A Study on the Meta-data Design for Long-term Digital Multimedia Preservation

Feng-Cheng Chang[1], Chin-Yuan Chang[1], and Hsueh-Ming Hang[1,2]
[1]Dept. of Electronics Engineering, National Chiao Tung University, Hsinchu, Taiwan
[2]Dept. of Computer Sci. and Inform. Technology, National Taipei University of Technology, Taipei, Taiwan
breeze@alumni.nctu.edu.tw, ralohcs@gamil.com, hmhang@ntut.edu.tw

Abstract

Due to the fast growth of multimedia content, the digital preservation emerges as an important technology that prolongs the life of digital objects. For audio-visual contents, we face the obsolescence problem in both supporting hardware/software and data representation formats. Migration strategy is a solution for achieving high-quality archival and easy access. In this paper, we investigate the concepts and structures in the Open Archival Information System (OAIS) and the MPEG-21 Digital Item Adaptation (DIA) specifications. By addressing the issues for a long-term audio-visual preservation system, we propose a combined meta-data scheme to record both the management and technical procedures and parameters. The detailed meta-data that keep the history of evolution versions can be used for subsequent migrations. We also implement a prototype system to realize our design. At the end of this study, we identify additional issues for future improvements.

1. Introduction

With the advances in digital devices, the number of newly created digital contents grows drastically nowadays. It is known that digital contents have many advantages over the analog ones, such as being resilient to environmental change and being easily processed by complicated algorithms. A number of projects were initiated to convert the analog contents into digital forms for easy access and preservation. For example, books and articles are digitized, and the electronic versions are distributed over the Internet.

One of the most complicated tasks is to build a digital library for archiving, accessing, distributing, and indexing. For text-based contents, there are well developed methods to maintain such kind of databases. However, the audio-visual contents are not only difficult to organize in a database, but also difficult to analyze their semantics. Even we can neglect these issues, there are the obsolescence issue in preservation. The importance of the latter is reflected by the large number of tutorials [1], organizations [2, 3], and standards [5, 6, 7, 8, 9, 10] that are dedicated to digital preservation researches.

In this paper, we first summarize the general digital preservation methods in Sec. 2. Based on these methods, the audio-visual specific preservation issues are also discussed. We then describe the MPEG-21 Digital Item Adaptation (DIA) [4] in Sec. 3. By combining these concepts, we propose a meta-data scheme for audio-visual contents in Sec. 4, which is suitable for long-term preservation. Then, we implement a prototype system to varify our design in Sec. 5. At the end, we conclude our work in Sec. 6.

2. Digital Audio-Visual Content Preservation

Due to the demand of digital information archives, a lot of efforts have been spent on standardizing meta-data schema and system architectures. Complementary to the industrial standards are the research activities on the digital preservation technology. Firstly, we describe the issues that a digital preservation system may have in Sec. 2.1. Secondly, we describe several strategies for preserving digital contents in Sec. 2.2. Thirdly, we discuss the issues specific to audio-visual preservation in Sec. 2.3.

2.1. Digital Preservation Issues

The digital preservation technology has been developed for several years. In the past decade, due to the limitation of the hardware and software and the small number of digitally preserved files at the beginning, its applications are not widely noticed. However, the maturity in the supporting devices and algorithms makes it a much more practical technology. Also, the drastically increasing amount of digital achieve files put it in a high demand.

In a digital archive system, several factors contribute to the preservation problems [1]. The two major types of problems are obsolescence and physical damage. The first type can be further divided into two sub-types. The file format

978-0-7695-3278-3/08 $25.00 © 2008 IEEE
DOI 10.1109/IIH-MSP.2008.261

and software obsolescence are due to the evolution of specifications and software versions. The format specification defines the layout of a digital object, and the associated software pieces define the run-time processing of the format. Their tightly coupled relationship implies that the evolution failure in either one would put the digital object into a useless state. A general guideline is to archive in non-proprietary formats and open specifications, in order to reduce the business/marketing effects. However, the problem still remains, especially when the format obsolescence issue is examined in a long-term scale. The other obsolescence is caused by the hardware and storage media. When a specific storage media phases out or the processing hardware is out-of-date, the cost of keeping the data accessible goes up.

