Preserving Historical Collections Using Multimedia Information Systems

Eduardo A. do Valle Jr.[1], Arnaldo de A. Araújo[1]

[1]NPDI/DCC – Universidade Federal de Minas Gerais
Caixa Postal 702 – CEP 30123-970 – Belo Horizonte – MG – Brazil

{edujr, arnaldo}@dcc.ufmg.br

Abstract. *We discuss the challenges of using multimedia information systems as a mean of preservation and access for collections of permanent value. We then describe the development of a multimedia information system that is being applied to historical collections. Our system will give access to a large collection of historic documents, in a variety of formats (manuscripts, maps, photos, audio samples, movies…), and also manage the workflow related to the digital assets in order to improve their long-term preservation. This work is being developed for* Arquivo Público Mineiro*, the state archive, in cooperation with* Universidade Federal de Minas Gerais *and* PRODEMGE*.*

1. Introduction

The preservation of document collections is a key activity for historical analysis and for the construction of the identity of the nations. However, this is a very challenging task: besides the difficulty in classifying and storing enormous documentary masses in a systematic way, the fragility of the artifacts imposes a severe compromise between conservation and access. To guarantee their preservation, valuable documents are kept in safe archives, available only to a few researchers. This, of course, is frustrating, since the protection of the documents at the cost of taking them away from the public hurts the goal of keeping memory alive. On the other hand, the direct and constant manipulation of the originals is an inevitable source of degradation. Digital technology appears as a mean to break this compromise, allowing ample access to high quality digital copies of the documents, and at the same time protecting the originals from unnecessary manipulation.

In addition, computerized systems have the potential of easing the organization and description of the collections. By using document and workflow management tools, it is possible to multiply the productivity of historians, archivists and conservation technicians. Well-conceived systems allow those professionals to automate and manage their complex activities in a rational way.

The collection consultants enjoy additional benefits, since the use of digital technology provides research facilities much more sophisticated and faster than those possible with paper-based finding aids, such as inventories and indices. The computer networks, in particular the *Internet*, with its *World Wide Web*, add the possibility of the remote access, expanding dramatically the universe of users of a collection.

However, though those benefits are very attractive, there are some challenges in the path of the application of information systems to historical data. Since the collections are highly valued, all the implications need to be addressed carefully and efficiently. The

consensus between archivists and other Information Sciences professionals is that inadequate technologies expose the collections to so many risks, that as to implement a naïve solution is preferable to abstain at all to implement any solution.

2. Precursors and Related Work

A prototype of a multimedia system for giving access to a sample of one of the fonds of the Arquivo Público Mineiro is described in [SPANGLER 98]. The success of this prototype stimulated us to pursue the current research.

There are already some available multimedia information systems in the *Internet*. The *Portinari Project* [LANZELOTTE 93] has a *catalogue raisonee* of the great Brazilian painter Cândido Portinari. The system allows a variety of research operations over the collection, providing search by theme, techniques, chronology, etc. [PORTINARI WEBSITE]. The *Joaquim Nabuco Project* [ROSA 94] was able to fulfill the development of an environment for acquisition, filtering, compression, consultation and impression of images that have both textual and iconographic value and currently hosts the collection of letters of the Brazilian politician and scholar Joaquim Nabuco [NABUCO WEBSITE].

All over the world, custodian institutions are taking advantage of the resources offered for computer sciences, in order to ease the research and increase the access to their collections. In the United States, the prestigious Library of the Congress [LOC WEBSITE] is at the same time a great sponsor and consumer of multimedia information systems applied to document collections. In Brazil, we can cite the pioneer work of the Biblioteca Nacional [BN WEBSITE], the Arquivo Nacional [AN WEBSITE] and the Aquivo Público do Estado do Paraná [APP WEBSITE].

3. The Challenge

3.1. Information Variety, Volume and Heterogeneity

Documents in a collection present a great diversity, which includes the difference of medias (texts, audio, video) and the variety of nature of the document. Correspondences, pamphlets and clippings, for example, even being all textual documents, are characterized by different metadata (a correspondence has sender and receiver, a clipping has source and so on). Although the universe of historical collection is guided by relatively strict international norms, that establish how the different document categories should be characterized, it is impossible to establish, a priori, all these categories. It is necessary to have a document architecture with at least some flexibility.

Another complicating factor is the heterogeneity of the description. The different subsets of the collections are described in different depths. Some subsets are detailed and completely inventoried, with complete and precise dates and indexes. Other subsets are so poorly described that the archivist may know as little as the general subject of a large box of documents as a whole. The system has to use to advantage all the available data, but it has to deal also with imprecision and lack of information in a graceful manner.

