Addressing Digital Preservation: Proposals for New Perspectives

José Barateiro*, **
jbarateiro@lnec.pt

Gonçalo Antunes*
goncalo.antunes@tagus.ist.utl.pt

José Borbinha *
jlb@ist.utl.pt

*INESC-ID, Information Systems Group, Rua Alves Redol 9, Apartado 13069, 1000-029 Lisboa, Portugal
**LNEC – National Laboratory for Civil Engineering, Av. Brasil 101, 1700-066 Lisboa, Portugal

ABSTRACT
Digital preservation aims at maintaining digital objects accessible over long periods of time, ensuring the authenticity and integrity of these digital objects. In this paper, we propose three different approaches to address the digital preservation problem. First, we survey the main requirements specific to the preservation arena. Next, we show how digital preservation can be approached as a specific case of System of Systems Engineering. Then, we introduce Enterprise Architecture as a framework which is regularly used to assist information systems design and maintenance, but can also be applied to System of Systems and consequently to digital preservation. Finally, in such complex environments, Risk Management is a key factor to assure the normal behavior of systems along time. Thus, we propose a Risk Management based approach to design and assess digital preservation environments, enclosing the definition of context and requirements, and the identification of threats and vulnerabilities, to be used as the basis of the definition of actions to deal with the risks associated with those threats and vulnerabilities. We generalize and survey the threats, vulnerabilities and techniques that can be applied in the scope of digital preservation.

Categories and Subject Descriptors
H.3.7 [**Digital Libraries**]: Systems Issues; H.3.4 [**Systems and Software**]: Distributed Systems

General Terms
Digital Libraries, Digital Preservation, Dependability, Data Grids, Interoperability

Keywords
Digital Preservation, Digital Libraries, Risk Management, Systems of Systems, Enterprise Architecture.

1. INTRODUCTION
Digital preservation aims at ensuring that digital objects remain accessible to users over a long period of time. This issue is gaining increasing attention, and important standardization efforts, such as the Open Archival Information System Reference Model (OAIS) [1], and the Preservation Metadata: Implementation Strategies (PREMIS)[1] data dictionary for preservation metadata, have contributed to its solution.

The Institute of Electrical and Electronics Engineers (IEEE) defines interoperability as the ability of two or more systems or components to exchange and use information [5]. In fact, digital preservation stresses the time dimension of interoperability, focusing on the requirement that digital objects must remain authentic and accessible to users and systems over a long period of time, thus maintaining their value.

In this paper we are proposing to bring into consideration three new perspectives that might increase the relevance and effectiveness of the "digital preservation" thinking. Thus, we propose to look at the digital preservation problem from a System of Systems Engineering (SOSE) perspective, an Enterprise Architecture perspective and also a Risk Management approach.

The emerging area of System of Systems Engineering rose from the need that current and future environments require capabilities and outcomes developed through the integration of existing legacy systems with potential new components or systems that provide the desired capability [14]. Thus, considering digital preservation as communication with the future, moving digital data to new choices of technology [15], we promote the idea that one must give more emphasis to the integration of efforts and body of knowledge between the digital preservation arena and the more generic area of System of Systems Engineering, where important concepts, regulations and best practices have emerged.

[1] http://www.loc.gov/standards/premis

The continuous growth of applications, components and size of collections, along with technology evolution (hardware, software, communication protocols, etc.) strongly increases the complexity of modern and future systems. This factor is crucial for digital preservation, since supporting solutions must evolve to be able to communicate with future environments. Model-Driven approaches such as the Model Driven Architecture[2], or Enterprise Architecture concepts to align organization's processes and activities with System of Systems development and maintenance, can be used to manage such complexity.

Based on the ideas proposed in [16] and [17], we motivate the use of Enterprise Architecture Frameworks to design and maintain digital preservation solutions, looking at a digital preservation system as a special case of a generic information system embodied in specific enterprise contexts.

In order to achieve the goals of digital preservation, repositories must "protect" digital objects against several threats that can affect their future interpretation. Actually, protecting digital objects against threats is equivalent to reduce the risk of those threats, which is the main goal of the broad area of Risk Management [10], which is also a challenge of System of Systems Engineering [18] [19], due to the system evolutions in the System of Systems context [20].

This paper also proposes a Risk Management based approach to design and assess digital preservation solutions, enclosing the definition of the context (e.g., digital preservation requirements), the identification of threats and vulnerabilities that may affect the achievement of the requirements to digital preservation, and the proposal of techniques to address the risks associated with those threats and vulnerabilities.

Even though this process must be optimized for any particular digital preservation scenario, in this paper we generalize this approach by proposing a taxonomy for digital preservation threats and vulnerabilities, and identifying a set of techniques that can be used in digital preservation systems to reduce the consequences of the identified threats and vulnerabilities. Moreover, we evaluate the application of those techniques, regarding the taxonomy of threats and vulnerabilities to digital preservation. We claim that the proposed threats, vulnerabilities and techniques can then be further used as the basis to design new digital preservation environments.

