Automation in Digital Preservation

Schloss Dagsthul, 18-23 July 2010

Organizers

Jean-Pierre Chanod (Xerox Research Center Europe - Grenoble, FR)
Milena Dobreva (The University of Strathclyde - Glasgow, GB)
Andreas Rauber (Technical University Wien, AT)
Seamus Ross (University of Toronto, CA)

Editor

Vittore Casarosa (ISTI-CNR, IT)

Final reports of the break-out sessions

Session 1: Preservation Ready Systems – Digital Preservation and Enterprise Architecture

(discussion facilitator and rapporteur Jose Borbinha)

The Problem

A system is a collection of components organized to accomplish a specific function or set of functions [1]. In each moment the systems' functions are perceived as outputs resulting from the combination of the state of its information model, of its inputs, and of the behavioural model defined for that system.

We can classify the systems' inputs and outputs as, in general, matter, energy or information. Information Systems are those conceived and created considering that our inputs, processes and outputs will deal with information. Since these systems are increasingly relevant in our actual technological paradigm, we recognize it is important to include in the core concern of Digital Preservation (DP) the scope of Information Systems. For that we need to bring to discussion two important concepts:

- Non-Functional Requirement: In Systems Engineering and Requirements Engineering, a Non-Functional Requirement is a requirement that specifies criteria that can be used to judge the operation of a system, rather than specific behaviours. This should be contrasted with Functional Requirements that define specific behaviour or functions [5]. A Functional Requirement defines what a system is supposed to do whereas a Non-Functional Requirement defines how a system is supposed to be. Non-Functional Requirements are often called qualities of a system, which sometimes are also called "constraints", "quality attributes", "quality goals", "quality of service requirements" and "non-behavioral requirements“ [5].
- System Architecture: In Software Engineering the architecture of a system is defined as “the fundamental organization of a system embodied in its components, their relationships to each other, and to the environment, and the principles guiding its design and evolution”

[1]. Yet, a successful systems' architecture has to reflect the concerns and interests of the stakeholders [1]. A stakeholder is defined as a viewer that perceives and conceives the universe, using his/her senses, in order to produce conceptions resulting from the interpretation of what is observed. While observing the universe, a viewer will be interested in a specific part of the universe, also called a concern, and might zoom-in to that part of its conception of the universe, also called a domain.

The traditional concerns of the DP community have been related to scenarios where Digital Preservation requirements are the main Functional Requirements. That means scenarios where the main business purpose has been digital preservation in itself. As a consequence, the community was able to develop a relevant body of knowledge to build Digital Preservation Systems, as also to advise on scenarios where Digital Preservation Systems are expected to interoperate with other business systems. These are the common scenarios in Cultural Heritage and other areas where "Information Archival" is a well defined concept, like in the Science, Technology and Medical (STM) publishing industry.

However, we must ask now what should happen in scenarios like for example Management Information Systems, Engineering Systems, Health Care Systems, and other domains where strong functional requirements already prevail (Table 1). In these scenarios, where all the information of the systems needs to be permanently available and can be changing at a very high frequency, the concept of "Information Archive" makes little or no-sense. But even if digital preservation is here not the main purpose, we believe it also can be especially relevant.

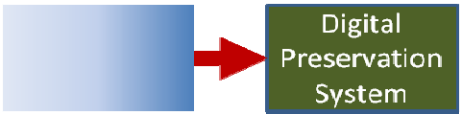
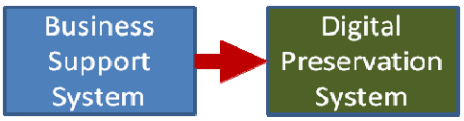
The "Digital Preservation System" (DPS): The business is digital preservation and the related concerns are transposed to functional requirements. The OAIS model was developed for scenarios of this kind.	
The "Systems of Systems" (SoS): The business is supported by a system that delegates the accomplishment of the digital preservation requirements in a complementary DPS. The OAIS model can be used in this scenario, eventually with compromises.	
The "Digital Preservation Ready System" (DPR): The business transposes the digital preservation concerns in non-functional requirements. So far, there are no identified relevant specific results from the Digital Preservation community for this scenario.	

Table 1: Scenarios of systems capabilities according to their digital preservation focus of concern

In these scenarios the digital preservation requirements are to be understood as Non-Functional Requirements. Therefore, the identification and elicitation of those requirements has to be properly addressed in order to add to these information systems new related levels of capability. We call those scenarios as "DP Ready", and we believe they can become a new specific area of research, with unknown challenges (Figure 1). In fact, it is where DP concerns will be not anymore the main purpose, but will have to compete side by side with other equivalent areas of concern.

