Assessing Digital Preservation Capabilities Using a Checklist Assessment Method

Gonçalo Antunes, Diogo Proença, José Barateiro, Ricardo Vieira, José Borbinha
INESC-ID Information Systems Group, Lisbon, Portugal
{goncalo.antunes, diogo.proenca, jose.barateiro, rjcv, jlb}@ist.utl.pt

Christoph Becker
Vienna University of Technology
Vienna, Austria
becker@ifs.tuwien.ac.at

ABSTRACT

Digital preservation is increasingly recognized as a need by organizations from diverse areas that have to manage information over time and make use of information systems for supporting the business. Methods for assessment of digital preservation compliance inside an organization have been introduced, such as the Trustworthy Repositories Audit & Certification: Criteria and Checklist. However, these methods are oriented towards repository-based scenarios and are not geared at assessing the real digital preservation capabilities of organizations whose information management processes are not compatible with the usage of a repository-based solution. In this paper we propose a checklist assessment method for digital preservation derived from a capability-based reference architecture for digital preservation. Based on the detailed description of digital preservation capabilities provided in the reference architecture, it becomes possible to assess concrete scenarios for the existence of capabilities using a checklist. We discuss the application of the method in two institutional scenarios dealing with the preservation of e-Science data, where clear gaps where identified concerning the logical preservation of data. The checklist assessment method proved to be a valuable tool for raising awareness of the digital preservation issues in those organizations.

Categories and Subject Descriptors

H.1 [**Information Systems**]: Models and Principles; J.1 Administrative Data Processing Government; K.6.4 Management of computing and Information Systems

General Terms

Management, Documentation, Measurement, Verification

Keywords

Repository Audit and Certification, Trust, Digital Preservation, Reference Architecture, Checklist Assessment

1. INTRODUCTION

Digital preservation (DP) has traditionally focused on repository-based scenarios, mainly driven by memory institutions. All the main reference models of the field such as the well-known case of the OAIS [1] have been developed with this concern in mind. These models define preservation processes, policies, requirements, and building blocks that can be used by institutions that host or want to host a repository system to effectively manage its implementation and/or its operation.

These references are widely considered valid for these kinds of scenarios. However, DP is starting to be acknowledged as a need by organizations from different walks of life in scenarios where common information systems are used for processing and managing data, and where no separate system for preservation is desirable, so that a repository approach is not applicable. These scenarios present emergent DP requirements, where DP is seen as a *desirable property of information systems*, and not as the main source of functional requirements. In that sense, those organizations execute information management processes that cannot be aligned with the functional aspects and information structures defined in the main reference frameworks of the DP domain. Despite the apparent shift, the main objective of preservation is maintained intact, which involves assuring that information that is understood today can be transmitted into an unknown system in the future and still be correctly understood then. Thus, besides the traditional repository scenario, an alternative scenario should be considered, where DP is seen as a capability that can be added to systems. Figure 1 depicts the two possibilities.

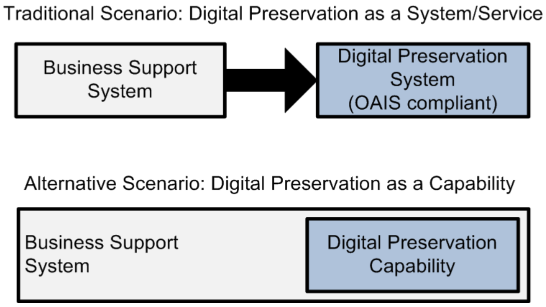

Traditional Scenario: Digital Preservation as a System/Service
Business Support System
Digital Preservation System (OAIS compliant)
Alternative Scenario: Digital Preservation as a Capability
Business Support System
Digital Preservation Capability

Figure 1. Digital Preservation Scenarios

With this in mind, a capability-based reference architecture was produced in the context of the SHAMAN[1] project and described in [3]. Reference architectures have the aim of capturing domain-specific knowledge and integrate that knowledge in a way that it can be later reused for developing new system architectures for the domain in question [4]. In that sense, the capability-based reference architecture captures knowledge from the DP domain, consolidates that knowledge taking into account reference models and best-practices of related or highly relevant domains, so that it can be reused for assessing and guiding the integration of DP capabilities in information systems. The purpose is to deliver value in organizations where DP is not a business requirement, but it required to enable the delivery of value in the primary business.