As the collection grows into tens of millions of items, the information system has to hinder its degeneration into an amorphous mass of badly classified and badly indexed data. When a database does not possess some semantic structure, its value for the users tends to diminish dramatically as it grows, since each search operation starts to return a

great amount of spurious answers, mixed to the useful information. In order to allow efficient research to the collection, it is necessary to describe the items and organize them into an arrangement. In archives, description is made through keywords taken from controlled vocabularies. The arrangement is set up grouping the documents into hierarchical groups that seek to preserve their original intellectual organization, in what is named *principle of respect des fonds* or *principe de provenance*. Research can be made not only by filtering by keywords or free-text, but also by the navigation through the arrangement. Actually, the two operations, filtering and navigation, need to be combined to compose a satisfactory consultation [GROSKY 94] .

3.2. Inherent Fragility of Digital Data and Technology Obsolescence

There is a consensus between archivists, that digital technology is still not safe for preservation purposes, in the long term. [CONWAY 96] People tend to believe that digital media have a life expectancy much longer than experience shows.

A recent study, by the American Commission for Preservation and Access showed the alarming conclusions that the best medias, in special storage conditions, have a life span of 20 to 30 years. In more ordinary conditions, the life expectancy can be as low as three to five years [VAN BOGART 95]. To this perishability, adds the special fragility of digital data in relation to sabotage, natural disasters and system failures.

But the greatest challenge is not the fragility of the storage artifacts, but the fast evolution of the technology. In contrast to a page of paper, that can directly be read, digital archives require sophisticated cooperating systems of hardware and software to become intelligible. Considering how quickly these systems are surpassed, it is possible to arrive at a scenario where a vast digital collection is lost because it is not possible to find the equipment and programs necessary to access it.

One has to accept that the digital data, without frequent intervention, does not survive. It is necessary to reevaluate the technological scenario every few years (five or ten), and to take measures to guarantee the survival of the collection in the new conditions, either through the migration of the data or the emulation of the old environments [BESSER 2000].

4. The Goals

In this work we intend to explore the benefits and challenges brought by the application of multimedia information systems to custodian institutions that have collections of permanent value. The systems that we will study do not limit themselves to make historical documents – manuscripts, pictures, movies, audio samples, etc. – available to the public. Instead, they also assist in the preservation of the digital assets, allow the creation of electronic and conventional finding aids and supply the archivist with a set of tools to ease and coordinate indexation, classification and reformatting efforts.

We are also implementing, in Arquivo Público Mineiro, an information system to give access, through the WWW, to a large collection of photographs, composed of tens of thousands of items, giving also support to other medias, and allowing the generation and distribution of finding aids.

4.1. Document Management

A multimedia system for historical collections is a special type of Document Management System (DMS). A DMS must provide functionalities such as classification, edition, storage, transmission and mechanisms of consultation to documents. Moreover, the system must capture the logical structure of the documents. This can be made through a catalogue of documents types, composing a hierarchy of generalization, much like the classes in the type system of an object-oriented language [FERREIRA 93].

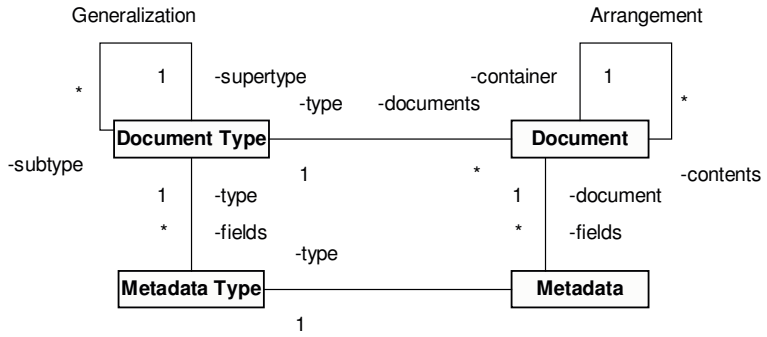

Figure 1. Document management support in our system.

The document architecture support in our system is explained in the UML diagram shown in Figure 1, where *Documents* and their *Metadata* (the attributes that characterize them) are related to *Document Types* and *Metadata Types*, two metaclasses that support the extensibility of the document architecture. The reflective association *Generalization* implements the hierarchy between *Document Types* while the reflective association *Arrangement* allows the nesting of documents, thus representing the arrangement of the collection.

