CHANNELS TO THE FUTURE

Gábor Magyar
Dept. of Telecommunications and Media Informatics, Budapest University of Technology and Economics
Budapest, Hungary

Keywords: Archiving documents, originality, metadata, Dublin Core, semantics.

Abstract: The long-term archiving of digital documents is a very challenging task, because of policy, legal, intellectual property rights, metadata, semantic support and other issues. This paper merges technical and sociotechnical approaches. As more research disciplines and societal sectors have come to rely on data-driven models and observational data, the archiving problem is growing, the shortcomings of current technologies have become apparent and the need to preserve historical material has become imperative. The variety and complexity of digital documents as information technology objects brings up a basic question: does it necessary to preserve the variety and complexity of the original objects? Our answer in general is 'no', essential attributes of a document are preserved when the document is transformed to different platforms. There are many reasons to change the format of a document. We use the categories of physical, logical, and conceptual layers in order to define generic properties that are true of all digital documents. This approach gives an overall framework for general preserving strategy managing technical obsolescence and semantic mutations.

1 DIGITAL DOCUMENTS

There is no a single conventional definition of the category „digital document". (Buckland, 1998)

In the „PC world" this term was originally used for a file created with a word processor. I.e. the term referred to a textual object. This conception has totally changed: documents can contain graphics, charts, and other objects (images, embedded sounds, animations, videos). A digital document can have an exactly described structure of its elements – this structure itself can have extra meaning. A word processing application can produce graphics, tables, hyperlinks, XML output and a graphics application can produce words. This trend has accelerated with technologies that allow an application to combine many components. What's more, many forms of digital information cannot be expressed in traditional hard-copy or analog media; for example, interactive Web pages, geographic information systems, and virtual reality models. Consequently, the term 'document' is used more and more to describe any file produced by an application. There is no single definition or model of a digital document that would be valid in all cases. Information technologists model digital documents in very different ways: a digital document can be a sequence of expressions in natural language characters or a sequence of scanned page images, a directed graph whose nodes are pages (what appears in a Web page), and so on. How documents are managed, and therefore how they are preserved, depend on the model that is applied.

Unlike a paper record where the symbols which encode the information are directly accessible to a reader, digital data is stored as a series of bits on a storage medium in a form which is intelligible (frequently not visible) to the human being. So a machine is required to retrieve the binary patterns from the device, and then a program is required to interpret the encoding format of the bit stream and render it in a human-intelligible form. The information is interpreted as the result of the action of these human made hardware/software tools on the data. The precise form in which the information is made available to a user depends as much on the action of these technological intermediaries as it does on the data on which they operate.

Two different software tools may render the same data in ways that give the user different views of that data as information. This may be true even for the same tool running in different environments or with different parameters. Typical word

Magyar G. (2007).
CHANNELS TO THE FUTURE.
In *Proceedings of the Ninth International Conference on Enterprise Information Systems*, pages 369-374

processing programmes incorporate multiple 'view' options which present quite different screen renditions of the same data. The user's experience of information becomes a complex product of the base data itself and the processes performed on that data. Extreme, but simple example is the difference between the views of the same textual content made by the same word processing applications but using different file format outputs (e.g. Rich Text Format and simple text).

This process of mediation also means that the user don't know the physical structure of a digital record (the way in which digital units of a representation are physically arranged on the storage medium). This will certainly change when a single representation is transferred from one medium to another, sometimes even on transfer from one instance to another of the same medium.

2 ORIGINALITY OF A DOCUMENT

What is meant by 'original' in the space of digital documents?

Almost any 'use' of a selected representation of a digital record involves the making of a copy in some way. If a user accesses a representation which is held on a networked server, a copy of that representation is transmitted to their client computer, and indeed multiple users could perform the same action simultaneously on that same representational form. There are files in the digital world that have the same properties. (The problem of versions, modifications are out of the scope of this paper. Here and now different versions are identified as different documents. Version management is a practical issue from this point of view. However, the three-layer model, described in this paper can handle this issue.)