The second type of preservation problem is due to the physical damage of the contents; examples are hardware worn-out, natural disasters, human improper handling, and etc. To reduce these problems, people define guidelines such as the rules regulating the handling of the media, specifying the preservation environment, backing up high priority data, and re-recording the data according to the media life.

2.2. Digital Preservation Methods

To reduce the impacts of the issues mentioned in the previous section, many preservation strategies have been proposed. Here, we briefly summarize some frequently used methods that reduce software and format obsolescence problems.

Technology preservation is to preserve the necessary environment for future access; both the hardware and the software are to be preserved. It extends the life of digital objects. However, the obsolescence of physical objects makes this approach stop working eventually.

Migration is to transfer the digital content from one technology to another. It preserves the contents by converting its format from the old one to the new one. This method can effectively overcome the obsolescence problems. The disadvantage is the cost and the complexity. Some people also criticize that neither authenticity nor integrity of a migrated digital object can be ensured.

Reliance on standards is a method to stay on a well-specified format to reduce the obsolescence problem. It assumes that a well-known and widely adopted standard would have better support when evolution occurs. The baseline is that the format is open and the software/hardware can be re-implemented as necessary. This approach is not reliable, since technology advances very fast in time.

Normalization is a variation of the previous one. It converts different kinds of contents to the corresponding internal standard formats.

Emulation is the method to reproduce the original behavior in the new environment. It usually needs both the software and the hardware technologies to emulate the old environment. It reduces the impact of obsolescence by trading with the costly long-term maintenance of complex emulators.

Encapsulation is to group all necessary meta-data with the digital object. The concept is to provide a self-contained package for accessing the content. It assumes that the meta-data for the object are so well-designed that no critical information for interpreting the object is missing. In the extreme case, the meta-data become an emulator described in the previous item.

2.3. Long-term Audio-visual Preservation

In this paper, we are interested in the long-term preservation for especially audio-visual contents. Several important facts need to be considered. The first noticeable fact is that audio-visual data size is often very large, and thus it is necessary to store them in a compressed format. Lossless compression can be used to preserve the original quality, but the compression ratio is low. In contrast, the lossy compression algorithms achieve high compression efficiency at the expense of sacrificing a small amount of quality loss. Either the lossless or the lossy compression techniques suffer from the obsolescence issues discussed earlier.

For a typical multimedia archive system, there is an additional requirement: the audio-visual contents should be also easily accessed, such as browsing and distribution. To satisfy this requirement, an up-to-date (current) file format is preferred, because it is most popular format used at the current time and thus is highly interchangeable between different systems. Therefore, after investigating the advantages and disadvantages of various long-term preserving strategies, we find that the migration strategy seems to be the most proper solution. The content format is updated whenever a new and well recognized format appears. Although the quality would degrade slightly between each migration, we assume that a newer (and superior) compression algorithm could minimize the quality loss with the help of some additional meta-data.

3. MPEG-21 Digital Item Adaptation

For audio-visual data, the migration process from one format to another is usually known as transcoding. A similar concept, though not entirely defined for format conversion purpose, is the MPEG-21 Digital Item Adaptation (DIA) specifications [4]. An application example of DIA is scalable data distribution. According to the bandwidth or the playback capability, the scalable coded bit-stream can be adjusted to match the target. For example, a sender can re-organize the bit stream to form a low-quality one; or a

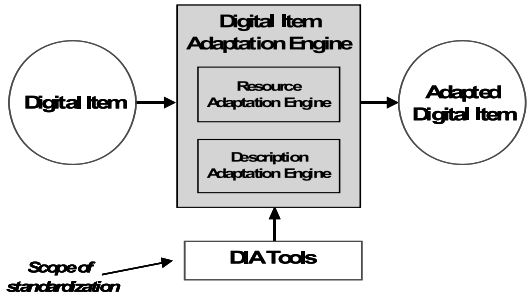

Figure 1. The concepts of DIA [4].

receiver can select the needed low-quality segments to decode.