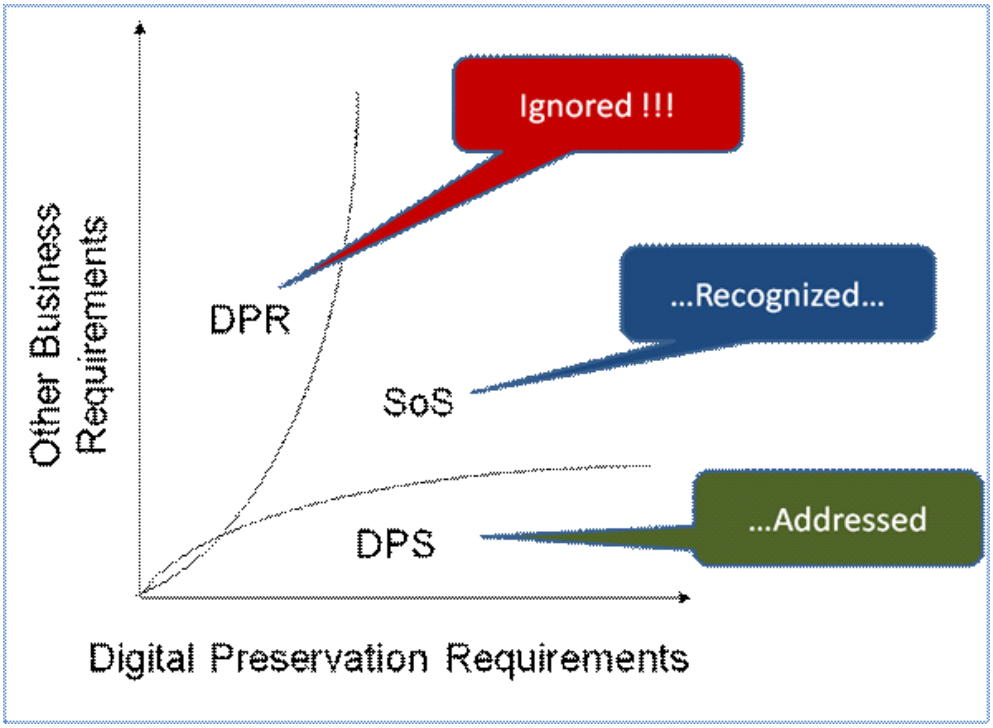


Figure 1: The traditional focus of the digital preservation community.

We can classify a system as "DP Ready" if we can prove that we effectively can, for that system, move the information base and the defined valid states' changes (the systems' behavioural schema, or business rules) to another instance of another information system. For this we need to bring to discussion one more important related concept:

- Enterprise Architecture is defined as a coherent whole of principles, methods, and models that are used in the design and realisation of and enterprise's organisational structure, business processes, information systems, and infrastructure [4]. An Enterprise Architecture framework is a communication tool to support the Enterprise Architecture process. It consists of a set of concepts that must be used as a guide during that process. Examples are the Zachman Framework for Enterprise Architecture [3] and The Open Group Architecture Framework, also known as TOGAF [2].

Concluding, there is a need for new knowledge to be produced (or motivated) by the DP community about how to give the capability of "DP Ready" to information systems in scenarios where DP concerns are perceived as Non-Functional Requirements! That must imply a concern with existing reference System Architectures and an involvement of the DP community with the Enterprise Architecture best practices.

Research Challenges

The research challenges to be identified related to this topic must contribute to the building of systems proving that they are "DP Ready (or in other words, that they are "resilient against change"). That implies to push for the consideration of DP as a new specific concern in Enterprise Architecture, with the purpose of giving to the systems new related capabilities. Consequently, examples of related research challenges are:

- What are the principle to use to identify cases and evidences of potential DPR scenarios where ignoring DP can be a threat or a cause of loss in Data, Information or Knowledge; or

a loss of Opportunity; or an Economic loss; or o degradation or even of loss of Life? What are the principles to use to prevent those scenarios?

- What are the principles that can assure longevity of the information/knowledge base of information systems, apart from those systems' primary purposes?
- We already have a fairly developed body of knowledge for the preservation of digital information. But what about the preservation of digital processes?
- How can we compare and align the actual DP references, such as OAIS, with other existing references already common in Information Systems environments, such as COBIT , ITIL (governance of IT infrastructures), MoReq (Records Management), etc.? Can we define, as a complement/alternative to OAIS, a kind of "MoRep-DP" (an ontology of generic requirements for DP to be taken in consideration in processes of Enterprise Architecture)?
- What strategic win-win scenarios and moves can the DP community envisage to approach and complement other relevant existing reference and standardization communities (such has for example the OASIS) with the DP concern?
- Research Question ...???...

Glossary

Architecture – The fundamental organization of a system embodied in its components, their relationships to each other, and to the environment, and the principles guiding its design and evolution [1].
Capability – An ability that an organization, person, or system possesses. Capabilities are typically expressed in general and high-level terms and typically require a combination of organization, people, processes, and technology to achieve [2].
Concern – A domain of interest (usually associated to at least one class of stakeholders).
Enterprise Architecture – A formal, highly structured, way of defining an enterprise's systems architecture [3].
Model – An abstract representation of a domain.
Stakeholder – An individual, team, or organization (or classes thereof) with interests in, or concerns relative to, a system [1].
System – A collection of components organized to accomplish a specific function or set of functions [1].
View – A view is a representation or description of the entire system from a single perspective. In contrast to a viewpoint, a view refers to a particular architecture of a system (i.e., an individual system, a product line, a system-of-systems, etc.). A view is primarily composed of models, although it also has additional attributes. The models provide the specific description, or content, of an architecture. For example, a structural view might consist of a set of models of the system structure. The elements of such models might include identifiable system components and their interfaces, and interconnections among those components [1].
Viewpoint - A specification of the conventions for constructing and using a view. A pattern or template from which to develop individual views by establishing the purposes and audience for a view and the techniques for its creation and analysis [1].

References

[1] IEEE Computer Society (2000). IEEE Std. 1471-2000: IEEE Recommended Practice for Architecture Description of Software-Intensive Systems. IEEE, New York.
[2] The Open Group. (2009). TOGAF Version 9. Van Haren Publishing.
[3] Zachman, J. (1987). A Framework for Information Systems Architecture. IBM Systems Journal, 26 (3), pp. 276 – 292.

[4] Lankhorst, M. (2005). Enterprise Architecture at Work: Modelling, Communication, and Analysis. Springer, Berlin/Heidelberg.
[5] http://en.wikipedia.org/wiki/Non-Functional_Requirements

Session 2: Beyond Metadata? Information Retrieval/Mining/Visualization/Context

(discussion facilitator Panos Constantopoulos, rapporteur Vassilis Plachouras and Michael Hartle)

In the breakout sessions, we dealt with four main areasof digital preservation, namely what to preserve, the nature of Digital Objects, the conceptual modeling of Digital Objects and Digital Preservation, and finally, techniques applicable to Digital Preservation. We identified three recurring challenges across the discussed areas of Digital Preservation. These challenges relate to the fluidity of preserved information, the preservation of context, and scalability issues in Digital Preservation. In the remainder of this draft document, we provide some initial insight on the identified challenges with respect to the discussed topics.

