Terminology and Wish List for a Formal Theory of Preservation

Giorgos Flouris [(1),(2)], Carlo Meghini [(2)]

[(1)] Foundation for Research and Technology – Hellas (FORTH)
Institute of Computer Science (ICS)
Vassilika Vouton, P.O. Box 1385, GR-71110, Heraklion ,Greece
EMail: fgeo@ics.forth.gr

[(2)] Consiglio Nazionale delle Ricerche (CNR)
Istituto della Scienza e delle Tecnologie della Informazione (ISTI)
Via Giuseppe Moruzzi, 1, 56124, Pisa, Italy
EMail: {flouris,meghini}@isti.cnr.it

ABSTRACT

The ability to access and understand digital information in the long term has been the subject of study of the digital preservation field, which is one of the most challenging research problems faced by the community of digital libraries today. The related research has produced a number of ad-hoc solutions which deal mainly with the static aspects of the problem. In this paper we take a different path in an attempt to formally describe the dynamic aspects of the problem. We describe and justify a number of desired properties of a good solution to the problem, as well as the basic steps required in order to get an adequate such formalism. This paper is part of a larger ongoing work towards developing a formal, mathematical, logic-based description of preservation as a scientific discipline, to the end of deriving a methodology resting on solid theoretical grounds.

Keywords: Digital Preservation, Information Preservation, OAIS, Theory of Digital Preservation

INTRODUCTION

One of the most difficult problems being faced by modern archivists is the rapid obsolescence of large volumes of digital (especially "born-digital") information. This problem is being addressed in the research area of *digital preservation* [11], [16], which deals with the problem of retaining the meaning of a digital object (file, image, database, document, etc) unaltered for an evolving community of readers. This goal can be achieved by defining suitable protocols, procedures and general methodologies that will allow accessing and understanding a digital object for a "sufficiently large" period of time or for a "sufficiently broad" audience (community of readers).

Preserving digital material is critical for modern digital libraries and archives [11], [16], so the problem has attracted a great deal of attention by researchers and practitioners alike. In particular, there has been a number of ongoing efforts dealing with the practical and methodological aspects of preservation (e.g., [7], [14], [15], [18], [19]), including projects such as CAMILEON [1], CASPAR [2], CEDARS [3] and PLANETS [17]. The relevant research efforts have resulted to a number of proposals to deal with the problem, including international standards, such as the OAIS [5], [13].

Despite these efforts, the preservation of digital information is still a very hard problem, not fully understood to date; one of the major gaps in the related research is that there are very few works in the direction of a formal description of the problem (the only effort the authors are familiar with is [6]). In particular, a commonly accepted formal model to describe the problem or a formal description of the required properties of a good solution is missing from the relevant literature. Thus, current solutions are mainly ad-hoc and application-specific. Moreover, the related research has focused mainly on the static aspects of preservation, greatly overlooking the dynamic aspects, i.e., the description of how a community of readers could evolve, and what effects different evolution patterns would have on the intelligibility of a digital object.

This paper presents some ideas towards filling these gaps by focusing on the theoretical dimensions of the problem and proposing some preliminary formal definitions and requirements for a theory of preservation. A central concept of our approach is the formal description of the dynamics of readers' understanding, upon which the theory is based. We will propose certain definitions and concepts which are part of a larger ongoing effort towards the development of a formal, mathematical, logic-based description of preservation as a scientific discipline, to the end of deriving a methodology resting on solid grounds.

We believe that the introduction of such a formal description would in many ways contribute to the research field of digital preservation. For example, a formal theory could allow the development (and proof) of impossibility and existential results: given the inherent difficulties associated with the problem, we intuitively expect some limitations on what types of digital objects can be preserved; we also expect certain types of readers' evolution to be such that no preservation is possible. In addition, a formal theory could allow the grounding of existing (and future) preservation methods upon a common formalism for comparison, and could result to a set of formal desirable properties for evaluating such methodologies [8], [9].

DISCUSSION

As already mentioned, digital preservation refers to the problem of retaining the meaning of a digital object unaltered for an evolving community of readers. In the standard terminology of the field, the creator of the digital object is referred to as the *producer*, whereas the reader of the information is called the *consumer* [5], [13]; in most cases, a digital object is intended to be read by a particular group of consumers who share some common characteristics (e.g., a scientific paper on digital preservation is usually meant to be read by scientists and archivists interested in the problem of preservation). Such a group of readers is usually referred to as the *Designated Community* (DC) of the digital object [5], [13].