As shown in Fig. 1, the bit stream adaptation can happen in two distinct processing paths: the data object path and the meta-data path. The resource adaptation transforms the data to the specified target format; and the description adaptation generates the meta-data that contain the adapted resource properties. The MPEG-21 DIA specifications do not specify the adaptation engines, because they are implementation dependent. Instead, the DIA standardizes the descriptions and format-independent mechanisms as the DIA Tools. There are eight categories of DIA tools, including (1) Usage Environment Description Tools, (2) BSDLink, (3) Bitstream Syntax Description tools, (4) Terminal and Network Quality of Service, (5) Universal Constraints Description Tools, (6) Metadata Adaptability, (7) Session Mobility, and (8) DIA Configuration Tools. Since the DIA Tools are used to specify an adaptation process, it can be used to specify the parameters/behavior of a given transcoder.

4. Meta-data Design

In a preservation system, the meta-data are as important as the data. A well known set of meta-data specifications is proposed by the Open Archival Information System (OAIS). It covers most of the usages in a digital repository [8, 9]. The OAIS meta-data are designed from the system management viewpoint. on the other hand, for audio-visual transcoding purpose, some technical details are important and should be kept in the meta-data although they may not be useful for management purposes. For example, the transcoding parameters (sampling rate, dynamic range, precision, etc.) are not meaningful for data query sessions, but they may be important for subsequent migration sessions to produce optimal results.

The OAIS specifications do not have detailed technical definitions for audio-visual contents. Because the MPEG-21 DIA concepts cover the format transcoding, format descriptions, and transcoder descriptions, they can be included to complement the OAIS. Thus, we extend the OAIS specifications with the DIA concepts to design a migration-based

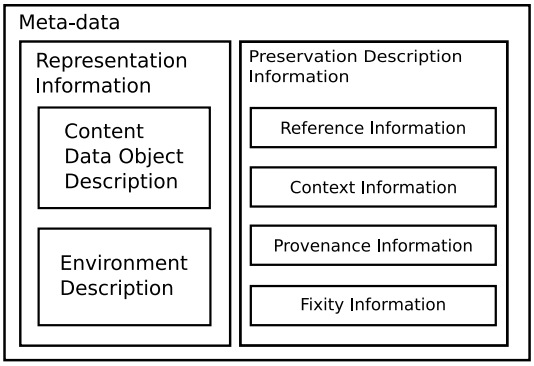

Figure 2. Meta-data for audio-visual objects.

audio-visual digital preservation system.

Figure 2 shows our proposed basic meta-data scheme for audio-visual objects. It comprises two major parts. The first part, the *Representation Information*, contains the *Content Data Object Description* and the *Environment Description*. The former specifies the static properties of the object, such as the content type and the file format. The latter specifies the environment used to playback or to render the content. The second part is the *Preservation Description Information (PDI)*, and it is also the focus of our design. The *Reference Information* records the identifiers of the object. The *Context Information* specifies the creation context and the relationships to the other objects. The *Fixity Information* records the signature for authentication and integrity check.

The *Provenance Information* is used to record the history of the content, including: (1) **Origin** records how the content object was created; (2) **Pre-Ingest** records the history before it was included in the preservation system; (3) **Ingest** records the procedures applied during the inclusion of the data object; (4) **Evolution History** records the history of the manipulation of the object; and (5) **Rights Management** records the details about the usage restrictions.

The evolution history is directly related to the migration sessions. Each Evolution History is a list of Evolution Records. Each Evolution Record contains the transcoding-related meta-data. Hence, we incorporate some of the DIA attributes into a record:

Reason of Evolution is used to specify the reasons that trigger the transcoding.

Evolution Versions is used to specify the version identifier, and the adjacent versions. The evolution history is arranged in the order of the version traversal.

Evolution Evaluation is designed to record the quality change from the previous version to this version.

Evolution Engine is used to record the transcoder meta-data, including the transcoding parameters.