What to preserve

While knowledge continuously evolves, preservation in the past resulted in a fixed view of the knowledge. The nature of digital information, however, allows the fluidity of knowledge to emerge. In this case, digital preservation may deal with the challenge of preserving the processes and dynamics related to digital objects.
The definition of what represents the context of a Digital Object that should be preserved constitutes an additional challenge for Digital Preservation. The digital nature of information allows to record evidence of the intent of a Digital Object, its production process, as well as its usage. Other processes related to the lifecycle of a Digital Object may also be captured, and enhance the context of the Digital Object.
When deciding what to preserve, however, scalability also becomes a challenge, because the preservation of metadata and the context of a Digital Objects introduces a recursive preservation of layers of information related to the Digital Object. Such recursive preservation of layers of information increase the complexity of the preservation process and may have economic implications.

Nature of Digital Objects

The fluidity of digital objects affects their nature in the sense that a digital object may have volatile aspects, which may change over time due to versioning, or migration to a new meduim. Furthermore, a digital object may have aspects endangered of being lost, because it is possible that in the future accessing the corresponding digital object may be impeded by the lack of the relevant knowledge or expertise required to understand it.
The possibility to preserve both a Digital Object as well as its context poses a challenge regarding the definition of the nature of Digital Objects. The context of a Digital Object may correspond to one or more Digital Objects, which are preserved along the initial object. In such a case, we can consider that we have a network of objects. For example, a Digital Object and its annotations form a network of objects that may be preserved as such. In an alternative scenario, preservation of Digital Objects and their associated contexts may correspond to their decomposition into basic elements, and the preservation of a database of elements rather than the Digital Objects.

The capability to associate a Digital Object with its context, and preserve both in the form of a network of objects, poses a scalabilitychallenge, due to the consequent increase of the amount of data to be preserved.

Conceptual Modeling

Regarding conceptual modeling, the challenge of fluidityappears in the form of the evolution of ontologies. As knowledge advances, ontologies may become outdated, hence, requiring adaptation or the development of new ones. The evolution of ontologies of Digital Objects may be counterweighted by the more likely stability of an ontology of relations between Digital Objects. Considering the context of Digital Objects as a network of objects, it raises the question on the different connections between Digital Objects. Two objects may be connected with a relation, which represents a typed connection. A link corresponds to an untyped connection, or a relation with an empty type. Furthermore, a similarity measure can be applied to generate connections between highly-similar objects.
The context and scalability also depend on the diversity of topics and the size of the designated community. For example, it may be possible to develop an ontology for a a small designated community focused on a small number of topics. On the contrary, a large designated community with a diverse range of topics may not be able to afford developing and maintaining an ontology. Furthermore, diplomatics can provide a scalable approach to represent relations between objects based on the similarity of their form.

Techniques

The techniques and procedures employed to implement Digital Preservation are required to handle the concept of fluidity. As knowledge advances, the metadata associated with a digital object may change at a fast rate. Hence, Digital Preservation techniques must have provision to handle the observed rates of change.
The context of Digital Objects can be enriched with the application of techniques, such as the statistical assignment of annotations, and diplomatics. Regarding the statistical assignment of annotations, the context must preserve information regarding the details of the employed algorithms, as well as the confidence of the algorithm's output. Diplomatics may also be used as means towards the statistical characterization of digital objects based on elements of form. Decisions on whether to employ techniques based on extraction or collection of knowledge have implications on the scalabilityof Digital Preservation. Techniques such as the statistical assignment of annotations mentioned above may be automatically applied to a large number of Digital Objects. Similar advantages may be obtained from the application of diplomatics in a statistical framework.

Overview

Next, we provide an overview of the four main areas of discussion and the identified challengesin the form of a table.

Areas Challenges	What to preserve	Nature of Digital Objects	Conceptual Modeling	Techniques
Fluidity	Process/Dynamics	Volatile/endangered parts of Digital Objects	Evolution of ontologies Ontology of relations	Rate of change in metadata

Context	Intent Usage Production Other processes	Network of objects - annotations Preserve DO vs.DB - decomposition of DO into elements	Ontology relations - Similarity - Linking - Relations Diversity/Community	Statistical annotation assignment Diplomatics
Scalability	Recursion of preservation	Network of objects Preserve DO vs.DB	Diversity/Community Diplomatics	Extraction/ Collection of Knowledge Diplomatics

Indicative research questions

In all cases below it is in addition required to: (i) define an appropriate metadata set for preservation purposes or demonstrate a mapping onto an existing metadata set; (ii) define the models and the metadata in such a way as to enable either minimization or streamlining of human input.

- Define a modelling framework for capturing the intended and the actual usage of digital objects, such that this may be documented in the course of creating or using the object respectively. A suitable classification scheme of usages is needed.
- Define modelling frameworks for specifying or describing processes of producing a digital object.
- Define modelling frameworks for specifying or describing processes of using a digital object.
- Develop a framework for representing situational collections of digital objects. These collections may be unstructured or they may have a structure induced either by untyped links or by specific relations, giving rise to networks of objects. Both the composition and structure of a collection may vary with time. An ontology of relations between digital objects is needed.
- Develop diplomatics-driven semi-supervised methods for metadata extraction from and annotation of digital objects.

Session 3: Storage Technologies and Protocols

(discussion facilitator Rudolf Gschwind, rapporteur Vassilis Plachouras)

Potential Research Topics and their Use Cases

In the current state, this document is still a draft, therefore incomplete, and still likely to contain errors and numerous misunderstandings of the Rapporteur. At the same time, it may contain potentially innovative ideas. Use at your own risk.

