (LOM) Working Group they derived a set of 'principles and practicalities' for building useful and sustainable metadata systems. Among 'principles' they outlined the metadata modularity, extensibility, refinement, and multilingualism as essential for defining effective metadata systems. Under 'practicalities' they provide advices on practical decisions that one is faced with when implementing metadata schemas for a particular domain.

However, it became clear that the issues around practical use of metadata reached far beyond the specification guidelines. For illustration, we point out important points raised by McClelland et al. [9]. If one decides to merge metadata from a different resource to an existing digital library and finds inconsistencies, such as missing elements in the imported metadata, or incompatible field sizes, under what circumstances can the imported data be altered? What should be the formal mechanism to communicate that alterations have occurred? How should copyright information of the particular object be distinguished from the copyright for the associated metadata? Thus, besides the standard concerns, it is the copyright of the metadata itself that requires attention.

Digitization of the physical artefacts introduces further questions and complexities. If the artefacts, like newspapers, journals, manuscripts, and other publications are scanned and processed using OCR software, the representation of the content may require a range of components, from a simple scanned image and OCR text to multiple scanned images at different resolution levels, with corresponding descriptors

of the content layout and content analysis. The quality of data representation and the captured metadata will determine the types of applications and services that can utilize such data.

3 Preserving Digital Media

The immediate use and management of digital information often overshadows the importance of systematic planning that is required for a long term preservation of digital content. Without an appropriate strategy for preservation, today's digital information, unlike paper documents, will not be accessible in 50, 20, even 10 years from now. Ironically, the digital content is in danger of becoming a victim of its own success. Continuous technological advances that facilitate authoring and use of digital content introduce new formats and new storage media. The use of old formats and applications quickly fades away. *Unless we use adequate technologies and best practices to ensure that the past digital content is compatible with new information environments, we will loose access to the material created in the digital era.* This is a challenging issue with serious implications on the collective memory of our civilization. If not addressed, it also undermines the very strategy we chose to preserve information from physical artefact, i.e., the digitization of paper and media documents.

3.1 Problems and Initiatives

According to Bergman [1] and Lyman and Varian [8], the estimated value of digital documents that are produced in the EU and in danger of digital obsolescence is in excess of €3 billion per year. This is a tremendous cost to businesses and governments. Furthermore, from the timely intervention to save the content of the important and visionary Domesday Project [5] lead by the BBC back in the 1986, we know that such an effort can be quite costly. It is important to incorporate plans for long term use of the content early in its lifecycle, ideally right at the authoring time.

Organizations with legal responsibilities for safeguarding digital information, such as national archives and libraries, have been active in educating about the current state of the art on preservation issues and encouraging innovation in building tools and designing procedures. However, meeting the challenges of preservation goes beyond the capabilities of any single institution. For that reasons the EU Commission has committed resources within the Sixth Framework Programme to address issues of access and preservation of cultural and scientific resources. The objective is to promote collaboration among the libraries, archives, and research institutions who can tackle the problem from different perspectives and with complementary skills.

Similarly, in 2000 the US Library of Congress established National Digital Information Infrastructure and Preservation programs. This effort was further strengthen through partnership with the National Science Foundations which in 2004 launched research grant programs to address digital repository models, tools, technologies and processes, and organizational, economic, and policy issues of digital content preservation.

3.2 Industry Involvement

Preservation issues equally concern businesses, public sectors, and individuals. Joint efforts between libraries and industry have resulted in new insights and innovative approaches to addressing the problem of digital content preservation.

For example, in 2000 the National Library of the Netherlands (KB, Koninklijke Bibliotheek) and IBM started building an electronic deposit system, the Digital Information Archiving System (DIAS). . In order to address the problem of durable and large volume storage with long-term preservation requirements, they initiated a Long-term Preservation Study (LTP Study). The study resulted in 6 reports on important aspects, including the content preservation approach based on the concept of the Universal Virtual Computer (UVC) [7]. The method comprises storing the data and a specifically designed program that decodes and provides a logical view of the data. The logical data view can be used in an emulated UVC environment, running on a real machine of the given time.

An alternative approach to data preservation involves conversion of proprietary document formats into a widely adopted standard. In order to ensure long term preservation of Office documents, Microsoft (MS) Corporation produced a specification of the Office Open XML format [10] that defines an XML schema and its semantics for the MS Office applications. It retains high-level information suitable for editing documents or undergoing transformations using XSLT and other XML-based languages or tools. The Open Office XML format is in the process of approval for industry standard.

4 Summary

Digital media has opened new opportunities for creating, publishing, and communicating information. It has connected the contemporary information with valuable archived content from physical artefacts. It provides new ways of dissemination knowledge beyond traditional boundaries and thus has an unprecedented impact on all aspects of our lives. However, it brings with itself a challenge that must be addressed – the volatility of the digital formats and computing environments in which it can be used. Identifying methods and strategies for ensuring the long-term preservation of the digital format is of utmost importance for the survival of data and information created in the digital era.

References

1. Bergman, M.K.: Untapped Assets: The $3 Trillion Value of U.S. Enterprise Documents. BrightPlanet Corporation White Paper, July (2005) http://www.brightplanet.com/pdf/DocumentsValue.pdf

2. Commission of the European Communities: Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions. i2010: Digital Libraries, Brussels, 30. Sept. (2005) http://europa.eu.int/information_society/activities/digital_libraries/doc/communication/en_comm_digital_libraries.pdf

3. Commission of the European Communities: Communication Staff Working Document. Annex to the Communication from the Commission. i2010: Digital Libraries. Questions for online consultation.

Brussels, 30. Sept. (2005) http://europa.eu.int/information_society/activities/digital_libraries/doc/communication/annex2_en.pdf

4. Commission of the European Communities: Results online consultation 'i2010: digital libraries', Brussels, 2. March (2006) http://europa.eu.int/information_society/activities/digital_libraries/doc/communication/results_of_online_consultation_en.pdf

5. Darlington, J., Finney, A., Pearce, A.: Domesday Redux: The rescue of the BBC Doomsday Project videodiscs. Ariadne, Issue 36, July (2003) http://www.ariadne.ac.uk/issue36/tna/

6. Duval, E., Hodgins, W., Sutton, S., Weibel, S.L.: Metadata Principles and Practicalities. D-Lib Magazine, Vol. 8, Num. 4, April (2002)

7. Lorie, R., van Diessen, R.: UVC: A Universal Computer for Long-Term Preservation of Digital Information. RJ 10338, IBM Almaden Research Center, San Jose, CA (2005)

8. Lyman, P., Varian, H. R.: How Much Information. The Journal of Electronic Publishing. December, 2000 Vol. 6, Issue 2, December (2003) http://www.press.umich.edu/jep/06-02/lyman.html

9. McClelland, M., McArthur, D., Giersch, S., Geisler, G. Challenges for Service Providers when Importing Metadata in Digital Libraries. D-Lib Magazine, Vol. 8, Num. 4, April (2002)

10. Open Office Specification 1.0. Committee Draft 1, 22. March (2004) http://xml.coverpages.org/OpenOfficeSpecificationV10-CD.pdf