To understand the process of "retaining the meaning of a digital object unaltered for an evolving community of readers", we must first understand the process of deciphering the meaning of a digital object. So, let us consider the example of an image I, produced by a producer P, in order to be read by a consumer C (who is a typical member of the DC). The goal of preservation is to ensure the intelligibility of I by C despite the many changes that can intervene as time passes by. Understanding in this context implies accessing, of course, but access alone is (usually) not enough. Below, we sketch the general steps required for the process; notice that most of these steps require the use of some artificial agent (software program, hardware device etc) to apply the relevant transformation:

- The original input is the physical storage (on some form of long-term storage media) of the sequence of bits which encodes the image I in some format.
- By reading these bits from the storage media, C obtains a sequence of bit values representing the image.
- By rendering these bits, C obtains an image that is some form of light that C's eyes can take in. Rendering presupposes some knowledge on the image format.
- By interpreting the image, C figures out its meaning, i.e., the worlds in which the portrayed scene can occur.

The fundamental divide in the above discussion is the separation between rendering the object, the image in our case, and understanding the object. This leads to our informal definition of preservation as: the ability to perform the interpretation of the rendering of a bit stream at any time. Notice that this involves three steps: producing the bit stream, rendering the produced bit stream and understanding the rendered object.

This results to a decomposition of the preservation task into sub-tasks, each corresponding to one preservation type. The first type, called *bit preservation*, refers to the ability to produce a particular sequence of bits from a storage media at any time; this can be achieved using error correction techniques, backups, RAID or mirrored disks, media refreshment and other technologies. The second type, called *data preservation* or *object preservation*, refers to the ability to render the produced bit

stream and produce a meaningful output from it at any time. This is the focus of most current approaches to the problem. The third type, called *information preservation*, refers to the ability to understand the rendered object at any time, i.e., to be able to understand its content by understanding the terms, concepts or other information that appears in it, by placing it in its correct context etc. This is the toughest type of preservation, and is often ignored by existing preservation approaches.

In what follows, we will not consider bit preservation; for some relevant discussion, refer, for example, to [18], [19]. Our work focuses on information preservation, even though most of the discussion presented here can be easily applied for data preservation as well.

Another important observation that can be made here is that the need for preservation can appear in both space and time dimensions. The "space" dimension refers to the fact that different people have different background knowledge and, consequently, may have trouble understanding each other's documents (e.g., an astronomer may have trouble understanding a scientific paper on computer software). Similarly, the "time" dimension refers to the fact that different people in different times (or even the same person in different times) have different background knowledge, so the same problem may occur (e.g., a telephone number written in a post-it note may be unusable if the writer forgets to whom this number belongs). We argue that the two "preservation dimensions" are, for all practical purposes, the same, so approaches to the problem of preservation can interchangeably handle both dimensions.

DIGITAL OBJECT AND UNDERLYING COMMUNITY KNOWLEDGE

It is clear by the above analysis that a central concept in preservation is that of the "meaning" of a digital object. The meaning of a digital object can be viewed as a special kind of mapping that associates a symbol with a particular real-world concept. This association is not always clear by looking at the digital object alone. For example, the text 'Carlo's birthday is on 02/08' could either mean that Carlo was born on the 2nd of August (European notation) or on the 8th of February (American notation). As a result, the exact association of this text to a real-world concept (the date of Carlo's birthday) is something external to the object that should be somehow recorded (unless the DC is already aware of this association).

In order to capture the concept of "meaning" we will assume that any digital object is expressed using the conventions described in some language, say L, which is different per digital object and provides the syntactic and semantic rules that allow us to formulate the digital object and associate it with some meaning. The language L should be a formal language of a logical nature, and can be seen as the encoding of the "language understood" by the consumer (or producer). There are several arguments in favour of the choice of L to be a logical language. First, L has to be formal, like logics are, otherwise no scientific theory of preservation can be developed; second, it must be able to express knowledge, and formal logic has been developed for exactly this purpose; third, it must be suitable for capturing implicit knowledge, and the inference relation of mathematical logic allows precisely that; and, finally, logic is a very well studied field of science, offering a very rich set of results from which to draw.