The remainder of this paper is organized as follows. Section 2 describes the specific requirements of digital preservation. In section 3 we present the vision of digital preservation as a special case of System of Systems Engineering. In section 4, we motivate the use of Enterprise Architectures Frameworks for digital preservation. Section 5, describes the Risk Management approach to digital preservation, including a taxonomy of threats and vulnerabilities to digital preservation and a proposal of techniques that can be used to address the identified digital preservation threats and vulnerabilities. Section 6 illustrates how the proposed perspectives can address digital preservation threats and vulnerabilities. Finally, in section 7 we list the open issues and conclude.

[2] http://www.omg.org/mda/

2. DIGITAL PRESERVATION REQUIREMENTS

Digital preservation combines policies, strategies and actions to ensure that digital objects remain authentic and accessible to users and systems over a long period of time, regardless the challenges of component and management failures, natural disasters or attacks [3].

Even though, it is impossible to define all the requirements applicable for all digital preservation needs, since digital preservation requirements depend, for instance, on the type, size and amount of data. It also depends on the goals of each organization, regarding the reuse of data. However, there are several generic and common requirements that can be surveyed, based on what someone in the future would require from information stored today.

First, digital preservation requires that a copy (or representation) of any preserved digital object survives over the system's lifetime, which is usually unknown, but may be as long as decades or even centuries. This can be defined as a **reliability** requirement. Therefore, a digital preservation system must be designed to store data indefinitely without suffering any data losses.

Second, a future consumer should be able to decide if the accessed information is sufficiently trustworthy. Usually, this requires the **authenticity assurance** of digital objects (which is already a common requirement for tangible objects). Also, the **provenance** of digital objects should be required, especially its creator or entity responsible for it. Moreover, it is crucial to assure the **integrity** of digital objects, guaranteeing that their informational content was not modified.

Third, digital preservation requires that future consumers are able to obtain the preserved information as its creators intended, **dealing with obsolescence** threats. This requirement encloses several challenges, since a digital object, to be explored, requires a technological context defined by specific software and, in some cases, even by specific hardware [3].

Finally, dynamic collections and environments for digital preservation require technical **scalability** to face technology evolution allowing, for instance, the addition of new components through incremental updates [1]. Existing static collections (with a fixed size) like, for instance, a digitized historical archive, where no new items will be added, will have a fixed data size. Although it will not be necessary to add new components to increase the storage capacity, it may be necessary to replace components by others with more recent technology (in order to achieve lower maintenance costs or simply because the initial technology was disrupted). This also implies a requirement for supporting **heterogeneity** (which is reinforced by the requirements for scalability).

Fortunately, some typical requirements of normal storage systems are not crucial in digital preservation. For instance, data updates are uncommon because, usually, objects in digital preservation systems are supposed to remain unchanged. Almost all write accesses to the repository are to either ingest new objects or re-write the exiting objects in new migrated formats.

In the next sections, we propose three perspectives that might help in surpassing the challenge of digital preservation. All three

perspectives are guided by the requirements presented in this section.

3. THE SYSTEMS ENGINEERING PERSPECTIVE

This Section describes how digital preservation can be seen as a specific case of the broad area of Systems Engineering. First, we introduce the concepts of the Systems Engineering field, with a strong focus on System of Systems Engineering. Second, we illustrate our vision of digital preservation as a case of Systems Engineering.

3.1 Systems Engineering

The International Council on Systems Engineering (INCOSE)[3] defines system as an integrated set of elements to consummate a specific objective. These elements may include hardware, software, firmware, people, information, techniques, facilities, services, and other elements to support the above mentioned components [25].

The area of Systems Engineering is a broad interdisciplinary approach that emerges with the aim to successfully develop and implement systems. It is focused on the definition of customer needs and the requirements of the system at an early phase in the development cycle, documenting requirements and proceeding with the design and validation of the system. The system must be designed taking into account several dimensions as, for instance, operations, costs, training or support, considering both the business and the technical needs with the goal to provide a quality product that meets the user needs [25].

The field of System of Systems Engineering has emerged to deal with the problem of System of Systems, which arises when it is required to integrate existing or legacy systems or even new components or systems [14].

The research interest on System of Systems Engineering has increased in recent years. Although there is not a unique and standard definition some convergence has been attained. In [14], the authors surveyed the most prominent definitions for System of Systems Engineering, considering the following proposals:

- System of Systems Engineering is the "transformation of higher order *metasystems* that must function as an integrated complex system to produce desirable results".
- The purpose of System of Systems Engineering is to "satisfy capabilities that can only be met with a mix of multiple, autonomous, and interacting systems".
- System of Systems Engineering is intended to "integrate the capabilities of a mix of existing and new systems into a system-of-systems capability".

3.2 Digital preservation and System of Systems Engineering

As a matter of fact, System of Systems Engineering is all about integration and/or federation [16] of multiple systems that must interoperate in order to achieve a common goal. Bounding the problem to a system of information systems, where the systems to integrate/federate are common information systems or components of information systems, we reduce the integration to the following three dimensions of information systems:

- Information entities (digital objects).
- Processes controlling the information entities (supported by computational services).
- Technological infrastructure required to run those processes.

Considering that a digital preservation system is a sort of information system that must be able to communicate (interoperate/federate) with some unknown system in the future, the ability to interoperate in the above mentioned dimensions are key factors for digital preservation. Thus, a digital preservation system requires the integration/interoperability/federation of information entities, processes and technological infrastructure, as following:

- Information entities: a future system must be able to interpret the representation of the preserved information entities, so that this information can be rendered as the original creator intended to;
- Processes: the alignment and traceability of processes manipulating digital objects during its entire lifecycle is crucial to be able to make assertions about provenance, integrity and authenticity.
- Technological infrastructure: the addition of new components into the preservation environment is required to support the growth of dynamic collections (incrementing the storage space) or to reduce the costs of digital preservation, refreshing components by newly ones with less administration and/or maintenance costs.