The variety and complexity of digital documents as IT objects brings up a basic question: does it necessary to preserve the variety and complexity of the original objects?

The answer in general 'no', essential attributes of a document are preserved when the document is transformed to different platforms. There are many reasons to change the format of a document, crossing technological boundaries (eg. platforms, operating systems, applications). For example, to ensure that written documents retain their original appearance, authors translate them from the word processing format in which they were created to Adobe's PDF format.

This paper gives a generic model of properties of digital documents. We use the categories of physical, logical, and conceptual layers in order to define generic properties that are true of all digital documents.

3 PHYSICAL LAYER

At the physical layer, a digital document is an inscription of signs on a medium. There are specific (coding) rules in any technical environment to determine the matching between a system of signs and the physical storage of bits. Those conventions vary with the type of the physical medium, media types and other factors. The physical layer of the model deals with physical files. The physical inscription of bits is independent of the meaning of the inscribed bits.

It is well known, that today's digital media solutions are not durable over long periods of time. The digital storage media degrades relatively quickly, when compared with the known durability of paper. This problem can be addressed through copying digital information to new media. Media refreshment or migration adds, of course a new cost element of digital preservation to the whole life cycle. However, this does not mean the continuous growth of the total costs, because digital storage effectiveness, especially recording density increases while costs decrease. Repeated copying of digital data to new media over time reduces per-unit costs. Based on the traditional rate of the growth of storage densities media migration yields a net reduction in operational costs. (Moore, 2000) In this context, the durability of the medium is only one variable in the cost equation: the medium needs to be reliable only for the length of time that it is economically advantageous to keep the data on it.

The physical preservation strategy must also include a reliable method for maintaining data integrity in storage and in any change to storage, including any updating of the storage system, moving data from inactive storage to a server or from a server to a client system, moving data between the elements of a distributed storage architecture or delivering information to a customer via the Internet, as well as in any media migration.

Physical preservation is necessary, but not sufficient in the archiving process.

4 LOGICAL LAYER

A digital document can be recognized as a logical object (or a specific set of logical objects) according to the logic of certain application software. The conventions of composing logical objects are independent of how the data are written on a physical medium. As we described before, at the storage level the interpretation of the bits is not defined. Similarly, at the logical layer the grammar is independent of physical inscription. Once data are read into memory, the type of medium and the way the data were inscribed on the medium are of no consequence. The rules that apply at the logical layer determine how information is encoded in bytes and how different encodings are translated to other formats; how the input stream is transformed into the system's memory and output for presentation.

The technologies of storage media and machine tools which read those media and the encoding formats. The programs which interpret those formats are permanently changing, driven by scientific advances, user requirements for improved cost effectiveness, and the commercial imperative of the suppliers. A file created using Microsoft Word, and stored in Word format can only be read by another version of that same program, or by a program which is able to encode the Word format. If the file is transferred to an environment where such a tool is not available, or if it is not accessed over a period of time and during that time the program becomes obsolete, then the information content of that file becomes inaccessible. This constant change is certain to continue: there is no evidence to suggest that a plateau of technological stability will ever be attained.

A logical object is a unit recognized by some application software. To preserve digital information as logical objects, we have to know the requirements for correct processing of each object's data type and what software can perform correct processing.

5 CONCEPTUAL LAYER

At the conceptual layer we see document as they are handled in the real world: documents are meaningful objects, such as books, reports, proceedings, photographs, contracts, maps - but in the digital space a document can be a mix of different media types.

The properties of the documents at the conceptual layer are those that are significant in the real world. A book has author(s), title, etc. A report has an author, a title, an intended audience, and a defined subject and scope. A proceeding has editor, authors, titles, etc. These properties of documents have meaning to human beings. Data elements of documents may have structure. Actually all meaningful textual documents have structure – words compose sentences, there are paragraphs, chapters, etc. Sometimes this structure is (at least partially) pre-defined: a good example is e-mail what has header (including 'to', 'from', 'subject', etc. fields) and (unstructured) message body. Web pages have links as structural elements. The information content of a table can not be recognized without the structure of this table. An Excel file is a set of tables, having cross-references between tables. This structured set of data can be seen as a database.