Self-Sufficiency

Self-explaining, -correcting, -replicating code

When exchanging a digital item as data, it has been always necessary to "explain" such data in some way to receivers, both in terms of its data format and the underlying semantics. These additional pieces of information again form digital items, leading to a highly problematic recursion regarding DP efforts.

Adding properties such as "self-correction" and "self-replication" marks a change **from passive to active data**, although it remains totally unclear how, if at all, to achieve such properties in general terms. Similarly, it can be assumed to be impossible to achieve completely, but a sort of "self-explanatory code" would be interesting, as it would provide for a **"recursion anchor" to the problem of recursive preservation**. This requires the **exploration of factual limits of self-**

explanation, be it for the use of **embedding textual or XML-based descriptions** into binary data, by trying to make it easy to infer meaning by explicitly **providing "help" to something like a cryptoanalysis in the future**, or something else entirely. Ideally, such a description is **multi-lingual** and **multi-coded**, potentially allowing the **inference of meaning** between multiple descriptions present in different languages (similar to the Rosetta stone which provided key clues to reading Egyptian hieroglyphs). Moreover, the descriptions should use **"semantic anchors"** that refer to topics which are likely to survive (human-based properties, mathematical or physical constants) and which are to be used redundantly in descriptions. A major difficulty regarding this research topic is the issue of its evaluation; potentially, evaluating the issue of self-explanation may be adequately substituted through a property that lends itself more to an evaluation.

Self-contained, embedded devices for writing, querying and reading contained data
Rather than depending on a complex and changing ecosystem of hardware and software to provide for writing, querying and reading a certain type of data, it may be interesting to have a self-contained, embedded device which acts as an interface for humans, and which provides these services in a closed manner.
Not directly a research topic on its own, this may nevertheless be of interest, and also includes development and engineering topics. A central aspect is the autonomous maintenance of data integrity within the device, ensuring that data is constantly checked and its detected health reported to the user.

Imperfect Digital Preservation

Perception-based graceful degradation in DP
Not every pixel in a digital photograph matters – some loss of information in audio-visual content may result in none or only a small perceived loss of information for a human audience. Ideally, a digital item provides for a perception-based graceful degradation in the first place.
This requires models for measuring the perceived loss of information for video and audio content, either based on or quite similar to existing psycho-acoustic models for the compression of audio data. Based on such models, this raises the question of how to design file formats with perception-based error resilience / graceful degradation in mind. Ideally, not only is the audio-visual content represented in a suitable fashion, but also those sections of bit streams that are of structural importance for processing. In a later stage, it may be interesting to consider whether it is possible to transpose perception-based graceful degradation from the audio-visual content to symbolic data beyond text-based summarization. This may potentially provide a means for answering which portions of a digital item can be deleted more easily.

Enabling Longevity through Redundancy
When longevity of data has to be ensured, redundancy is key – not only redundancy in the actual representation based on Coding Theory, but also through spatial redundancy by storing data at different locations. To ensure redundancy for very large amounts of data on the scale of petabytes, potentially stored on a large number of disks (10.000+ units) in a distributed scenario, said redundancy has to be managed automatically in a transparent fashion by an autonomous system, inspiring references to the Google File System and related approaches.
Ideally, such an approach for transparent management of redundancy on various levels has a mathematical foundation and thus allows the simulation of different parameterizations including measurements of key metrics, also including aspects such as Total Cost of Ownership (TCO), bandwidths for simultaneous archival and restoration during live operation, or Quality of Service

(QoS) parameters. The research topic may potentially yield something like a “Statistical Preservation Theory” in this context.

Summarization and Forgetting in DP
Not every piece of preserved digital information is equally important. Assuming a database of Walmart including each and every sales transaction, the loss of an individual transaction may be acceptable, as long as the overall proportions of sales transaction do not become grossly misrepresented.
In this regard, aspects such as summarization and aggregation are of importance. Likewise, on the other hand, it is necessary to prioritize information in order to select information that may be forgotten “more easily” if loss can potentially be steered. Although unclear, somewhat contradictory and potentially impossible, it may possibly even be desirable to “forget data for a while”, for example when data first loses its commercial importance, only to gain historical value at a later point in time.

Information Handling

Secure, gradual release of confidential information over time
It is sometimes necessary to ensure that sensitive information is accessible only after a certain period of time, often due to a historic interest. Practical examples are confidential documents providing background on political decisions, which cannot be released in a timely manner to protect strategic national interests, or sensitive documents containing personal information on a donor, where a timely release invades the privacy of the donor. The conventional approach to this problem is to use a trusted third party, a custodian, who releases such information after a designated period of time. Rather than simply trusting the security of a custodian, it is desirable to actually ensure that confidential information is represented in such a way that confidentiality is provably maintained over a defined period of time, and rescinded afterwards.
Ideally, a suitable solution both guarantees a) the confidentiality of protected information before the intended date of release, and thus withstands “time-travel attacks” where an attacker tries to manipulate references to the current time, substituting it with a later point in time to gain early access, and b) the availability of protected information after the intended date of release, and thus withstands “denial-of-service attacks” where an attacker tries to provoke a prolonged or permanent protection of information. This research topic has close ties with Cryptography, potentially depends on the use of Trusted Computing infrastructure, and potentially has to handle the preservation of keys and cryptographic methods.

Provable deletion of information
Proving the deletion of a digital object to a third party can become necessary due to a multitude of reasons: due to legal obligations, to guarantee that the right, unaltered digital object was deleted, or to ensure that the authoritativeness of a migrated digital item is not undermined by a left-over original.
Actually giving proof of a deletion itself is a non-trivial task that may depend on the support of technical infrastructure known from Trusted Computing, and has to cope with issues such as non-erasable media. It is unclear whether the deletion of a digital object can actually be proven; it may or may not be necessary to explicitly keep track of copies, possibly even through Digital Rights Management (DRM) systems. For guaranteeing the deletion of the right digital item, and to prevent the accidental deletion of wrong digital items, the use of digital signatures may be necessary as well.