From a technical point of view, System of Systems Engineering raises several challenges, especially in the scope of interoperability, information technology and technical integration [20].

The MIT Engineering Systems division has defined a list of non-functional requirements for engineering systems. In [16], the authors consider *Adaptability, Flexibility, Agility, Scalability, Modularity* and *Sustainability* as non-functional requirements for System of Systems Engineering. When thinking about the long-term, all the above mentioned non-functional requirements are critical to be able to run digital preservation solutions in the future, since the solution for digital preservation is not complete if we are not able to preserve the digital preservation system.

Model-driven approaches (MDA) are adequate to face System of Systems Engineering problems, since they intend to face the challenges of business and technology changes. The main idea of MDA is to clearly separate the business and application logics, from platform technology, using platform-independent models (PIM) to document business processes and application functionalities. In order to deploy the system in specific environments, a PIM is transformed into a platform-specific model (PSM), which depends on the specific runtime environment and implementation language.

Resuming, we propose to look at digital preservation problems considering the currently proven methodologies of the Systems Engineering approach, from the problem definition (e.g., needs, constraints, requirements); the study and analysis of cost-effective

[3] http://www.incose.org/

solutions; the process planning, including the identification of technical tasks, efforts and associated risks; the process assessment to evaluate the intermediate developments and measure to progress of the overall solution; and evaluation, including, for instance, unit and global tests, analysis of deployed products or users' satisfaction. As a consequence, digital preservation can be seen as a specific case of System of Systems Engineering, guided by specific requirements as those listed in section 2.

4. THE ENTERPRISE ARCHITECTURE PERSPECTIVE

The ANSI/IEEE 1471-2000 standard [24] defines architecture as "the fundamental organization of a system, embodied in its components, their relationship to each other and the environment, and the principles governing its design and evolution". Similar to this, Enterprise Architecture is a way to align business and technology, managing operations and future development in an organization.

Enterprise Architecture provides a common view of the primary resources of an organization (people, process and technology) and how they integrate and collaborate to provide strategy to the organization. It allows the organization to be flexible in face of changes since the configuration of a system might have to change at any moment, removing, adding or replacing services on the fly in order to align with new business requirements.

Digital preservation can also be seen as a business activity. Taking into account that the ultimate goal of the implementation of a digital preservation system is to be able to offer solutions to address problems in a proper manner, then it should be recognized that such solutions must be always a mix of an organizational structure with the related set of activities and services.

In this section we motivate and promote the use of the Enterprise Architecture for the benefit of the digital preservation area.

Table 1 - The Zachman Framework

View	What (Data)	How (Function)	Where (Network)	Who (People)	When (Time)	Why (Motivation)
Scope	Things important to the business	Processes the business performs	Locations the business operates	Organizations important to the business	Events significant to the business	Business goals/strategies
Business Model	e.g., Semantic Model	e.g., Business Process Model	e.g., Business Logistics System	e.g., Work Flow Model	e.g., Master Schedule	e.g., Business Plan
System Model	e.g., Logical Data Model	e.g., Application Architecture	e.g., Distributed System Architecture	e.g., Human Interface Architecture	e.g., Processing Structure	e.g., Business Rule Model
Technology Model	e.g., Physical Data Model	e.g., System Design	e.g., Technology Architecture	e.g., Presentation Architecture	e.g., Control Structure	e.g., Rule Design
Components	e.g., Data Definition	e.g., Program	e.g., Network Architecture	e.g., Security Architecture	e.g., Timing Definition	e.g., Rule Specification
Instances	e.g., Data	e.g., Function	e.g., Network	e.g., Organization	e.g., Schedule	e.g., Strategy

4.1 Enterprise Architecture Framework

In a generic definition, a framework can be described as "a set of assumptions, concepts, values, and practices that constitutes a way of viewing the current environment" [21].

Frameworks can be used as basic conceptual structures to solve complex issues. In this perspective the definition of a conceptual framework must give preference to scopes, goals, requirements and processes, in the sense that such concepts are already common in Enterprise Architecture frameworks.

The need to rationalize resources, to apply standard governance models and business processes, to comply with strict legal and auditing requirements has prompted central administration services, public services and enterprises in general to adopt Enterprise Architecture Frameworks.

One of the first and most comprehensive Enterprise Architecture frameworks is the Zachman framework[4], defined as "…a formal, highly structured, way of defining an enterprise's systems architecture. (…) to give a holistic view of the enterprise which is being modeled". Table 1 resumes the framework in simple terms, where each cell can be related with a set of models, principles, services, standards, etc., whatever is needed to register and communicate its purpose. The meanings of the lines in that table are:

- Scope (Contextual view; Planner) defined the business purpose and strategy;
- Business Model (Conceptual view; Owner) describes the organization, revealing which parts can be automated;
- System Model (Logical view; Designer) outline of how the system will satisfy the organization's information needs, independently of any specific technology or production constraints;
- Technology Model (Physical view; Builder) tells the system will be implemented, with the specific technology and ways to address production constraints;
- Components (Detailed view; Implementer) details each of the system elements that need clarification before production;
- Instances (Operational view; Worker) give a view of the functioning system in its operational environment.

