DATA PRESERVATION

Carlo Meghini

Istituto di Scienza e Tecnologie dell' Informazione (ISTI)"Alessandro Faedo", Italian National Research Council (CNR), Via G. Moruzzi 1, 56124 Pisa, Italy
Email: carlo.meghini@isti.cnr.it

1 STATE-OF-THE-ART

Digital information is a vital resource in our knowledge economy, valuable for research and education, science and the humanities, creative and cultural activities, and public policy (The Blue Ribbon Task Force on Sustainable Digital Preservation and Access, 2010). New high-throughput instruments, telescopes, satellites, accelerators, supercomputers, sensor networks, and running simulations are generating massive amounts of data (Thanos, 2011). These data are used by decision makers for improving the quality of life of citizens. Moreover, researchers are employing sophisticated technologies to analyse these data to address questions that were unapproachable just a few years ago (Helbing & Balietti, 2011). Digital technologies have fostered a new world of research characterized by immense datasets, unprecedented levels of openness among researchers, and new connections among researchers, policy makers, and the public (The National Academy of Sciences, 2009).

But digital information is inherently fragile and often at risk of loss. The project PARSE.Insight (http://www.parse-insight.eu/) has conducted a survey that has identified the main threats to the permanent access to the data collected by different scientific disciplines and stakeholders. According to the PARSE.Insight FP7 Project (2010), these threats are:

1. Users may be unable to understand or use the data,
2. Lack of sustainable hardware, software, or support of computer environment may make the information inaccessible,
3. Evidence may be lost because the origin and authenticity of the data may be uncertain,
4. Access and use restrictions (e.g., Digital Rights Management) may not be respected in the future,
5. Loss of ability to identify the location of data,
6. The current custodian of the data, whether an organization or project, may cease to exist at some point in the future, and
7. The ones we trust to look after the digital holdings may let us down.

Digital preservation aims at countering these threats by maintaining digital information accessible, independently understandable and usable by a designated community, and with evidence supporting its authenticity, over the long term (Giaretti, 2011).

Digital preservation is a relatively young discipline, whose importance increases as the amount of knowledge encoded exclusively in digital form increases. The basic concepts underlying digital preservation, including its very definition, are set by the Open Archive Information System (OAIS) Reference Model (Lavoie, 2004; CCSDS, 2002), an ISO standard (ISO 14721:2003, see http://www.iso.org/iso/catalogue_detail.htm?csnumber=24683) that plays a central role in digital information preservation. According to OAIS, digital information is preserved for a designated community that is an identified group of potential consumers that may be composed of multiple user communities and whose definition may change over time. The objective of preservation is to enable these potential customers to access, understand, and use a particular set of information without having to resort to special resources not widely available, including named individuals. Moreover, digital preservation spans the long term, that is "a period of time

long enough for there to be concern about the impacts of changing technologies, including support for new media and data formats, and of a changing Designated Community" (Giaretti, 2011), on the information being preserved. This period extends into the indefinite future.

Digital preservation requires the collection, management, and use of metadata about the object to be preserved. The central problem is which metadata. There is already a substantial body of literature on the subject (e.g., Gladney, 2007; Deegan & Tanner, 2006; Brown, 2006; Borghoff, Rodig, Scheffczy, & Schmitz, 2003) addressing the preservation of various types of objects. To date, the OAIS Reference Model remains the most general conceptual framework for digital preservation: it defines the goals of preservation, the functions that are required for realizing it, and the types of metadata that are necessary for achieving digital preservation independently of the object type or context of application. In essence according to OAIS, preservation requires metadata concerning the understandability of objects and their origins, context, and restrictions (Giaretti, 2011). The metadata required for understandability are termed in OAIS as *Representation Information.* Representation Information is defined as "the information that maps a Data Object into more meaningful concepts" and is further decomposed into:

- *Structure* Information, essentially the encoding format of the Data Object, and
- *Semantic* Information, understood as any description that captures enough semantics of the Data Object.