Storage Technologies

Visualizing magnetization of magnetic media for DA

When hardware for reading magnetic media becomes obsolete, access to these media is lost. In an effort related to Digital Archaeology, it is desirable to have a method for visualizing the magnetization of magnetic media, thereby switching media access from the visual into the magnetic domain. Potential means of realization may involve magnetic liquids, magneto-optical effects and other suitable approaches from physics or chemistry.

Enabling actual long-term storage technology

Current storage technology is optimized for aspects such as access speed and throughput, but not for long-term storage of data. For DP, it would be attractive if it was possible to dynamically decide for a suitable storage technology whether to prioritize access speed over durability of storage or vice versa. As an analogy, today's projectors typically have at least two different modes of operation, which either allow the projector to be run in Eco mode with reduced brightness, but with an increased lifespan of the internal lamp, or running in standard mode, delivering full brightness at the expense of reduced lamp lifespan.

Total Cost of Ownership (TCO) models for comparing storage technologies in DP

Different types of storage technology have varying Quality of Service (QoS) parameters and Total Cost of Ownership (TCO). In order to properly decide in favor of a specific storage technology, it would be helpful to have actual models for simulating a specific DP setup, for computing both the resulting TCO and QoS parameters that can be expected, and for comparing different setups.
As a by-product, adequate models may also help to test alternative setups (eg. increasing redundancy, increasing throughput, increasing storage capacity), evaluate resulting properties over time (eg. risk of data loss) and to optimize certain properties (eg. initial cost, maintenance) of a setup without increasing the risk of losing data to be preserved. Regarding this research question, close cooperation with Economics and from the domain of Mathematical Modelling may be beneficial.

Distributed Systems and Protocols

Distributed peer-to-peer archival system

For DP, it is desirable to ensure that data to be preserved is stored in a redundant manner, similar to what LOCKSS is doing. At the same time, there is a progress on storage technologies in terms of lower cost and increased storage capacity, where there is a need to hide the heterogeneity of individual storage nodes. In such a complex, dynamic environment, it is desirable to have a distributed system which can handle the addition and removal of nodes with storage capacity, and which manages the redundant storage of data according to policies (advancements relative to LOCKSS should be detailed more clearly).
The problem of distributing contents over large networks has been successfully addressed in the domain of Peer-to-Peer (P2P) networking, and thus lends itself for an automated, distributed version of LOCKSS using P2P technology. Automatic distributing of data and management of redundancy is desirable to minimize the effect of human errors during operation. Moreover, defining a suitable protocol has the additional benefit of encouraging multiple different implementations. Practical examples are the Google File System regarding its distributed character, and the "Time Machine" backup system from Apple regarding its simplicity. A suitable approach should provide a rigorous mathematical foundation.

Synchronisation of distributed storage data in case of bit errors

When data is stored in a distributed, redundant fashion, loss of data is bound to occur on individual nodes in the system, either due to disk failures or individual bit flips triggered by Cosmic Rays. Given that such a distributed system may easily carry data in the range of Petabytes, nodes need to be able to mutually correct these errors without actually being able to exchange large portions of data.
In part, this is a problem of synchronizing data in different storage locations which has been addressed by protocols such as rsync and others, yet these have not yet been considered in the context of multiple nodes a P2P system or for large volumes of data in the scale of Petabytes.

New Frontiers

DNA as Data Carrier

Biological life itself can be considered as a role-model for DP regarding the preservation of data, since data in DNA has been "preserved" successfully for millions of years. There are potential opportunities to learn how life itself has solved the preservation of DNA-based representation of information, where interesting properties such as self-repair are built-in. Creating artificial life, for example by constructing information-carrying bacteria which act as custodians of their in-built DNA, may be a way of storing data in a way which caters for its active repair. On the other hand, maintaining the authenticity of such data in terms of mutations is entirely unclear.
In this still quite speculative area, there are numerous opportunities for cross-domain cooperation with Biology and Bio-Informatics, and this will require actual supervision through experts from these domains.

DP in the age of Quantum Computing

Due to the advent of Quantum Computing (QC), the question arises whether QC is going to have effects on DP itself, e.g. regarding the preservation of QC programs, or preserving the result of QC computations. In contrast to other research questions, this is highly speculative in its nature, as it is unclear whether such effects are actually present.

Autonomous Agents as Data Carriers

Computer viruses can be considered as prototypes for artificial life, which focus on their replication, thus ensuring the redundancy of their code as data. Applying this concept to DP, the idea is to utilize autonomous, distributed agents to preserve portions of data, similar to biological cells. Through coordinated operation and replication, these agents are to ensure the redundancy and accessibility of entrusted data, turning this into a hard problem for research.
As well, this idea is highly speculative, and brings a number of interesting legal and ethical problems of its own.

Session 4: Policy and Rule Management

(discussion facilitator Jens Ludwig, rapporteurs Jens Ludwig and Cal Lee)

Implementation of policies and rules is important to digital preservation for a variety of reasons. They describe how digital preservation is ensured by an institution and digital preservation systems must realize them. It is the current state of the art to explicitly formulate policies and rules and express them seperately from the underlying system in order to make them more manageable, adaptive and auditable. Digital preservation policies and rules also provide important

contextual information for evaluating the authenticity of preserved objects and for communicating service expectations and requirements between stakeholders.