Concerning the meanings of the columns, "What" refers to the system's content, or data; "How" to the usage and functioning of the system, including processes and flows of control; "Where" to the spatial elements and their relationships; Who to the actors interacting with the system; "When" represents the timings of the processes; "Why" represents the overall motivation, with the option to express rules for constraints where important for the final purpose.

Thus, the Zachman framework is suitable to address digital preservation problems, since it includes models that clearly

[4] Originally conceived by John Zachman at IBM, this framework is now in the public domain, through the Zachman Institute for Framework Advancement. For more details see http://www.zifa.com

separate business processes from applications design and technological solutions. Consequently, it is also suitable to address the challenges imposed by future business or technology changes.

Many other Enterprise Architecture frameworks for specific areas have been developed. Those have been developed by research entities (such as E2A[5]), governmental bodies[6,7,8] (such as FEAF, TEAF, TOGAF, etc.) and private companies (such as the IAF[9], from Cap Gemini).

4.2 Digital preservation and Enterprise Architecture

As a reference framework, OAIS (ISO 14721:2003) is a simple and effective model of concepts and functional entities aimed to provide an abstract quality measure for the design of archival systems or repositories. However, OAIS concepts are too generic and, for some real scenarios, at a far distance of the implementation level. To fill that gap we need to better understand the specific preservation processes and requirements that our preservation environment will have to serve. For example, a preservation environment must comprise not only data manipulation services, but also management and audit of the execution of the defined policies. We need to know how today's preservation processes are related to preservation processes applied in the past, to make assertions about integrity and authenticity.

In fact, to bridge the gap between reference models, processes, systems and people, we should move from the perspective of the engineer (responsible for specific system design) to the perspective of the architect, planning and developing specifications to integrate multiple processes, systems and people. As a consequence, we should motivate digital preservation stakeholders to emphasize on the need for a better integration with the more generic area of enterprise systems.

This way, Enterprise Architecture can mainly support digital preservation in two possible ways. First, Enterprise Architecture can be used as a methodology for engineering design. In this sense, the Enterprise Architecture defines guides of the information required and it complements enterprise methodologies. A methodology defines the steps that the different persons (mainly analysts) must follow in order to generate that pieces of knowledge and fill the Enterprise Architecture [22].

Second, it can be used as a management/audit tool. After the engineering phase, a manager could visualize the relationships between any artifact at different levels and dimensions within the Enterprise Architecture (relationship between processes, resources, people, information, strategy, information systems and so on). Moreover, one can determine if the existing preservation processes are aligned with the requirements of preservation and, in case not, proceed to the alignment.

[5] http://www.enterprise-architecture.info/ (Institute for Enterprise Architecture developments)

[6] http://www.eagov.com

[7] http://www.eaframeworks.com/frameworks.htm;

[8] http://www.whitehouse.gov/omb/egov/a-1-fea.html

[9] http://www.capgemini.com/services/soa/ent_architecture/iaf/

With this purpose, it is necessary to previously have defined relationships between the different architectures and artifacts contained in the Enterprise Architecture [23].

Using the Enterprise Architecture to address digital preservation is equivalent to looking at a digital preservation system as a special case of a generic information system with very specific requirements.

5. THE RISK MANAGEMENT PERSPECTIVE

Risk Management is a continuously developing arena whose ultimate goal is to define prevention and control mechanisms to address the risk attached to specific activities and valuable assets, where risk is defined as the combination of the probability of an event and its consequences [12] It is recognized that Risk Management is concerned with both positive and negative consequences of risks.

ISO/FDIS 31000 [13] is a risk management standard currently under development (the first version is expected during 2009). It intends to define the principles and implementation of Risk Management to control the behavior of an organization with regard to risk, and is based on the principle that Risk Management is a process operating at different levels, as shown in Figure 1. The Risk Management process encloses the limitation of the context, risk assessment (identification, analysis and evaluation of risks) and risk treatment. This process requires a continuous monitor and review activity to audit the behavior of the whole environment allowing, for instance, the identification and treatment of an unexpected vulnerability.

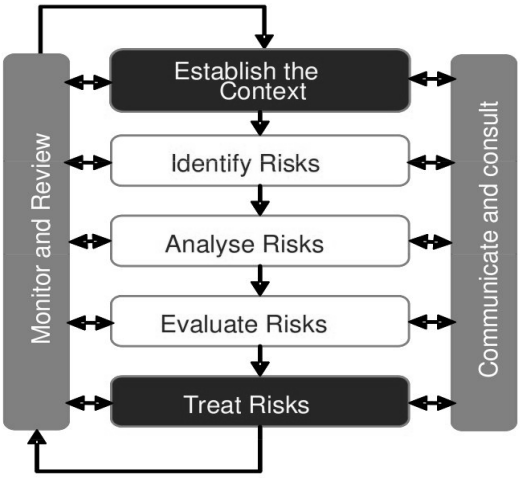

Fig. 1 - Risk Management Process

First, defining the context is crucial to identify strategic objectives and define criterions to determine which consequences are acceptable to this specific context. Second, today's organizations are continuously exposed to several threats and vulnerabilities that may affect their normal behavior. The identification, analysis and evaluation of these threats and vulnerabilities are the only way to decide on the appropriate techniques to handle them. The identification of threats, vulnerabilities and risks is based on events that may affect the achievement of goals identified in the