As already mentioned, L has a dual purpose. Firstly, it provides the formulation rules for creating a digital object. These rules can be viewed as the format specification of the object and are encoded using the syntactical formulation rules of the logic at hand. Secondly, it provides the semantics of a digital object. These semantics are provided using the logic's semantics. In addition, any contextual, background or commonsense information, as well as any terminology used by the digital object can (and should) be captured using logical constructs in a logical theory, formulated under L.

The language L, along with the theory T that encodes the above information will be called the *Underlying Community Knowledge* (UCK). A *digital object* D, is a "piece of knowledge" represented using the "language understood" by (the UCK of) the producer, i.e., a set of formulas of L. This way, each digital object is associated with a certain UCK which provides its meaning.

An alternative proposal for the definition of a UCK and a digital object, can be found in [9]; in that work a central concept is that a digital object is defined as a set of queries and their answers, based on the language L. It is easy to show that the approach presented here is more general, but the approach of [9] is closer to everyday practice, as it is based on the realization that it is not usually necessary (or possible) to preserve the entire information carried by a digital object. Instead, we could isolate and preserve the

object's most "useful" or "important" information, which is just the questions modelling the digital object. Determining the information worth preserving for the object at hand is not an easy task, as it depends on the object type, its content, legal issues as well as on the needs of the creator and the reader of the information. A great aid in this task is provided by preservation models, such as the OAIS [5], [13]; the role of such a model in this respect is to provide a methodological framework and a "best practices" approach towards the aim of determining the most important pieces of information contained in a digital object.

DESCRIBING DC EVOLUTION

Based on the analysis of the previous section, we can identify a producer with the language he understands, i.e., its UCK. Similarly, the knowledge of a DC can be viewed as another UCK, and a consumer-member of that DC, can be also identified with that UCK, as he is assumed to be a typical representative of the DC. A digital object is represented using the "language understood" by the producer, so it is "associated" with (the UCK of) the producer. Since a consumer will (in general) understand a different UCK, we need to make the necessary amendments to the digital object so that it is understood correctly by the consumer. In other words, we could say that the purpose of preservation is to "translate" the digital object from the "language understood" by (the UCK of) the producer into the "language understood" by (the UCK of) the consumer.

In order to achieve the above goal, we need to encode the dynamics of languages, that is to be able to describe (formally) the differences between the producer's and consumer's UCKs. This "delta" could represent the evolution of knowledge over time, or it could represent the differences in understanding and terminology between two people with different backgrounds.

The main focus of a formal approach to preservation should be to define a formal structure that will allow us to represent and describe these differences (delta); this structure will be called the *UCK Evolution Structure* (UCKES). A UCKES should be powerful enough to capture changes in UCK's language (which could be either semantic or syntactic), as well as changes in UCK's theory. For the latter type (changes in the theory), the ideas developed in the well-established research fields of belief revision [10] and ontology evolution [12] could prove helpful; this is not true for the former type (changes in the logic), as, to the authors' knowledge, the problem of language dynamics has not been studied in the literature.

Given the producer's UCK and a UCKES, we should be able to regenerate, step-by-step, the changes that the producer's UCK went through in order to get the consumer's UCK; thus, the first important function of a UCKES is to allow us to construct the consumer's UCK.

The second important function of a UCKES is to allow us to define a number of associations between the producer's and consumer's UCKs. Given that meaning itself cannot be represented (consider the mental experiment with a Chinese-to-Chinese dictionary appearing in [4] – a non-Chinese-speaking person can never learn Chinese using a Chinese-to-Chinese dictionary alone), our only option in order to understand something unknown is to map it to (or associate it with) some symbol whose meaning we already know. These associations will allow the consumer to understand the meaning of a formula (or term, structure, symbol etc) in the producer's UCK, as it shows how to express it in terms of formulas (or terms, structures, symbols etc) understood by the consumer (consumer's UCK).