For making progress in the area of digital preservation, the following research questions about policy and rule management have been identified:

- A variety of different sources and methods is conceivable to derive policies and rules for digital preservation, e.g. they can be based on digital preservation or domain specific standards, theoretical works, human-language expressions of needs statements/requirements or higher-level goals. Often no explicit policies have been formulated (or in the case of data archaeology they may have been lost), and a derivation based on observations and measurements of system and human behavior might be useful. What are the resulting policies and rules and how do these different methods compare to each other? How can these methods be refined or supported (e.g. interactive donor agreements)?
- Policies are often the result of negotiations between stakeholders and are used to communicate intentions. Can policies play effective roles as boundary objects in decision making and can policies and rules serve as sufficient documentation? Can policies effectively model (and can rules reflect) the results of negotiations? Can policies/rules support negotiations between stakeholders? What is the minimum set of decisions that one needs to make in order to specify the policies for a digital preservation environment?
- Many variables influence policy creation. It would be very useful to create several policies for different exemplary institutions. These policies could be tested in order to answer a variety of questions. What are the differences between preservation policies and rules for dedicated preservation environments and those for production systems with preservation capabilities? Are domain-specific languages required or are generic policy/rule languages sufficient for digital preservation purposes? How content-specific do the policies/rules need to be? What are the categories of policies that are required at different levels of granularity (e.g. for specific agents or collections)? How can one express and determine different compliance levels related to implementation of a given policy?
- The expression and language of policies and rules for digital preservation becomes particularly relevant if one wants to achieve automation and machine-actionable rules. How should policies and rules be expressed in order to facilitate the definition and automation of preservation processes? Can we develop controlled vocabularies or even ontologies for expressing policies? What should be the level of granularity of entities, and what should be the atomic operations? One answer could be that an entity is needed whenever a measurement needs to be performed.
- Policies and rules need to be managed continuously because they will change and evolve with their environment. How should policies that change over time be managed and versioned so that they can be appropriately associated with the digital objects of the time (e.g. the security measures of a collection at the time were so low that the user may have less trust in the objects in the collection)? How can this process be supported? Can policies and rulesets be found which are resilient to change?
- Policies and rules are highly complex, and the results of their applications may not always be evident in advance. We would like to reason about them and ideally prove that they are consistent and complete. This requirement can become very important if different institutions with different policies and rules have to cooperate. How can policy conflicts which prevent successful preservation be detected and resolved? How should policies and rules be formulated so that we can reason about them?

Session 5: Ethics, Privacy, Security and Trust

(discussion facilitator Cal Lee, rapporteur Yunhyong Kim and Cal Lee)

Research Challenges	**Systematically Formulating as DP Research Problem**	**RQ re: How to Respond**
Controlling collection and access	Modeling different categories	How can we model and detect different categories and level of permissible collection and usage?
Controlling machine-interpretability		How to develop various representations of digital information with different levels of machine-interpretability
Applying cryptography in long-term DP situations	Advantages/disadvantages for: • Ensure integrity • Control access	Methods for: • Ensuring integrity • Controlling access
Integration of security measures into DP activities	Advantages/disadvantages of doing so	Means of doing so • Standard digital signature models do not account for key management over long periods – would need to keep timestamps of key succession over time • Security strategies (e.g. user roles, access rights) to be applied over very long time periods? • How can one reconcile the rights of individuals across accounts and systems?
Characterize levels of trust in a system in terms of assurance of digital object integrity	What constitutes a trustworthy system?	Expressing specific levels or types of trustworthiness related to specific assurances (more fine-grained and less binary than e.g. TRAC-compliance)
Identify “original” among multiple instances of digital objects (e.g. pictures)	What constitutes an original for a given context/purpose	

Data leakage	Modeling forms of data leakage at various levels of representation	• Defining and mitigating specific data leakage risks • Pruning out or redacting underlying data (e.g. embedded data in office file) – requires assigning semantics to parts of bitstream or digital object
Managing sensitive data	Modeling forms of confidentiality violations at various levels of representation	Defining and mitigating specific confidentiality risks • Detecting sensitive data using NLP, machine learning
Forgeries of data and renderings	• What are the DP requirements for identifying and detecting forgeries and how to detect evidence of forgeries? • Modeling types of forgeries at various levels of representation	• Defining and mitigating forgery risks • Detecting evidence of forgeries o Inconsistencies IP addresses and domains o Embedded object metadata, timestamps o Comparison of data across sources o Comparison of other factors across collections
DP-Aware Digital Rights Management	What is to be preserved about DRM?	How to develop DP-aware DRM (functional requirements)
Inappropriate reuse of intellectual property	Modeling types of cases	Detecting instances
Aggregate-level concerns	What are ethical considerations related to aggregates that don't emerge with individual items (e.g. data mining, data correlation)	
Consent involving more than one party, including one party acting on behalf of another • Cultural ownership arrangements • Heritage arrangements • Cases involving minors		

Ensuring forgetting	Modeling forms of forgetting at various levels of representation	• Overwriting media and data removal
Controlled data release	Modeling forms of release at various levels of representation	Controlled release regimes based on: • time periods • trigger events • use context • documentary context (associated digital information that is available)

Session 6: Evaluation and Benchmarking in Digital Preservation

(discussion facilitator and rapporteur Christoph Becker)

Potential Research Topics and their Use Cases

In the current state, this document is still a draft, therefore incomplete, and still likely to contain errors and numerous misunderstandings of the rapporteurs. At the same time, it may contain potentially innovative ideas. Use at your own risk.

Address necessary expertise for each Research Question

Motivation

Several key issues provide the foundation for the urgency of measurements...

Digital preservation will have to scale down and up substantially in the near future. These goals of scalability require reliable, repeatable and efficient measurement of the decision factors that underpin all DP and PP operations, such as

- Desired properties of digital objects,
- Properties of digital objects that have to be kept through changing representations and environments, Properties of object formats and other representation information networks, and
- Operational properties of systems and processes.

Can we just measure less and use approximation as a stochastic approach? Simply reducing the coverage of measurements to a level that is practically feasible on the technical level is not sufficient in many cases where high requirements are posed on trust and authenticity. In many scenarios, legal requirements impose strict constraints regarding authenticity and verification of processes to avoid litigation. This implies that we need a high degree of trust also in the measurements used in the course of the process.