4.2. Workflow Management

In addition to the core requirement of retrieving documents in the collection, it is necessary to have well-established procedures, with good support tools, for the addition of new registers. The documents of interest for historical databases do not exist originally in digital form, forcing the user to digitize from the original, analogical, artifacts. This reformatting, for valuable documents, is extremely complex and costly, demanding a degree of planning, operational care and quality control much above those commonly applied to the processing of ordinary documents, in commercial institutions. Without tools to support the archivist, often the task can become infeasible.

In addition to the digitization, there are many other necessary activities: including acquisition of metadata, generation of controlled vocabularies, indexation of documents using the vocabularies, classification of the collections in hierarchic sets, creation of finding aids, microfilming and restoration of the collections, etc.

In order to coordinate all those activities, the information system must provide workflow management facilities. Workflow management tools allow detailed specification of the stages that compose the processes, control and coordination of the tasks execution, and monitoring of the results. The most sophisticated ones even provide special methods and environments for process modeling and reengineering [CICHOCKI 98].

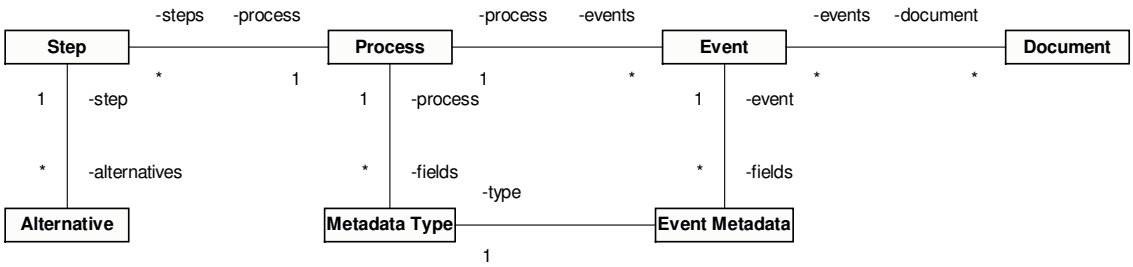

Figure 2. Workflow management support in our system.

Our system support to workflow management is depicted in Figure 2. There are facilities to represent *Processes*, decomposing them into logical *Steps* and considering every *Alternative* a *Step* can lead to. The enacting of a process generates an *Event*, often related to *Documents*. A set of *Event Metadata* is stored, for managerial purposes.

4.3. Access

In the universe of archives and libraries, access is always an essential question. A great deal of the working time of the archivist is expended in the preparation of finding aids, i.e., references such as indices and inventories that help the consultant to find the desired information in a large collection. A finding aid is a much more sophisticated access guide than a simple listing or a free-text search: elaborated normally by researchers who know the collections deeply, they portray the intellectual organization of the documents and make the consultants work much easier.

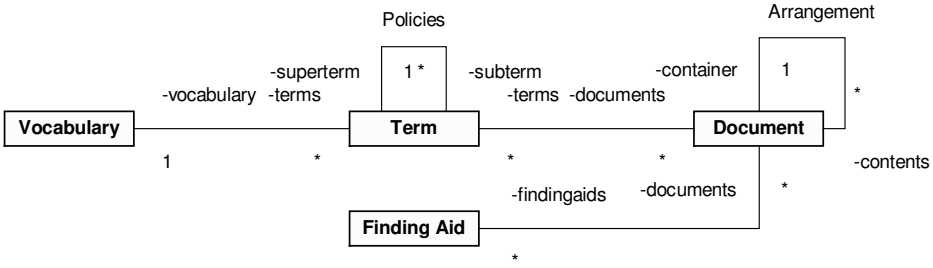

Figure 3. Indexing, arrangement and finding aids.

Figure 3 shows the mechanisms of access provided by our system. *Documents* are indexed by *Terms* taken from *Vocabularies*. The user can use filtering operations asking for documents that are related to certain terms. Since the *Arrangement* reflects the original intellectual organization of the documents, it helps also in the research, in our system it is used mainly for navigation purposes. If special *Finding Aids* are available for a subset of the collection, the user can take advantage of them too.

5. Our Approach

5.1. Activity Planning

Our project has four main lines of action: the multimedia application development, the treatment of the collection, the installation of the support infrastructure and the creation of politics for technological management in Arquivo Público Mineiro. Our plan explored the opportunities of parallelism between these activities, and can be seen in the diagram of Figure 4.