Some of the properties are tagged and stored as metadata and metadata elements are organized into data-schema(s). Advanced word processing applications do have metadata management facility.

Metadata often serves as integration platform. Formalized metadata is on the rise, leading to significantly better data management and exploitation capabilities. Metadata will make it much easier for machines to process data automatically, and it is exactly this capability that can drive many other benefits: interoperability, cost-cutting, better data quality, transparency, better decision support and new business opportunities. Metadata standardization is key to interoperability. Dublin Core Metadata for Resource Discovery seems to become the most relevant candidate for common metadata platform of different media-types. (DC, 1998)

Metadata tags will also aid search engines and content processing intelligence software. They can carry additional information that can be used as additional hooks for searching, whether based on keywords or taxonomy. And the tags can facilitate better categorization and predictive analysis. Google's Froogle for example requires catalog suppliers to tag their content according to the W3C's RDF. (RDF, 2006)

There are many problematic issues concerning metadata. First, metadata tends to be scattered and there are often conflicting approaches for describing the same things. Second, creating coherent metadata can be difficult and expensive. And third, it is not easy to interpret different proprietary metadata schemes in various application modules. For metadata to become more cost-effective, it must be shared and reused – not only within one organization but within the whole circle of co-operating institutions.

The content and structure of conceptual document properties must be contained somehow in the logical document(s) that represent that document in digital form. However, the same conceptual content can be represented in very different digital encodings, and the conceptual structure may differ substantially from the structure of the logical document. The content of a document, for example, may be encoded digitally as a page image or in a character-oriented word processing document. There are different metadata schemas. The conceptual structure of a report - e.g., title, author, date, and introduction - may be reflected only in digital codes indicating differences in presentation features such as type size or underscoring, or they could be matched by markup tags that correspond to each of these elements.

Can we state that one of the possible digital formats (Microsoft Word, Adobe PDF, WordPerfect, HTML, a scanned image, etc.) is the true or correct logical representation of the document? As the archivist's ultimate aim is to preserve the document exactly as it was created the most basic criterion is whether the document that is produced when the digital file is processed by the right software is identical to the original. In fact, each of these encodings, when processed by software that recognizes its data type, will display or print the document in the format in which it was created. So if the requirement is to maintain the content, the structure, and the appearance of the original document, either digital format is suitable.

Since we have a variety of digital formats that are equally suitable for preserving the conceptual object(s), this rule can be extended to more complex types of documents, including databases and electronic transactions as well, where the documents are not necessarily presented to human beings but are found only at the interface of two computer applications.

6 THE RICH RELATIONS OF THE THREE LAYERS

The complex nature of a digital document having distinct physical, logical, and conceptual properties gives rise to considerations for digital preservation.

To preserve a digital document, the relationships between layers must be known or knowable. To retrieve a paper archived as master and subdocuments, we should know that it is stored in this way and we must know the identities of all the logical components. To retrieve a specific certification of examination results for a student, you don't need to know where all of the data for that student's educational activities are stored in the database. You only need to know how to locate the relevant data, given the logical structure of the database.

In general: to preserve a digital document, we must be able to identify and retrieve all its digital components – including the meaningful structure of it.

The digital components of a document are the logical and physical objects that are necessary to reconstitute the conceptual object. These components are not necessarily limited to the objects that contain the contents of a document. Digital components may contain data necessary for the structure or presentation of the conceptual object, like style sheets, form specifications and more complex ones, like name spaces.

To identify and retrieve the digital components, one must process them correctly. Digital preservation is not a simple process of preserving physical objects but one of preserving the ability to reproduce the objects. The success of digital preservation can be proved only by re-creating the document in some form that is appropriate for human use or for computer system applications.

Remember the first general question of this paper: what is „original" in the space of digital documents? Does it necessary to preserve the variety and complexity of the original objects? In the context of the three-layer model we should ask: does it necessary to preserve the physical and logical components of a digital document and also their interrelationship, without any alteration?