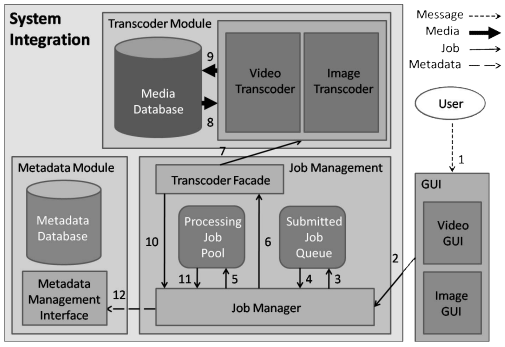

Figure 3. Prototype system structure.

5. Prototype System and Discussions

To verify our design, we construct a prototype system as shown in Fig. 3. It consists of the transcoding module, the meta-data database module, and the migration job management module. A user can set up a migration job for a given audio-visual object through the GUI. Then the job management schedules the tasks and controls the work flow. After the job is completed, the meta-data are written into the database.

In the implementation, we integrate two transcoders from the other two sub-projects, which are parts of this integrated project. One is an MPEG-2 to MPEG-4 video transcoder, and the other is a configurable still image transcoder. Even though we have investigated the required meta-data structure and the related transcoder technology, several problems appear in the process of constructing a practical digital preservation system. These problems are described below.

- We assume that the information about the transcoding is available. Sometimes it takes a lot of efforts to obtain the necessary information. For example, to obtain the Evolution Evaluation attribute, either the transcoder or a separate tool is required to compute the quality change.
- The migration strategy is criticized in its non-confidence of authenticity and integrity. When a migration action is necessary and a re-migration (under fine-tuned control) is in place, the updated meta-data would cause additional complexity for the meta-data management.
- For long-term audio-visual preservation, migration is not frequent because we should carefully choose a widely adopted and flawless format for the next evolution. Once a migration is decided, all the old-format data in an achieve disposal have to be transcoded. In this situation, not only the transcoding process should be efficient, but also the updating system should be sufficiently robust and stable.

6. Conclusions

In this paper, we studied the long-term preservation problem and solutions for digital audio-visual objects. After investigating the conventional digital library projects and the related standards, we learned the strategies and the meta-data structures for long-term preservation purpose. For audio-visual objects, migration is a good strategy as long as widely adopted and flawless standard formats of each generation are carefully chosen. It reduces the impact of obsolescence, and maintains the content under a popular format for easy access and distribution.

Technical meta-data for migration are as important as the management meta-data. Therefore, we proposed the meta-data scheme by extending the current OAIS meta-data with the MPEG-21 DIA specifications. The management part is covered by OAIS definitions, while the technical part is a variation and extension of the DIA definitions.

We also implemented a prototype system to verify our design. During the integration, we identified several practical issues that are not covered in the previous meta-data study and the transcoder study. These issues can be the topics for future study..

7. Acknowledgements

This work was partially supported by the NSC, Taiwan under Grants NSC 96-2422-H-007-003.

References

[1] Digital preservation management: Implementing short-term strategies for long-term problems. http://www.library.cornell.edu/iris/tutorial/dpm/eng_index.html.
[2] Online computer library center. http://www.oclc.org/.
[3] Research libraries group. http://www.rlg.org/.
[4] *ISO/IEC JTC1/SC29/WG11 N5845, Text of ISO/IEC 21000-7 FCD Part 7: Digital Item Adaptation*, Trondheim, Norway, July 2003.
[5] *NDIIPP Technical Architecture Version 0.2*, 2004.
[6] *Metadata Encoding and Transmission Standard (METS)*, 2007.
[7] *PREMIS (PREservation Metadata: Implementation Strategies) version 2.0*, March 2008.
[8] The Consultative Committee for Space Data Systems. *Reference Model for an Open Archival Information System (OAIS)*, 2001.
[9] The OCLC/RLG Working Group on Preservation Metadata. *Preservation Metadata and the OAIS Information Model: A Metadata Framework to Support the Preservation of Digital Objects*, June 2002.
[10] RLG Inc. *Trusted Digital Repositories: Attributes and Responsibilities*, Mountain View, CA, May 2002.