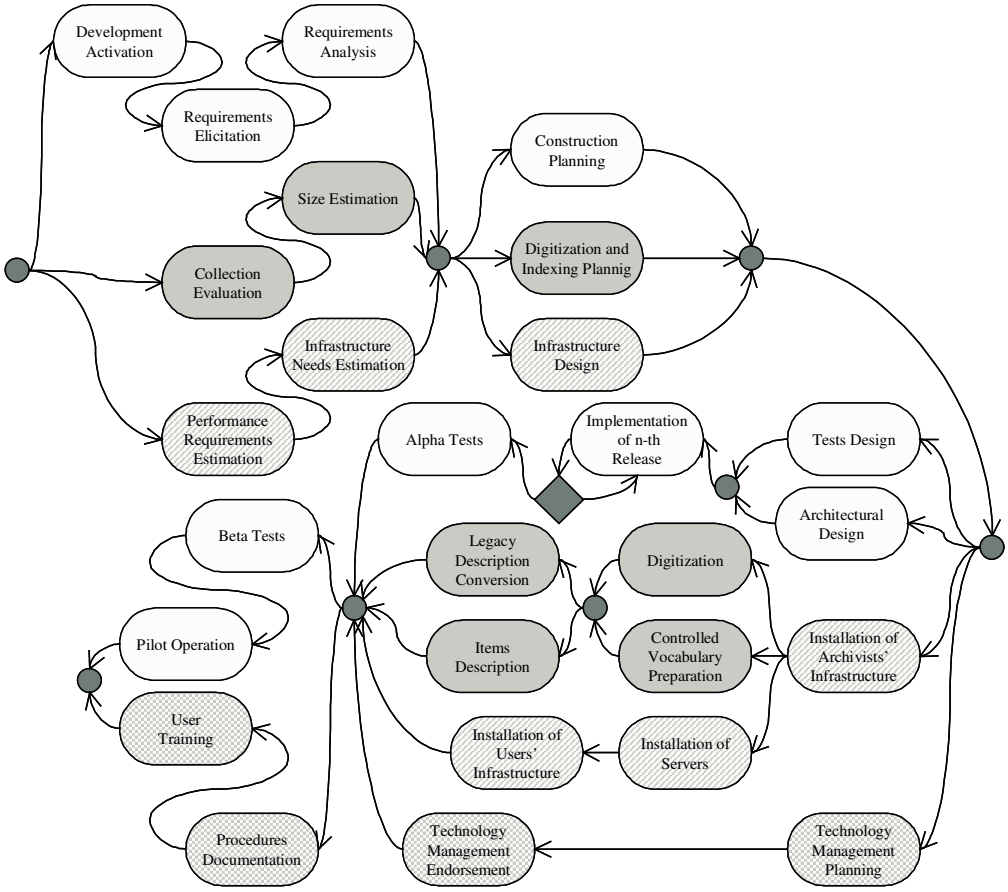

Figure 4. Project activities. Each track is represented in a different pattern, and the small circles represent synchronization points

The system construction followed the stages prescribed by PRAXIS, a simple development methodology described in [PAULA 2001].

Perhaps more complex than the construction of the system itself is the preparation of the data that will be fed to it. In the first stage of the project it is planned the digitization of more than 14,000 pictures, of different sizes and formats, some of them more than 100 years old. The pictures also must be indexed and classified. During the description process, the controlled vocabulary must be maintained.

In order to provide the functionalities required with acceptable performance, a platform of robust hardware and software is necessary. It is also necessary to acquire, install and maintain scanners and workstations for the archivists.

5.2. Development Methodology – Object Orientation

Object-orientation is a great ally in the development of multimedia information systems. It is much easier to model and design Document and Workflow Management Systems using the new object paradigms than the old structured analysis of the 70'. The raise in interest for document management using multimedia information systems has stimulated many researchers to develop methodologies and models for representing this kind of information and guiding the development of project. This is consistent with a general raise in interest on modeling and software engineering practices.

An approach to the modeling of Web applications (multimedia or not) can be found in [CONALLEN 99]. [PAULA 2000] adapted PRAXIS specifically for multimedia purposes. A framework to support the development of hypermedia applications, which was applied to the Portinari Project, is described in [SCHWABE 98].

Interestingly, all the above methods are based on *Unified Modeling Language* (UML), today's *de facto* standard for software engineering, which came as the result of the convergence of many object-oriented methodologies. UML enforces the reduction of the semantic gap between analysis, design and implementation of software, providing visual representations for the system through models that capture all the structural and behavioral aspects of the software [BOOCH 99].

5.3. Extensive Use of Metadata

Metadata is information concerning a document, in addition to its contents, like title, authorship, dates, administrative history, donor, original size, scanning resolution, etc... Metadata are an invaluable for the management of a digital collection, helping to solve matters of workflow (when a document was scanned, by who, etc.), of access (keyword indexation, arrangement, etc.) and preservation (structure of the digital file format, media longevity, etc.). In collections that have navigation structure (hypermedia), the hyperlinks are sometimes considered portion of the metadata.