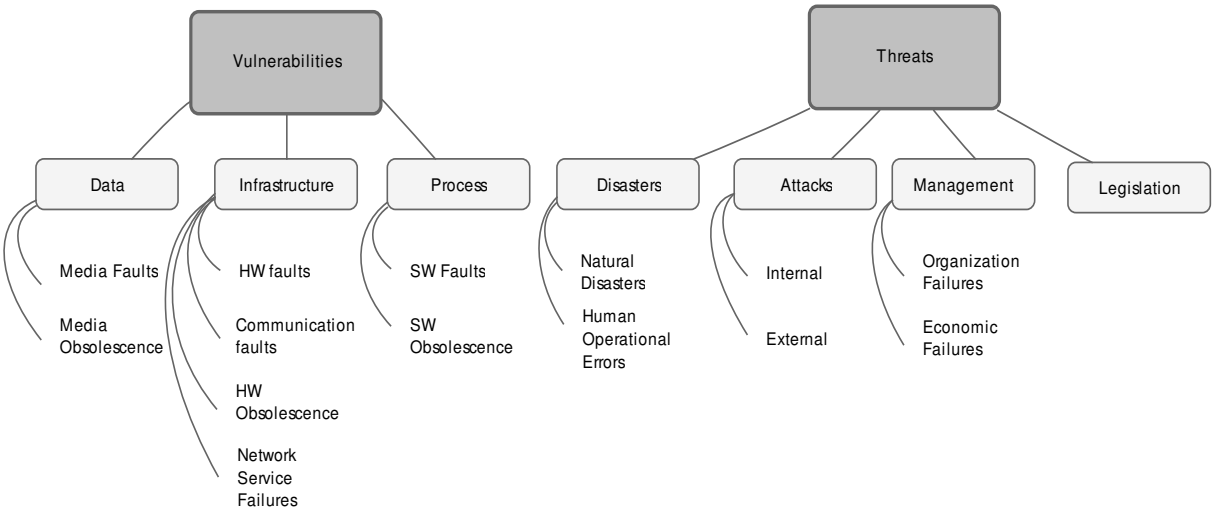

Fig. 2 - Taxonomy of vulnerabilities and threats to digital preservation

first phase. After that, the risk analysis and evaluation estimates the likelihood and impact of risks to the strategic goals, in order to be able to decide on the appropriate techniques to handle these risks (Treat Risks).

Currently, the digital preservation arena just uses Risk Management concepts to assess repositories. The Trustworthy Repositories Audit and Certification (TRAC) Criteria and Checklist[10] is meant to identify potential risks to digital content held in repositories. It takes OAIS as its intellectual foundation, and as the benchmark for measuring success in terms of trustworthiness. It establishes appropriate methodologies for determining the soundness and sustainability of digital repositories.

The Digital Repository Audit Method Based on Risk Assessment (DRAMBORA) [11] process focuses on risks, and their classification and evaluation according to individual repositories' activities, assets and contextual constraints.

In this paper, we intend to go beyond and propose a perspective where Risk Management can be used to assess existing solutions, but also to conceive digital preservation environments, which encloses three steps: (*i*) establish digital preservation requirements (context and strategic objectives); (*ii*) identify digital preservation vulnerabilities and threats, and (*iii*) address digital preservation threats and vulnerabilities (treat risks).

Following, we survey the main requirements to digital preservation, present a taxonomy to represent digital preservation threats and vulnerabilities, and survey a set of techniques that can be applied to address these threats and vulnerabilities.

5.1 Digital preservation threats and vulnerabilities

In this section we present a revision of the taxonomy of threats to digital preservation presented in [3], which was based on papers that point out different threats [2, 7].

[10] The TRAC checklist is available at http://www.crl.edu/PDF/trac.pdf

Figure 2 presents our revised taxonomy, which is based on the Risk Management terminology, considering vulnerabilities[11] and threats[12] to digital preservation. Thus, vulnerabilities are weaknesses (potential points of failures) in the environment and threats are events that affect the normal behavior. For instance, a natural disaster threat may exploit several vulnerabilities in the preservation environment.

Like common information system's architectures, we consider a preservation environment as the aggregation of different components, namely: (*i*) the information entities, including preserved objects and metadata; (*ii*) processes controlling the information entities (can be supported by computational services)

and (*iii*) the technological infrastructure that supports the preservation environment. Based on that assumption, each of these components may present several vulnerabilities. Thus, we propose a classification of vulnerabilities in: (*i*) **data vulnerabilities**, affecting the information entities; (*ii*) **infrastructure vulnerabilities**, enclosing the technical problems in the infrastructure's components; and (*iii*) **process vulnerabilities**, affecting the execution of processes (manual or supported by computational services) that control information entities.

Data vulnerabilities include **media faults** that occur when a storage media fails partially or totally, losing data through disk crashes or "bit rot". **Media obsolescence** is a different kind of failure that occurs when the representation format becomes obsolete and unable to be rendered, even if the "bit stream" survives over time.