No. You can change the way a conceptual object is encoded in logical objects and stored in physical objects without having any negative impact on its preservation. E.g. in a repository of staff members data in a university database CVs are defined as textual and embedded image (photo) files. The photographs of staff members can be stored in separate image files (for different applications – ensuring single storage for multiple applications), and there are only links in the CV files to the appropriate image file. However, the image file could be embedded in the word processing file without altering the report as such. One can produce PDF version of this CV.

At first sight change and preservation are opposite categories. On second thought the possibility of preserving a digital document while changing its logical encoding or physical inscription

promises benefits, especially in long-time preservation. Technology creates the possibilities for change, but we should determine what changes are permissible, beneficial, necessary, or harmful.

To make such determinations, we have to consider the ultimate purpose of preservation. What is the goal of digital preservation of documents?

For libraries, archives, and other organizations that are for preservation of digital documents over time, the ultimate outputs are authentic preserved documents. According to the previous parts of this paper the output of a preservation process must be identical in all essential aspects, to what went into that process. Identical - in all essential aspects.

The ideal preservation system would be a communications channel for transmitting information to the future. This channel should not corrupt or change the messages transmitted in any way. The process of preserving digital documents is essentially different from that of preserving physical objects such as traditional books on paper. To access any digital object, we have to retrieve the stored data, reconstituting, if necessary, the logical components by extracting or combining the bytes from physical files, reestablishing any relationships among logical components, interpreting any syntactic or presentation marks or codes, and outputting the object in a form appropriate for use by a person or a business application. We don't want to preserve a digital document as a physical object; instead we need to ensure the ability to reproduce the document for future users. The preservation of an information object in digital form is complete only when the object is successfully reconstructed. In fact, the original document is not retrieved, but „copied", as it is reproduced by processing the physical and logical components using software that recognizes and properly handles the files and data.

Paper degrades, ink fades. (Lorie, 2000) In general we are not able to assert with complete assurance that no substitution or alteration of the object has occurred over time. Authentication of preserved objects is ultimately a matter of trust. There are ways to reduce the risk entailed by trusting someone, but ultimately, you need to trust some person, some organization, or some system or method that exercises control over the transmission of information over space, time, and technological boundaries.

Can an object change and still remain authentic? Common sense suggests that something either is or is not authentic, but authenticity is not absolute. Authenticity depends on use. (Thibodeau, 2001) The criteria for authenticity depend on the intended use of the object.

A document known to be in someone's handwriting, but containing text he copied from a book, does not reveal his thoughts. Oppositely, the final testimonial of a person can be written down by his secretary. Authenticating something as someone's writing depends on how we define that concept.

There are contexts in which the intended use of preserved information objects is well-known. For example, many corporations preserve records for long times for taxation purposes. It is a clear case: we know the exact aim of the preservation and the intended use as well. Libraries and public archives, however, usually cannot prescribe or predict the future use of their collection. Such institutions generally maintain their collections for access by anyone, for whatever reason. Users and their behaviors are not known in advance, you must assume that any valid intended use must be somehow consonant with the original nature and use of the document. Anyway, given that a digital document is not something that is preserved as an inscription on a physical medium, but something that can only be constructed or reconstructed by using software to process stored inscriptions, it is necessary to have an explicit model that is independent of the stored object and that provides a criterion, or at least a benchmark, for assessing the authenticity of the reconstructed object.

7 CHANNELS TO THE FUTURE

A preservation system will act as a communications channel for transmitting information to the future only if it systematically supports the preservation of the original context of the document(s). That is why you should manage the semantics of the document, what can be done in the model described above – at the Conceptual Layer.

We use metadata for contextual description. (Magyar, 2004) The contextual information serves to provide a more complete understanding of the document(s). The most important method for contextual description is taxonomy. Taxonomy is a classification of information components and their interrelationships that supports the discovery of and access to information. Metadata and taxonomies can work together to identify information and its features, and then organize it for access, navigation and retrieval.