The metadata required for origins, context, and restrictions are termed in OAIS as *Preservation Description Information* and are further decomposed into:

- *Reference* Information, dealing with identification of the Data Object,
- *Provenance* Information, dealing with the history of the Data Object,
- *Context* Information, dealing with the relationships of the Data Object with its environment,
- *Fixity* Information, dealing with the mechanisms for ensuring that the Data Object has not been altered, and
- *Access Right* Information, dealing with the access restrictions of the Data Object.

Beyond these definitions, the fact remains that understandability is a particularly difficult notion to capture; it is hard to define in general, and moreover, it is a relative concept: what is understandable for one person may not be so for a different person. OAIS overcomes these difficulties by referring Representation Information to a designated community. This solves the problem of relativity (or at least confines it into a well-defined scope) and also allows the qualification of the understandability with respect to a specific language and body of knowledge: the language spoken by the designated community and the body of knowledge encoded in that language possessed by the designated community.

Once an object is stored on a safe memory device and endowed with the necessary metadata, its accessibility and understandability are ensured until something changes in the technological architecture that makes the object accessible and in the linguistic and epistemological architecture that makes the object understandable. Naturally, the passage of time brings change in both of these architectures: some hardware or software component may become obsolete, making the technological architecture no longer effective; the language or the knowledge of the designated community may change, changing or destroying the meaning of the metadata associated with the object. Actions must be taken in order to preserve accessibility and understandability over time. These actions constitute preservation.

In the United States, in the year 2000, the Library of Congress was charged to create the National Digital Information Infrastructure and Preservation Program (NDIIPP) to develop a strategy to meet the challenge of digital preservation. A network of institutions committed to preserving the nation's digital heritage is now poised to carry forth this strategy. The Library of Congress (2011) summarizes the program's accomplishments to date and outlines its next steps.

In Europe, research and development in digital preservation has been identified as an important topic by the European Commission starting from the 5th Framework Programme, launched in 1998. To date, the Commission has funded 18 projects covering several aspects of digital preservation. Expenditure had grown from €0.9m in the 5th Framework Programme to more than €68m to date in the 7th Framework Programme, which is currently active (Billenness, 2011). Thanks greatly to this investment, progress has been made in the development of tools and methodologies that address preservation of digital assets for certain communities. These communities are, however, fragmented and not well connected and have not yet engaged widely beyond public memory institutions and research bodies. The base skills in digital preservation are also expanding, but there is not yet a corpus of Best Practice to refer to.

2 TEN-YEAR VISION

As we move away from the research world and from the few pioneer memory institutions, we observe a lack of awareness of the importance of digital preservation. As Billenness (2011) puts it, neither society in general nor large areas of business with a probable duty of long-term care for data are yet inspired by preservation. On the one hand, society seems to be unaware of the fragility of digital information. On the other hand, industrial partners who produce digital material are aware of the fragility of their products, but the long-term nature of digital preservation is incompatible with current economic modelling in which investment requires short-term returns. Digital Preservation is therefore seen as a problem to be addressed 'someday' but not as a high priority (Billenness, 2011).

This situation is going to change. The pervasiveness of technology in the life of citizens and in the operation of administrations is going to raise awareness of importance of digital preservation, along with the many initiatives that are being undertaken by national governments and the European Commission. The increase in demand will stimulate the creation of industrial-scale preservation services, leading in turn to the development of a new profession in digital preservation with well-defined roles and its own qualifications and training.