In a DMS, many metadata are determined by the document architecture, where they appear as the attributes of document types. Today, many consortiums are trying to standardize the metadata for hypermedia applications. [W3C 99] [DCMI 2001]

6. Conclusion

Including imaging methods for enhancing and restoration, new paradigms for complex databases, and facilities for retrieval and distribution of the information, digital technology brings new and exciting possibilities to the universe of collections preservation. However, its application in devices of permanent value must be led carefully, in order to avoid the issues arisen by the fragility of digital data. The extensive use of metadata and the application of special development methodologies are of great help in order to build a successful application.

7. Acknowledgements

The authors wish to thank CNPq, CAPES and the SIAM Project DCC/PRONEX for supporting this work. Mr. do Valle is sponsored by a scholarship provided by CNPq.

References

AN WEBSITE. Website of Brazilian National Archive – Arquivo Nacional, June/2002 at http://www.arquivonacional.gov.br/

APP WEBSITE. Website of Public Archive of the State of Paraná, June/2002 at http://www.pr.gov.br/arquivopublico/

BESSER, H. (2000) "Digital Longevity" in Handbook for Digital Projects: A Management Tool for Preservation and Access. Northeast Document Conservation Center.

BN WEBSITE. Website of Brazilian National Library, June/2002 at http://www.b n.br/

BOOCH, G, RUMBAUGH, J. and JACOBSON, I. The Unified Modelling Language - User Guide. Reading, Addison-Wesley Longman Inc., 1999.

CICHOCKI, A. et al. Workflow and Process Automation: Concepts and Technology. Kluwer Academic Publishers, Boston, 1998.

CONALLEN, J. (1999) "Modeling Web Applications with UML", June/2002 at http://www.conallen.com/whitepapers/webapps/ModelingWebApplications.htm

CONWAY, P. (1996) Preservation in The Digital World. Commission on Preservation and Access, Washington.

DCMI (2001). Dublin Core Metadata Initiative. "Using Dublin Core", June/2002 at http://www.dublincore.org/documents/2001/04/12/usageguide/

FERREIRA, B. (1993) Projeto e Implementação de Um Sistema de Gerência de Documentos Segundo o Paradigma de Objetos. M.Sc. Dissertation. Universidade Federal do Rio Grande do Sul, Brazil.

GROSKY, W. (1994) "Multimedia Information Systems". In: IEEE Multimidia, Spring 1994, p. 12-24.

LANZELOTTE, R., MARQUES, M. et al. (1993) "The Portinari Project - Science and Art Team Up Together to Help Cultural Projects", In: Proceedings of the Second International Conference on Hypermedia and Interactivity in Museums, England.

LINS, R. and FRANÇA NETO (1995), L. "Projeto Nabuco Processamento de Imagens de Documentos Históricos", In: Anais da XXI Conferencia Latino-Americana de Informática - PANEL'95, Brazil.

LOC WEBSITE. Website of American Library of Congress, June/2002. http://www.loc.gov/

NABUCO WEBSITE. Website of the Joaquim Nabuco Project, June/2002. http://www.cin.ufpe.br/~nabuco

PAULA Filho, W. Engenharia de Software: Fundamentos Métodos e Padrões. Livros Técnicos e Científicos Editora, Brazil, 2001.

PAULA Filho, W. Multimídia: Conceitos e Aplicações. Livros Técnicos e Científicos Editora, Brazil, 2000.

PORTINARI WEBSITE. Website of the Portinari Project, June/2002. http://www.portinari.org.br/

ROSA, L. and LINS, R. (1994) Nabuco: Uma Base de Dados para Documentos Históricos. MSc. Dissertation, Universidade Federal de Pernambuco, Brazil.

SCHWABE, D. and ROSSI, G. (1998) "An Object Oriented Approach to Web-Based Application Design", Theory and Practice of Object Systems 4. Wiley and Sons, NY.

SPANGLER, N. (1998) Sistema de Informações Multimídia. M.Sc. Dissertation. Belo Horizonte, Fundação João Pinheiro e Universidade Federal de Minas Gerais.

VAN BOGART, J. (1995) Magnetic tape storage and handling: a guide for libraries and archives. Washington D.C., Comission on Preservation and Access.

W3C (1999). World Wide Web Consortium. "Resource Description Framework (RDF) Model and Syntax Specification". http://www.w3.org/TR/REC-rdf-syntax