Terms and taxonomies are so subjective. Different users and different applications may use a variety of terms for the same concept. Humans can also easily relate the meanings of two terms as "similar", "the same", "more general" or "more specific". (Magyar, 2005) Computers can handle only strings, not concepts. The semantics of a domain model is machine-usable only if it is expressed using an agreed upon vocabulary and syntax. The heterogeneity of the information environment also relates to the coexistence of structured, semi structured and unstructured information. (There are, of course essential differences between these categories in terms of their conceptual schemas and their inherent machine and human understandable representations.) (Magyar, 2004)

In the case of the structured databases, the addition of lexical relations to the semantic relationships already defined by the schema. In the class of semi-structured documents, in addition to their classical role for assigning and searching metadata, thesauri can be used for searching and retrieving free text or integrated into data mining, knowledge extraction, or other related applications. (NISO, 1993) Thesauri can be used for assigning metadata to different kinds of materials, such as images, videos, sounds, etc. in rich multimedia documents. Thesauri can serve for linking all different types of information and their representations in the domain of digital documents. (Kosovac, 1998) Thesaurus services can be used to complement the modeling technology by extending mechanisms for achieving semantic interoperability to the level of human-understandable information representations.

8 CONCLUSION

A comprehensive model of long-term archiving of digital documents was presented. The model is applicable to all kinds of rich content multimedia documents. The importance of preserving semantics of document was emphasized.

Semantic approaches try to construct and use formalized imprints of human conceptual interpretation results. Semantic tools tend to be part of the so-called content-infrastructure. Using that kind of semantic resources conceptual objects could be interpreted perhaps not only by human beings but by software agents as well.

As a final conclusion for future long-term archival systems we must assume that semantic interoperability needs integrated solutions at all the three layers of the model described in this paper

REFERENCES

Buckland, 1998; Michael Buckland: What is a "digital document"? Document Numérique (Paris) 2, no. 2 (1998): 221-230

DC, 1998; Dublin Core Metadata for Resource Discovery. IETF #2413 / Weibel, S.; Kunze, J.; Lagoze, C.; Wolf, M. The Internet Society, 1998. September 1998.(http://purl.org/DC/index.htm)

Lorie, 2000; Lorie, Raymond A. 2000. The Long-Term Preservation of Digital Information. http://www.si.umich.edu/CAMILEON/Emulation%20papers%20and%20publications/Lorie.pdf.

Thibodeau, 2001; Kenneth Thibodeau: Building the Archives of the Future. D-Lib Magazine. February 2001. Volume 7 Number 2. ISSN 1082-9873

Kosovac, 1998; Kosovac, B (1998). Internet/Intranet and Thesauri, Canadian Institute for Scientific and Technical Information, Internal Report, National Research Council Canada, Ottawa, Canada. <http://www.nrc.ca/irc/thesaurus/roofing/report_b.html>

Magyar, 2001; Magyar, G., Szakadát: Metadata System of National Audiovisual Archive in Hungary. Invited Paper. 20th Conference of the Audio Engineering Society: Archiving: Restoration and New Methods of Recording. Budapest, 5-7 October 2001. Mira Digital Publishing Inc., 2001

Magyar, 2004; Tikk D., Kardkovacs Z, and G. Magyar, „The hungarian deep web searcher project," Internation Journal on Information Technology, vol. I, pp. 191–197, Dec. 2004.

Magyar, 2005; Tikk D., Szidaroszky F. P., Kardkovács Zs., Magyar G.: Entity Recognizer in Hungarian Question Processing. In: Lecture Notes in Computer Science, 2005, Publisher: Springer-Verlag GmbH, ISSN: 0302-9743

Moore, 2000; R. Moore et al. Collection-Based Persistent Digital Archives, D-Lib Magazine, March 2000, Volume 6 Number 3 [Part 1] <http://www.dlib.org/dlib/april00/moore/04moore-pt2.html>.

NISO, 1993; National Information Standards Organization (1993). ANSI/NISO Z.39.19-1993. Guidelines for the Construction, Format, and Management of Monolingual Thesauri, Bethesda, MD: NISO Press.

RDF, 2006; Resource Description Framework (RDF) / W3C Semantic Web Activity. http://www.w3.org/RDF