Notice that these associations could be very simple when the changes performed are syntactic, but in the general case they could be quite complicated; for example, if there is no formula in the consumer's UCK that corresponds to the exact same meaning as the given formula in the producer's UCK, then we need some complex association mechanism that will allow us to describe the exact relationship between the "real-world meaning" of the given formula (producer's UCK) and the "real-world meaning" of formulas in the consumer's UCK. This complex mechanism should be based on a more expressive UCK-like language (an "interlingua"), that will allow us to encode such complex associations; this language will be called the *UCK Mapping Structure* (UCKMS).

As already mentioned, the purpose of preservation is to evolve a digital object in a manner that will not alter its meaning, despite the changes that the underlying UCK may have been through (i.e., despite the

differences between the producer's and consumer's UCKs). The UCKMS is the tool that can help us in this task, as it encodes the associations between the various elements that appear in the old digital object (which are elements from the producer's UCK) with the respective elements in the consumer's UCK. Using this structure, we could replace each element of the producer's UCK that appears in the digital object with its "counterpart" in the consumer's UCK.

The latter observation leads us to the introduction of the notion of "replacement of elements", which gives rise to two different related concepts. The first concept captures an exact replacement of element; it can be viewed as a relation, called the *identification relation* (and denoted by ~), which associates the elements of the two UCKs that carry identical meanings. The important point here is that elements associated via ~ should carry identical meanings; unfortunately however, associations cannot always be perfect. In addition, we could sometimes be happy enough to replace a formula of the original digital object with a formula from the consumer's UCK that is, in a sense, "equivalent" to the given formula; notice that this "equivalence" does not refer to standard logical equivalence, as it associates elements from different logics. This relaxed form of identification can be again encoded as a relation that will be called the *preservability relation* and denoted by $\cong$.

The preservability relation should be also applicable (or extended) to digital objects. Using it, we will say that a digital object D of the producer's UCK *is preserved by* a digital object D′ of the consumer's UCK if and only if $D \cong D'$; if there is such a D′, then D will be called *preservable*. Thus, the preservability relation provides the means to determine whether a digital object preserves another and allows us to fine tune the notion of "preserving a digital object" by making it as liberal (or as strict) as we want it to be. Obviously, a more restrictive preservability relation implies that the meaning of a digital object will be carried over more faithfully during the transition, but also implies that there will be more cases (i.e., more types of UCK evolution) in which certain digital objects will not be preservable.

In the case where all possible digital objects of the producer's UCK are preservable, we can define a *migration function* associating each digital object in the producer's UCK with a digital object in the consumer's UCK that preserves it; in the general case, a migration function may be applicable only upon preservable digital objects. Notice that a digital object might be preservable by more than one digital objects in the consumer's UCK; in this case, we should define a *preservation policy* (which is modelled using the migration function) in order to determine which of the alternative digital objects should be selected as the output of the migration function. A *preservation system* is a system that implements the migration function. Preservation is *successful* for a particular digital object D if and only if D is preservable and the preservation system correctly implements the migration function, i.e., it gives in the output a digital object D′ for which $D \cong D'$.

CONCLUSIONS

Digital preservation is a critical problem faced by modern digital libraries and refers to the problem of retaining the meaning of a digital object unaltered for different readers. In most cases, the problem appears due to the passage of time that makes old formats, terminology, jargon etc obsolete, so it has often been identified as the problem of assuring "interoperability with the future". This paper focused on an effort to define a set of desired properties for a formal framework that will describe the problem; unlike current approaches to the problem, our approach was centred around the dynamic aspects of the problem, and attempted to remain as general as possible so as to be application-independent (as opposed to the standard practice which usually employs application-oriented, ad-hoc solutions to the problem).

Our methodology identified a digital object with a logical structure based on a logical language (UCK) that represents the background knowledge of, and the language understood by, the creator of the digital object (producer). This way, the problem of preservation can be described as the problem of changing the original digital object in such a way that the new digital object can be understood by (the UCK of) the consumer and that the meaning understood by the consumer while reading the new digital object (using his own UCK) is the "same" (under the notion of preservability) to the one intended by the producer for his original digital object (under his own UCK).