Infrastructure components can suffer **hardware faults** by transient recoverable failures, like power loss, or irrecoverable

[11] The existence of a weakness, design, or implementation error that can lead to an unexpected, undesirable event compromising the security of the computer system, network, application, or protocol involved [13].

[12] Any circumstance or event with the potential to adversely impact an asset through unauthorized access, destruction, disclosure, modification of data, and/or denial of service [13].

failures, such as a power supply unit burning out. Similarly to media formats, hardware components can become obsolete and unable to communicate with other components (**hardware obsolescence**).

Communication faults occur in packet transmission, including detected errors (e.g., IP packet error) and undetected checksum errors. Other **network services failures**, such as DNS problems, can compromise the system availability.

Processes supported by software services can be affected by **software faults**, usually known as bugs that can cause abrupt failures in the system. For instance, a firmware migration error can cause an unexpected data loss. Again, **software obsolescence** can limit the execution of processes due to the impossibility to interact with other components (infrastructure or data).

We propose the classification of threats to digital preservation into **disasters**, **attacks**, **management** and **legislation**. Management failures are the consequences of wrong decisions that produce several threats to the preservation environment. Disasters and attacks correspond, respectively, to non-deliberate and deliberate actions affecting the system or its components. Finally, legislation threats occur when digital preservation processes or preserved data violate new or updated legislation.

An organization responsible for a preservation system may become unable to continue operating at the desired level due to sudden financial limitations (**economic failure**), political changes or any other unpredictable reason (**organization failure**). Moreover, failures can also occur due to incompetent management.

Natural disasters, such as earthquakes or fires can cause failures in many components simultaneously. For example, an earthquake may cause a data center to be destroyed, producing, for instance, hardware faults and media faults. Accidentally **human operational errors** might introduce irrecoverable errors. For instance, people often delete data by mistake. Additionally, humans can cause failures in other components such as hardware (accidentally disconnecting a power cable) or software (uninstalling a needed library).

Attacks might encompass deliberate data destruction, denial of service, theft, and modification of data or component destruction, motivated by criminal, political or war reasons, including fraud, revenge or malicious amusement. Systems connected to public networks are especially exposed to **external attacks**, such as those caused by viruses or worms. Similarly, **internal attacks** might be performed by internal actors (e.g., employees) with privileged access to the organization and to the physical location of the components.

Some vulnerabilities and threats cannot be detected immediately, remaining unnoticed for a long time. For instance, a damaged hard disk sector can remain undetected until a data integrity validation or hard disk check is performed. Moreover, we cannot assume threat and/or vulnerability independence.

5.2 Addressing digital preservation threats and vulnerabilities

Following the Risk Management process, this section proposes techniques to address the threats and vulnerabilities identified in Section 4.2 (see Table 1).

First, we would like to remark that auditing is used to quickly detect vulnerabilities and threats to the preservation environment, allowing the rapid execution of corrective techniques. For example, faults that cause data loss may only be detected when the data is accessed. This can be done by auditing the system periodically. Thus, the auditing technique is equivalent to the monitor and review activity of the Risk Management process.

The goal of migration is to keep digital objects in recent media formats. Lossless migrations of data maintain exactly the same contents as the original version, while loss migrations might imply the loose of some information in the process. As a consequence, the aim of migration techniques is to reduce the risk of media obsolescence.

Emulation techniques are also used to reduce the risk of media obsolescence, where original hardware and/or software conditions of fruition for which the information objects were initially conceived (production environment) are simulated in more recent systems.

The extra information required to be able to deal with obsolescence of media formats is usually the technical metadata. In the simplest scenario, the requirement of storing the object along with its technical metadata (encapsulation) may suffice. In this scenario the preservation system is not required to do anything special with that data, except give it back when required. In a more complex scenario, the system might be required to support specific input and output formats, to make it possible to ingest objects in one or more formats and retrieve them in other different ones (whether the transformation inside the system is done in advance or in real time is a question of implementation to be dictated by other, for now irrelevant, requirements).

Each of these scenarios will be adequate to different preservation strategies to deal with format obsolescence. For example, the first is the adequate for emulation, while the second might be required for migration. Remark that usually, emulation is required for dynamic objects (like games or other software), while migration is widely used in static objects (e.g., images, text).

Redundancy techniques make use of a basic attribute of digital information: it can be copied without any loss of information. This means that several copies of the data can be stored across many components. As a consequence, if a data loss does not affect all the valid replicas of a digital object, this object can be recovered from the undamaged replica. Thus, the risk of the media fault vulnerability, as well as the threats imposed by attacks, human operational errors and natural disasters, can be reduced adopting redundancy techniques.

The risk of media faults can be reduced by refreshing the media supports used (e.g., replace hard disks). Moreover, redundancy, metadata and auditing support the recover from media faults, in the sense that auditing allows the quick identification of undetected media faults. Metadata is also required, for instance to verify the integrity of corrupted objects and finally, redundancy is crucial to obtain an undamaged copy.