A fundamental responsibility will be the choice of what to preserve. The continuing and high rate of growth in volumes of data suggests that this year a total of 1,800 Exabytes will be produced, with an annual growth rate of 60%. The rate of growth in storage technology, at 25% per annum, is being outstripped by the rate of growth in data at 60% per annum. It is therefore not an option to preserve everything (Gantz et al., 2008). In order to offer the best possible support to the decision makers in selecting the data to be preserved, two fundamental breakthroughs are needed:

1. Economic models that enable cost-benefit analysis of digital preservation, highlighting the real value of information and the costs of its loss. There are at present no general, well-founded models to this end, just pioneers addressing the problems in their own specific fields with *ad hoc* methods.
2. Technology for assessing the level of fragility of digital information and the technical actions required for its preservation, to be used as a basis for cost estimations required by the economic models.

The effective implementation of preservation strategies will face the problem of sustainability. In order to reduce the costs of preservation, *sharing* at all levels will be required. In this respect, infrastructures may, and indeed should, play a crucial role in sharing.

Service infrastructures can support sharing the functionalities required for preservation by offering these functionalities as core services made available to their users similar to the services for the analysis, the transformation, or the exchange of data. Extraction of preservation metadata from digital objects, migration of objects from obsolete formats, and the emulation of applications running on obsolete hardware and software architectures are simple examples of the services for preservation that can be shared through an infrastructure.

Knowledge infrastructures can support sharing the knowledge required for preservation. The sharing may involve: (a) sharing ontologies (typically used for Semantic Representation Information) thereby reducing the cost of creating the Representation Information while avoiding the semantic interoperability problem that typically arises when accessing objects preserved by different archives and (b) sharing single records of Representation or Preservation Description Information (for instance, science data captured by the same device type will have the same provenance information, or documents issued under the same licensing scheme will have the same access right information).

Specific communities will create and manage their own infrastructures, as is already happening in several areas of science and society. The Web, on the other hand, will continue to offer a generic infrastructure, with web services as core middleware. In particular, the semantic web may play a crucial role in the sharing of knowledge by supporting the universal access to ontologies formalized in Resource Description Framework (RDF) Schema (Manola & Miller, 2004) or in the Ontology Web Language (McGuinness & van Harmelen, 2004; W3C OWL Working Group, 2009) and to records formalized as RDF graphs. Moreover, the Linked Open Data (Heath & Bizer, 2011) initiative may be very promising for access to both.

3 CURRENT CHALLENGES

- To raise awareness of the fragility of digital information.
- To devise effective criteria and associated tools for the selection of the digital assets to be preserved.
- To reduce the cost of preservation for all types of digital assets. Reducing the cost of storing materials, developing sustainable sources of energy to power preservation systems, and engineering ways to lower the cost of preserving, curating, and providing access.
- To expand and better integrate communities of interest.
- To have a deeper engagement between the engineering disciplines within computer science and the social science disciplines, which have driven the early research into digital preservation.
- To develop well-expressed business models to support investment in digital preservation.
- To consider newer format of digital objects – for example the outputs from social networks, in which it is as important to preserve the process as the objects themselves.
- To move from the preservation of data to the preservation of knowledge.
- To demonstrate the authenticity of data stored in preservation systems.

4 RESEARCH DIRECTIONS PROPOSED

At the object level, we envisage the following directions:

1. Extend the object **life cycle** to include the long-term, as defined by digital preservation. The extension should lead to an object model that can accommodate all information relevant for preservation. This information may be generated at different stages of the object's life-time, namely:
    a. *before* the object is created, the design and architectural decisions that have led to the object should be recorded as they form the technological architecture that keeps the object alive and must therefore be preserved.
    b. *during* the object's existence, it is important to record the operations that change the object because they constitute the provenance of the object, and provenance is required for preservation.
    c. *after* the object exits its usage context, it enters preservation and should therefore be endowed with the additional preservation metadata, such as Representation and Preservation Description Information.

2. Consider the preservation of **Complex Objects**. Complex objects in this context means web sites, interactive games, and complex data such as aeronautics. The preservation of a complex object of this kind

requires several techniques to be jointly employed to keep the object accessible, understandable, and usable in the long-term. Many of these techniques have been developed and tested on simple objects. Their applicability to complex objects is still an open problem, and substantial work is needed to reach a mature technology. Moreover, open standards must be created to facilitate interoperability and the creation of a market.