The input to our methodology is the original digital object, the producer's UCK, and the description of the "delta" (UCKES) between the producer's and consumer's UCK. Given the producer's UCK and the UCKES, we are able to calculate, as intermediate products, the consumer's UCK and the UCKMS, i.e., the associations between the two UCKs. The final product, which is also the output of our methodology, is the new digital object, which should be such that (a) it can be understood by the consumer (i.e., it should be formulated using the language represented by the consumer's UCK) and (b) it "preserves" the original digital object, per the preservability relation. The determination of the digital object that preserves the original one is heavily based on the association information provided by the UCKMS and on the preservation policy, which allows us to choose among the various alternative results (in the cases where there is such a choice involved).

Notice that our methodology presupposes that the differences between the producer's UCK and the DC, for which the digital object is intended to be preserved, are known in advance, so as to know the UCKES. Also, preservation is assumed to be taking place while there are still people who are aware of the producer's UCK, or that the producer's UCK has been somehow recorded along with the digital object.

We are currently in the process of implementing our "wish list" in the sense of defining the formal constructs that will model the above structures and satisfy the requirements presented in this paper. This paper is not meant to be the final word on the subject, as the ideas presented here constitute just the first of a long series of steps towards the development of a formal theory of digital preservation; instead, our goal was to outline our planned research path and present the initial results of our research in this direction.

REFERENCES

[1] - CAMILEON: Creative archive at Michigan and Leeds, emulating the old on the new. NSF, JISC funded project. http://www.si.umich.edu/CAMILEON

[2] - CASPAR: Cultural, artistic and scientific knowledge for preservation, access and retrieval. EU funded project (FP6-2005-IST-033572). http://www.casparpreserves.eu

[3] - CEDARS: curl exemplars in digital archives. JISC funded project. http://www.leeds.ac.uk/cedars

[4] - A. Cregan: Symbol grounding for the semantic web. In Proceedings of the 4th European Semantic Web Conference (2007)

[5] - ISO 14721:2003 CCSDS 650.0-B-1: Reference Model for an Open Archival Information System (OAIS). Blue Book, Issue 1, (2002) http://ssdoo.gsfc.nasa.gov/nost/isoas/ref model.html

[6] - J. Cheney, C. Lagoze, P. Botticelli: Towards a theory of information preservation. In Proceedings of the 5th European Conference on Research and Advanced Technology for Digital Libraries (2001)

[7] - M. Factor, D. Naor, S. Rabinovici-Cohen, L. Ramati, P. Reshef and J. Satran: The need for preservation aware storage: A position paper. ACM SIGOPS Operating Systems Review, 41(1), pp. 19–23 (2007)

[8] - G. Flouris, C. Meghini: Steps towards a theory of information preservation. In Proceedings of the International Workshop on Database Preservation, Invited Talk (2007)

[9] - G. Flouris, C. Meghini. Some preliminary ideas towards a theory of digital preservation. In Proceedings of the 1st International Workshop on Digital Libraries Foundations (2007)

[10] - P. Gardenfors. Belief revision: An introduction. In P. Gardenfors, editor, Belief Revision, pp. 1–20, Cambridge University Press (1992)

[11] - H. Gladney: Preserving digital information. Springer-Verlag (2007)

[12] - P. Haase, Y. Sure: D3.1.1.b state of the art on ontology evolution (2004) http://www.aifb.unikarlsruhe.de/WBS/ysu/publications/SEKTD3.1.1.b.pdf

[13] - B. Lavoie: The open archival information system reference model: Introductory guide. In DPC Technology Watch Report 04-01 (2001)

[14] - C. Lynch: Canonicalization: A fundamental tool to facilitate preservation and management of digital information. D-Lib Magazine, 5(9) (1999)

[15] - P. Mellor, P. Wheatley, D. Sergeant: Migration on request, a practical technique for preservation. In Proceedings of the 6th European Conference on Research and Advanced Technology for Digital Libraries, pp. 516–526 (2002)

[16] - A. Pace: Coming full circle, digital preservation: Everything new is old again. Computers in Libraries, 20(2) (2000)

[17] - PLANETS - Digital Preservation Research and Technology. EU funded project. http://www.planets-project.eu

[18] - D. Rosenthal: Engineering issues in the preservation of databases. In Proceedings of the International Workshop on Database Preservation, Invited Talk (2007)

[19] - M. Roussopoulos: A fresh look at the reliability of long-term digital storage. In Proceedings of the International Workshop on Database Preservation, Invited Talk (2007)