Table 2 - Addressing digital preservation threats and vulnerabilities. r: reduces de risk of the threat/vulnerability; R:required for recovery; -: does not fit

Threats and vulnerabilities			Techniques							
			Redundancy	Migration	Emulation	Refreshing	Diversity	Inertia	Metadata	Auditing
Vulnerabilities	Process	Software faults	-	-	-	r	r	-	-	R
		Software obsolescence	-	-	-	r	r	-	-	R
	Data	Media faults	R	-	-	r	-	-	R	R
		Media obsolescence	-	r	r	-	-	-	R	R
	Infrastructure	Hardware faults	-	-	-	r	r	-	-	R
		Hardware obsolescence	-	-	-	r	r	-	-	R
		Communication faults	-	-	-	r	r	-	-	R
		Network service failures	-	-	-	r	r	-	-	R
Threats	Disasters	Natural disasters	R	-	-	-	r	-	-	-
		Human operational errors	R	-	-	-	r	r	R	R
	Attacks	Internal attack	R	-	-	-	r	r	R	R
		External attacks	R	-	-	-	r	r	R	R
	Management	Economic failures	-	-	-	-	r	-	-	R
		Organization failures	-	-	-	-	r	-	-	R
	Legislation	Legislation changes	-	-	-	-	r	-	r	-

The risk of obsolescence and faults in components of the preservation infrastructure can be reduced diversifying the infrastructure components (reducing the probability of correlated failures) and refreshing the components by recent and robust ones. Similarly, the risk of vulnerabilities that affect processes supported by computational services (software) can be reduced diversifying and refreshing these computational services, since common software components present the same vulnerabilities to virus, bugs, etc.

With respect to disaster threats, diversity, especially of the physical location, is the technique that can be adopted to reduce the risk of natural disasters like earthquakes or fires, since natural disasters may affect several components in a limited area. Human operational errors can produce the same consequences of deliberate internal and external attacks. Thus, diversifying the administration and physical location limits the potential malicious actions that can be performed by users. Moreover, if inertia is also applied, the speed of destruction is also lowered. Furthermore, physical location diversity also reduces the risk of terrorist attacks to the infrastructure, and auditing, metadata and redundancy are crucial to detect the effects of operational errors and attacks, also supporting the recover to a normal state.

Management failures occur when an organization responsible to run a preservation environment becomes unable to continue operating, due to financial limitations, political reasons or other organizational problem. From the surveyed techniques, diversifying funding or even organizations involved in the digital preservation environment are techniques to reduce the risk of this threat.

Sudden legislation modifications may demand changes in object rights, modification of preservation processes, etc. The risk imposed by legislation changes can be reduced diversifying responsible entities (managed by different laws) and the adequate cataloguing of rights metadata.

6. ANALYSING THE DIFFERENT PERSPECTIVES

In previous Sections we have shown how the digital preservation problem can be addressed from three different perspectives, namely: System of Systems Engineering, Enterprise Architecture and Risk Management. Although these are distinct research areas that follow different lines of investigation, they are not mutually exclusive.

As a matter of fact, the proposed perspectives are all guided by specific digital preservation requirements, which are the basis to start System of Systems Engineering, Enterprise Architecture and Risk Management processes. In section 2, we proposed a set of

general digital preservation requirements but, for specific scenarios, the proposed requirements must be refined and/or extended in order to accurately represent the specificities for each scenario.

However, the relationships between the proposed perspectives go further than the common ground on the definition of requirements. Actually, Enterprise Architecture can be seen as a methodology that may help in the engineering of solutions for System of Systems, dealing with the complexity of such kind of systems. In [20], the authors propose a combination of Enterprise Architecture initiatives with strategic planning, engineering the business and systems as a whole for an Enterprise Systems Engineering, to be able to deal with the complexity of System of Systems Engineering and respect strong non-functional requirements. As a consequence, there is a strong connection between System of Systems Engineering and Enterprise Architecture Frameworks.

Furthermore, System of Systems is usually supported by high-speed Internet networks, where trust can not be assured, since systems may be affected by multiple viruses and worms, or even attacks compromising data. As a consequence, Systems of Systems are also characterized by a risk and uncertainty that may affect their normal behavior. Thus, the process of Risk Management and assessment should constitute an integral part of the engineering of this kind of systems [19]. Once more, this is also a common ground between System of Systems Engineering and digital preservation, as the ultimate goal of digital preservation is the reduction, as much as possible, of the risk associated with being unable to obtain digital information as its creator intended to.

7. CONCLUSIONS

This paper presents an approach where digital preservation can be seen from three different perspectives: as a specific case of the more generic area of System of Systems Engineering; as a specific case of an Enterprise Architecture; and as a problem of Risk Management.

We present the main characteristics and requirements of System of Systems and motivate the use of Enterprise Architecture Frameworks to address the embodied complexity of digital preservation. Moreover, Risk Management activities comprise a common ground between System of Systems Engineering, Enterprise Architecture and digital preservation.

Since the ultimate goal of digital preservation consists in the reduction of risks associated with a data loss, we propose a Risk Management based approach to digital preservation, consisting in three different phases: establishing digital preservation requirements, identifying digital preservation threats and vulnerabilities, and treating the risks associated with the identified threats and vulnerabilities. We surveyed the main requirements to digital preservation and classified the threats and vulnerabilities that might endanger preservation using a taxonomy of threats/vulnerabilities. Finally, we propose common techniques used in digital preservation and show how these techniques can be used to address the identified threats and vulnerabilities.