3. Consider **autonomic** computing as a paradigm for digital preservation. The growing size and complexity of digital objects suggests a new need for mechanisms able to automatically adapt to new, unexpected scenarios. A possible solution could be to enable digital objects to manage themselves without direct human intervention. In a self-managing autonomic system, the human operator takes on a new role: he does not control the system directly. Instead, he defines general policies and rules that serve as an input for the self-management process. The idea consists in having an autonomic object that senses the operating environment and takes action to change the environment or its own configuration with a minimum effort. This approach leads to objects that are self-preserving. The idea is clearly far-fetched; however, its pursuit is expected to lead to new ways of looking at the problem of digital preservation and possibly to new approaches.

At the knowledge level, we envisage the following directions:

1. Consider the **preservation of knowledge**. Preserving the technological architecture that makes an object accessible is a relatively well-understood problem even though we do not yet have a mature technology for it. However, the preservation of the language and epistemological architecture that makes the object understandable is still an open problem. This architecture consists of the vocabularies that are used in the preservation metadata; these vocabularies list the terms used in metadata and fix their meaning. Now, vocabularies are in continuous evolution: the vocabulary of science, for instance, evolves as a result of discoveries that reveal new aspects of known concepts or that bring up new concepts; likewise, the vocabulary of technology evolves as a reflection of the evolution of the products of technology. An undesired effect of the evolution of vocabularies is that preservation metadata may change their meaning if they use some term whose meaning has changed, or they may become incomprehensible if they use some term that is no longer in use. How to avoid this? In other words, how to preserve the original meaning of the preservation metadata across the change in the language in terms of which the metadata were expressed and in the knowledge base of the designated community on top of which these metadata rely? At present, the problem is solved by manually evolving preservation metadata, but this is not sustainable on the long term. Research is required in order to devise techniques that would allow evolving metadata in at least a semi-automatic way.
2. Consider the **automatic extraction** of preservation metadata. The general observation here is that often the information required for the preservation of an object is available from some digital source, be it the signal (text, image, video, or audio) carrying the content of the digital object or be it external, such as Wikipedia (for semantic information), system logs (for provenance information), configuration files (for context information), and others. The extraction of preservation metadata from these sources is therefore possible and also paramount for the sustainability of preservation. A wide range of techniques needs to be employed, depending on the type of the source and the type of sought information, ranging from machine learning, to data mining, to information discovery. The overall goal is to maximise automated extraction, giving rise to challenges over scalability, quality, process efficiency, and interoperability. These become increasingly difficult at higher levels of semantic extraction.

5 RECOMMENDATIONS

Make preservation sustainable by supporting the sharing of services and knowledge required for preservation. As remarked earlier, infrastructures, including the web, play a central role for the effective achievement of this recommendation.

Derive economic models for assessing the value of digital assets and inform the process of selecting the digital assets that need to be preserved.

Preservation should be seamlessly integrated into the life cycle of digital objects. The life cycle of digital objects should be expanded to include the early stages of life, namely object design and configuration, in order to obtain the technological architecture making the object accessible. The changes made to the object during the period of use should be recorded in order to derive the provenance and authenticity information of the object.

Automatic techniques should be devised for obtaining preservation metadata automatically or semi-automatically from the objects themselves (analysis of multimedia content) or from external sources.

Automatic or semi-automatic evolution of metadata should be actively researched in order to increase the level of automation and reduce the costs of preservation.

6 REFERENCES

Billenness, C. (2011) Report on the Proceedings of the Workshop "The Future of the Past" – The future of Digital Preservation Research Programmes, organised by The Information Society and Media Directorate General of the European Commission. Luxemburg.

Borghoff, U.M., Rodig, P., Scheffczy, J., & Schmitz, L. (2003) *Long- Term Preservation of Digital Documents*. Springer Verlag.

