Terminology and Wish List for a Formal Theory of Preservation

Giorgos Flouris, Carlo Meghini

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

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 information preservation. Preserving digital information is a very hard problem, not fully understood to date; in particular, there is no commonly accepted formal model to describe it or a formal description of the required properties of a good solution. This paper presents some ideas towards filling this gap by focusing on the theoretical dimensions of the problem and proposing some preliminary formal definitions and requirements for a theory of preservation.

We begin with a general analysis which identifies the need to define three different types of preservation. The first type, *bit preservation*, refers to the ability to read a particular sequence of bits; this can be achieved using error correction techniques, backups, RAID or mirrored disks, media refreshment and other technologies. The second type, *data preservation*, refers to the ability to render the digital object and produce a meaningful output from the data. This preservation type is the focus of most current approaches to the problem. The third type, *information preservation*, refers to the ability to understand the rendered file, 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. We argue that a complete preservation model should handle all three preservation types.

We continue by presenting a number of examples which show that a central concept for preservation is that of the “meaning” of a digital object. Our analysis shows that any digital object is given meaning using certain assumptions related to its format, context, terminology and other commonsense and background knowledge, most of which is often implicit. We provide convincing argumentation that this background information can and should be captured using some kind of logical formalism (not necessarily the same for all digital objects) plus a logical theory, expressed in terms of this formalism, which captures the community’s background knowledge. This structure is called the *underlying community knowledge*. Thus, each digital object is associated with a certain underlying community knowledge which provides its meaning.

Regarding the digital object, we argue that it is not usually necessary (or possible) to preserve the entire information carried by it; instead, we could isolate and preserve the object’s most “useful” or “important” information. Determining the information worth preserving for the object at hand is not an easy task; 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, which we embrace in this work. 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.

Based on these fundamental notions, we formally define concepts like preservation policy, preservation system, successful preservation, emulation, migration etc, as well as OAIS-related notions, such as the consumer, the producer, the designated community etc. We define desired properties for a preservation system and provide some discussion on the issues related to the evolution of the underlying community knowledge and the effects of such an evolution on the associated objects. Finally, we show that, under certain conditions, this evolution can be modelled using techniques from the well-established research fields of ontology evolution and belief revision.