Another crucial aspect to consider is that in practical terms, there can be no competition without a metric and no marketplace without comparability and no quality assurance without evaluation.

A number of approaches have addressed the second aspect. Practical experience indicates that instead of fundamentally canonical approaches to ensuring authenticity, which encounter tough challenges hidden in the small but abundant complexities and variations of format implementations, a more pragmatically viable way is to define a roadmap of aspects that need to be measured and address them on a prioritisation basis.

Some observations on objects, environments and dependencies

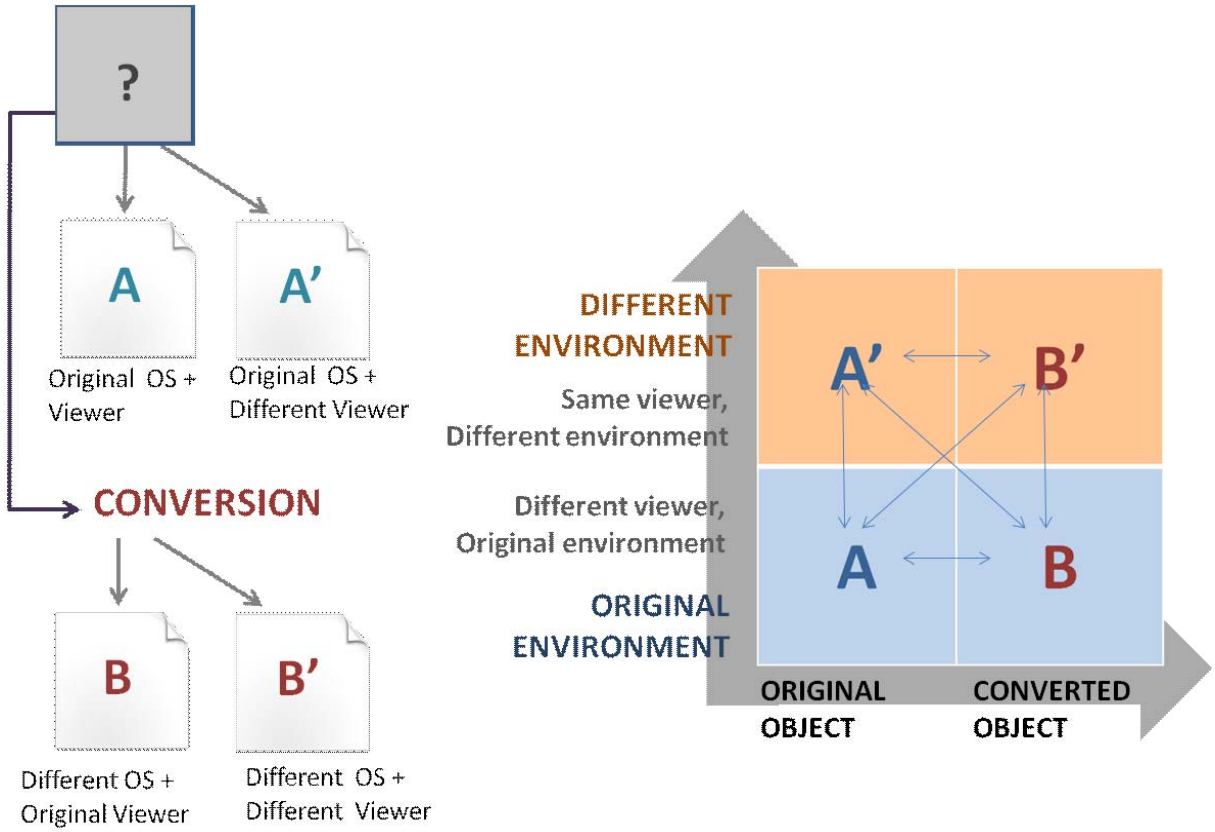


Figure 1 Objects and environments

Considering the digital ecosystem as the object and its environment (in the sense of the set of artefacts it depends on, which in turn show dependencies), for instance

- the feature set / look and feel of games on different machines,
- rendering of text documents on systems with varying font configurations
- ...

Prior to exposure to an interpretation, the properties of any object are entirely unknown.
It is questionable whether the distinction in these two dimensions holds for certain classes of objects and environments (think Mashups....).

Research Challenges

The first issue that naturally arises is the notion of subjectivity or the differentiation between objective measurements and the subjective value that corresponds to a measurement. Can there be an objective characterisation of objects at all? How can we achieve it?

The first and central building block of evaluation being the specification of criteria, several questions arise.

- How can we decide on the completeness of a set of criteria at a certain level?
- How can we address the fact that significance of certain criteria may change over time and that new criteria will be added to the set in the future, e.g. when the context of usage changes, the user community shifts, etc.?

Properties modelling

RQ: Can we define classes of objects concerning their properties relevant for DP? How can we describe this in a way that is preservable in itself?

- How can we describe an object according to these properties?
- What is the minimum set of properties needed to define a similarity metric between two objects of a certain class?
- Given an object and its set of properties, is it possible to automatically discover new properties? (reasoning, ML, IR...)

Measurement devices as sensors, filters, ...
Substantial lack of coverage of existing measurement devices (% rate!)
Criteria distribution... motivation to focus on object properties

Measurements and calibration

RQ: Suppose we have a measurement device for comparing two objects. How can we know that the measurements are correct?

- formally proof the correctness of the method used in the device
- multiple devices, pick a second (third) method - statistical approximation
- self-correction and adjustment over time
- How can we achieve (and reflect) confidence in a measurement?
- How can we model uncertainty in measurements and handle it...?
- If there is no specified ground truth, we do not have a reference frame to facilitate calibration
- Is it feasible and sufficient to employ crowdsourcing for approximating and/or verifying measurements for certain criteria?