Brown, A. (2006) *Archiving Websites*. Facet Publishing.

CCSDS (Consultative Committee for Space Data Systems) (2002) Reference Model for an Open Archival Information System (OAIS). Consultative Committee for Space Data Systems Blue Book 650.0-B-1. Retrieved from the World Wide Web, June 27, 2013: http://public.ccsds.org/publications/archive/650x0b1.PDF

Deegan, M. & Tanner, S. (2006) *Digital Preservation*. Facet Publishing.

Gantz, J. et al. (2008) *The Diverse and Exploding Digital Universe*. IDC White Paper. Retrieved from the World Wide Web, June 27, 2013: http://www.emc.com/collateral/analyst-reports/diverse-exploding-digital-universe.pdf

Giaretta, D.L. (2011) *Advanced Digital Preservation*. Springer Verlag.

Gladney, H.M. (2007) *Preserving Digital Information*. Springer Verlag.

Heath, T. & Bizer, C. (2011) Linked Data: Evolving the Web into a Global Data Space (1st edition). *Synthesis Lectures on the Semantic Web: Theory and Technology* 1:1, 1-136. Morgan & Claypool.

Helbing, D. & Balietti, S. (2011) From Social Data Mining to Forecasting Socio-Economic Crises. White Paper of the EU Support Action Visioneer. Retrieved from the World Wide Web, June 27, 2013: http://arxiv.org/abs/1012.0178

Lavoie, B. (2004) The Open Archival Information System Reference Model: Introductory Guide. DPC Technology Watch Series Report 04-01.

Manola, F. & Miller, E. (2004) RDF Primer. W3C Recommendation, WWW Consortium. Retrieved from the World Wide Web, June 30, 2013: http://www.w3.org/TR/rdf-primer/

McGuinness, D. & van Harmelen, F. (2004) OWL Web Ontology Language Overview. W3C Recommendation, WWW Consortium. Retrieved from the World Wide Web, June 27, 2013: http://www.w3.org/TR/owl-features/

PARSE.Insight FP7 Project (2010) PARSE.Insight Science Data Infrastructure Roadmap. Deliverable D2.2. Retrieved from the World Wide Web, June 27, 2013: http://www.parse-insight.eu/downloads/PARSE-Insight_D2-2_Roadmap.pdf

Thanos, C. (2011) Global Research Data Infrastructures: The GRDI2020 Vision. Report of the GRDI2020 project funded under the 7th Framework Programme, Capacities – GÉANT & eInfrastructures. Retrieved from the World Wide Web, June 27, 2013: www.grdi2020.eu

The Blue Ribbon Task Force on Sustainable Digital Preservation and Access (2010) Sustainable Economics for a Digital Planet: Ensuring Long-Term Access to Digital Information. Final Report. Retrieved from the World Wide Web, June 27, 2013: http://brtf.sdsc.edu/biblio/BRTF_Final_Report.pdf

The Library of Congress (2011) Preserving Our Digital Heritage: The National Digital Information Infrastructure and Preservation Program 2010 Report. A Collaborative Initiative of the Library of Congress. Retrieved from the World Wide Web, June 30, 2013: http://www.digitalpreservation.gov/documents/NDIIPP2010Report_Post.pdf

The National Academy of Sciences: Committee on Ensuring the Utility and Integrity of Research Data in a Digital Age (2009) Ensuring the Integrity, Accessibility, and Stewardship of Research Data in the Digital Age. Retrieved from the World Wide Web, June 27, 2013: http://www.nap.edu/catalog.php?record_id=12615.

W3C OWL Working Group (2009) OWL 2 Web Ontology Language, W3C Recommendation, WWW Consortium, October 2009. Retrieved from the World Wide Web, June 27, 2013: http://www.w3.org/TR/owl2-overview/

(Article history: Available online 30 July 2013)