In a digital preservation system, components may fail in a correlated manner, since some threats may cause the failure of multiple components with similar configurations. Moreover, each preservation scenario has its own specificities, making it impossible to determine which technique is better suited to all the scenarios. Moreover, even if one can specify that a specific technique is the most adequate, there are several potential applications of this technique (e.g., which format to migrate, where to put replicas and how many replicas in redundancy strategies).

To effectively assess and measure adequate risk treatment for digital preservation scenarios, we proposed a simulator [8, 9], that can be used to evaluate the risk of threats (natural disasters) and infrastructure failures, on a preservation environment using redundancy and diversity techniques. We plan to extend this under-development simulator to follow the Risk Management based approach to digital preservation, developing an effective risk analysis tool for preservation using the proposed taxonomy of threats/vulnerabilities and digital preservation techniques.

8. ACKNOWLEDGMENTS

This work is partially supported by the projects GRITO (FCT, GRID/GRI/81872/2006) and SHAMAN (European Commission, ICT-216736), and by the grant from FCT (SFRH/BD/23405/2005) and LNEC to José Barateiro.

9. REFERENCES

[1] Consultative Committee on Space Data Systems, ISO 14721:2003 - Reference model for an open archival information system, 2003.

[2] M. Baker, M. Shah, D. S. H. Rosenthal, M. Roussopoulos, P. Maniatis, T. J. Giuli, and P. P. Bungale. A fresh look at the reliability of long-term digital storage. In EuroSys, pages 221-234, 2006.

[3] J. Barateiro, G. Antunes, M. Cabral, J. Borbinha, and R. Rodrigues. Using a grid for digital preservation. In International Conference on Asia-Pacific Digital Libraries, pages 225-235, 2008.

[4] Metadata encoding and transmission standard. Primer and reference manual. Digital library federation, 2007.

[5] IEEE. IEEE Standard Computer Dictionary: A Compilation of IEEE Standard Computer Glossaries, 1990.

[6] P. Maniatis, M. Roussopoulos, T. J. Giuli, D. S. H. Rosenthal, and M. Baker. The lockss peer-to-peer digital preservation system. ACM Trans. Comput. Syst., 23(1):2-50, 2005.

[7] D. S. H. Rosenthal, T. Robertson, T. Lipkis, V. Reich, and S. Morabito. Requirements for digital preservation systems: A bottom-up approach. CoRR, abs/cs/0509018, 2005.

[8] G. Antunes, J. Barateiro, M. Cabral, J. Borbinha, and R. Rodrigues. Preserving Digital Data in Heterogeneous Environments. In JCDL, 2009 (to appear).

[9] J. Barateiro, G. Antunes, F. Freitas and J. Borbinha. Challenges on preserving scientific data with data grids. 1st International Workshop on Data Grids for E-Science. 2009 (to appear).

[10] AIRMIC, ALARM, IRM. A Risk Management Standard. 2002.

[11] A. McHugh, R. Ruusalepp, S. Ross and H. Hofman. Digital Repository Audit Method Based on Risk Assessment. Edinburgh: DCC and DPE. 2007.

[12] ISO/IEC Guide 73:2002 – Risk Management vocabulary guidelines for use in standards, 2002.

[13] ISO/FDIS 31000 – Risk Management principles and guidelines, 2009.

[14] A. Sousa-Poza, S. Kovacic and C. Keating. System of systems engineering: an emerging multidiscipline. International Journal of System of Systems Engineering. Volume 1 - Issue 1/2, 2008.

[15] R. Moore. Towards a Theory of Digital Preservation. International Journal of Digital Curation. Volume 3 – Issue 1, 2008.

[16] R. Valerdi, E. Axelband, T. Baehren, B. Boehm, D. Dorenbos, S. Jackson, A. Madni, G. Nadler, P. Robitaille and S. Settles. A research agenda for systems of systems architecting. International Journal of System of Systems Engineering. Volume 1 - Issue 1/2, 2008.

[17] J. Borbinha: It Is the Time for the Digital Library to Meet the Enterprise Architecture. In ICADL, pages 176-185, 2007.

[18] F. Baiardi and C. Telmon. Risk management of an information infrastructure: a framework based upon security dependencies. International Journal of System of Systems Engineering. Volume 1 - Issue 1/2, 2008.

[19] Y. Haimes. Models for risk management of systems of systems. International Journal of System of Systems Engineering. Volume 1 - Issue 1/2, 2008.

[20] P. Chen and J. Clothier. Advancing systems engineering for systems-of-systems challenges. Journal of Systems Engineering. Volume 6 Issue 3, Pages 170 – 183, 2003.

[21] OASIS - Organization for the Advancement of Structured Information Standards. Reference Model for Service Oriented Architecture. Committee Specification 1. 2 August 2006.

[22] V. Anaya and A. Ortiz. How enterprise architectures can support integration. In Proceedings of the First international Workshop on Interoperability of Heterogeneous Information Systems. Germany, 2005.

[23] C. Braun and R. Winter. Integration of IT service management into enterprise architecture. In Proceedings of the ACM Symposium on Applied Computing. Korea, 2007.

[24] IEEE: IEEE Standard 1471-2000 IEEE Recommended Practice for Architectural Description of Software-Intensive Systems –Description. 9 October 2000.

[25] INCOSE. Systems Engineering Handbook: A How To Guide For All Systems Engineers, Release 1.0, 1998