Benchmarking and ground truth

Any improvement on the coverage and precision of measurements is doomed to fail if it cannot rely on substantial benchmarks and well-known ground truth to be validated against.

RQ1: Define an annotated benchmark set and metrics to compare different measurement devices in scenario X
RQn: Define a benchmark to compare preservation (-ready) environments

The building blocks needed for benchmarking include

- A representative sample distribution
- A probability distribution ...
- A second order change probability distribution of the change rate of the object area
- Evaluation criteria and metrics
- Ground truth

An essential barrier so far and a main reason why despite repeated efforts, there is no benchmark corpus yet available for any DP scenario, lies in the fundamental problem of the object as a black box that is only meaningful within a certain environment: If we search for real-world sample objects and characterise their properties ex-post to define the ground truth to associate with a benchmark set of objects, we need first to have confidence in the accuracy of these properties.

Further considering ground truth and validation, an interesting analogy arises when considering QA in industrial production processes. In these scenarios, products are automatically tested for conformance to a specified set of criteria. The substantial difference in the case of DP is the fundamental lack of such specifications on the level of specific objects and processes.

RQ: Is it feasible to create sets of reference objects for certain classes that adequately represent their class?

- We need a mechanism for describing the desired set of properties (cf. Above)
- We need a mechanism for generating objects from this description
- Create a repository of reference objects for certain classes
    - Collect and filter
    - Specify and search

RQ: Based on a set of extracted properties, can we construct a minimum representation (a minimally simple test object that contains all properties)?

- Describe input
- Minimise properties
- Generate output
- MST etc.

While in this case we are looking for sets of objects to use as a reference point for verifying the correctness of measurements, on a different level we may strive to incorporate reference aspects or components into objects and/or processes.
In photography, colour cards are sometimes included in the scenery of the picture taken so that they are embedded in the photograph as a reference point and can be used for verifying the accurateness of colour representation, luminance etc.

Similarly, sample behaviour of games may be a reference point for an emulator, etc....

RQ: Can we create the equivalent of a colour card as it is being used in photography, to embed in objects and/or processes for quality assurance?

- What are the desired properties of such an artefact?
- How can we generate these and embed them?
- Can we define preservation processes that can take them into account?

Can we create links between extracting the semantics and extracting the content/formatting...

... like a checksum used in communication protocols

Preservability and risk assessment

Intuitively, the preservability of an object depends on

- Its properties and their relevance - what is the set of significant properties?

- The dependency of the object on an environment or other objects for being properly rendered
- The internal structure, complexity...

For each of these aspects, certain threats apply that comprise risks to preservability to be handled. Preservability as multi-dimensional property thus is directly related to risk assessment.
RQ: How can we define a (exhaustive) list of factors that influence the preservability of a certain object?
RQ: Can we define a unified scale of risk assessment and/or preservability of a certain object?

Supporting quality assurance

RQ: How can we enrich processes (production, ingest) to produce information that supports present and future QA?

- What is the minimum set of properties that have to be defined?
- for processes e.g.: context, test data e.g. for services/processes... Specifications, boundary conditions.
- Given processes and services in a system with known purpose, function, boundary conditions, test data, expected outcomes. How can we address aspects such as dependency management and change propagation over time while securing preservation requirements?

living system... lifecycle model .. include dynamics, change, change rates of a system...

Human experience and reference points

The challenge of preserving human experience is a substantial one. In theory, it would be desirable to formally express human experience reference points, linking emotions, intent, provenance, process and perception. This is a question addressing a number of fields outside core computer science research, such as media science, social science and cognitive science.

Provided such questions are addressed, can we apply perceptual hashing approaches and tools to create these reference points, and can we build perceptual hashing tools once we know the reference points?

On a more computer-science oriented viewpoint, we can focus on systems that are simpler than humans:
RQ: Can we model the behaviour and interactions between systems in such a way that allows us to assess faithfulness of preservation of a piece of/or a system?

Feedback and Learning

RQ: Create a preservation (-ready) system that incorporates (user) feedback about preservation quality and learns from mistakes to improve quality

- Can we feasibly incorporate user feedback into access modules and feed that information into an improvement cycle?
- Can we address the relation between objective measurements and subjective judgement using such crowdsourcing and approximation?

“Machine learning” for DP: Be able to handle and select from multiple strategies, redundancy …

Session 7: Application Domains

(discussion facilitator and rapporteur Andreas Lange)

Given their complexity, computer games can be seen as indicators for future developments in digital preservation in a number of application domains. Some of the research challenges given by computer games are given here below.

1)Games are application, whose cultural value consists to a good portion in the way the user interact with it.
RQ1: How can historical interactions/ user behavior be preserved?
RQ2: How can the applications be preserved, that future users can be enabled to interact with the application environment in a way the historic users did?

2)There is a clear tendency, that games were distributed more and more in parts rather than in one application. Also the service based model, in which the application is running on a server and is sending only a video stream over the internet back to the user, went into business recently.
RQ1: Are we facing a turn from a download model towards a service based model of distribution?
RQ2: What strategies must be developed to preserve artifacts, which are distributed in such a way.

3)DRM systems become more and more worrying for archivists, because the archivist is depending on factors he/she cannot control in the moment. E.g. one of the latest DRM trend is, that the game/ application is only running, when it is connected on-line with a special dedicated server from the IP holder.
RQ: How can DRM systems be developed, which satisfy the needs of the IP holders to protect their IP and offers special features to allow preservation, too?

4)Considering the capabilities of mobile devices, mixed/augmented reality games applications will become more important.
RQ: How can augmented/ mixed reality environments be preserved?

5)Since the early days of the Web, the gamer community set up great resources of meta data and game archives on a user generated basis. This could become an example, which might be applied to other domains as well.
RQ: How can user generated content be used in DP on a systematical and institutional level?