Several assessment methods are currently available in the DP domain for evaluating the effectiveness of DP in repository-based scenarios. Works like the Trustworthy Repositories Audit & Certification: Criteria and Checklist (TRAC) [5], DRAMBORA [6], or the freshly published ISO 16363:2012 [7], allow the assessment of a repository system and the surrounding

[1] http://shaman-ip.eu/

organizational environment using several different perspectives. However, their application in non-traditional DP scenarios is difficult, mainly due to the assumption that a repository system is present and that once data enters such system, it will only be accessed again in the long-term. This work proposes a checklist assessment method based on the capability-based reference architecture. The checklist itself is based on the assessment methods already existing in the DP domain, but significantly reworked and aligned with the capability approach, so that it can be applied to any scenario. It contains sets of criteria organized per capability. The implementation was made through a spreadsheet that can be configured by the user in order to concede different weights to different criteria according to the concerns of the stakeholder filling the checklist. In that way, the current DP capabilities can be identified and their levels assessed, and a gap analysis between the current and the desired situation can be performed, which can support decision making on improvements to the current situation.

This paper is structured as follows. Section 2 describes the related work in terms of assessment checklists in the DP domain and in other relevant domains. Section 3 describes a capability-based reference architecture for DP. Section 4 describes a method for assessing the current DP capabilities of an organization and a companion checklist for performing the assessment. In Section 5, the application of the checklist assessment method to two institutions dealing with the issue of preserving e-Science data is described. Finally, Section 6 discusses lessons learned, and draws conclusions.

2. RELATED WORK

The usage of assessment checklists is widely spread, being used in various areas. In the DP domain, the Trustworthy Repositories Audit & Certification: Criteria and Checklist (TRAC) [5] is one example. Its purpose is to be an audit and certification process for the assessment of the trustworthiness of digital repositories, and its scope of application it's the entire range of digital repositories. It is based on the OAIS model [1]. The final version of TRAC was published in 2007, based upon the experience and findings of various test audits by the Center for Research Libraries from 2005 to 2006. It contains 84 criteria which are divided into three main sections: Organizational infrastructure; Digital object Management; and Technologies, technical infrastructure, and security. Within each of this sections are various subsections and under the subsections are the criteria. A successor version of TRAC, a standard for Trusted Digital Repositories (TDR), was published by ISO in February 2012, the ISO16363:2012 standard [6].

In the DP domain there are other assessment tools, for example, the Northeast Document Conservation Center self-assessment tool [8]. This tool aims at helping the museums, libraries, archives, and other cultural organizations to begin thinking about long-term sustainability of their digital collections and complements the DP readiness assessment developed by the same center. It covers the following topics: (1) Mission and Goals; (2) Policies and procedures; (3) Staffing; (4) Finances; (5) Digital content; (6) Technology; (7) Access and metadata; (8) Digital preservation and (9) Rights Management.

A different approach for the assessment of repositories has been taken by DRAMBORA [6], a digital repository audit method based on risk assessment. DRAMBORA characterizes digital curation as a risk-management activity, because it recognizes the job of a digital curator as the rationalization of the uncertainties and threats that inhibit efforts to maintain digital object authenticity and understandability, transforming these into manageable risks. There are six stages within the process. The first stages require that auditors develop an organizational profile, describing and documenting the repository's mandate, objectives, activities and assets. Then, risks are derived from each of these, and assessed in terms of their likelihood and potential impact. In the end, auditors are encouraged to conceive of appropriate risk management responses to the identified risk.

There are other domains which make use of checklist in order to assess a certain capability. For example in the IT domain, ISACA provides an IT Governance Self-Assessment checklist [9] in order for the management to determine, for each of the COBIT [10] processes: (1) How important they are; (2) Whether it is well performed; (3) Who performs and who is accountable; (4) Whether the process and its control is formalized and (5) Whether it is audited.

Other domains of usage include teaching [11], for example, to record observed performance of students while working in groups, to keep track of progress over time or even help students fulfill task requirements.

In conclusion, assessments using checklists are well spread in numerous domains, including the DP domain, applied for example in healthcare institutions [13], pharmaceutical industry, and manufacturing, and many other areas as described in [14] and [15]. Checklists are proven to be a successful tool to verify the state of certain aspect, in an organization, class room or even yourself.

However, DP assessment checklists assume the presence of a repository system and that once data enters the repository it will be seldom accessed. Despite that being desirable for a wide range of scenarios (e.g., cultural heritage), the existence of such solution might not be adequate for determined organizations, where data management processes are well-defined and established and specialized information systems are in place. In other words, this work aims to bridge that existing gap through a proposal of a capability assessment checklist that can be applied to any organization. Additionally, while existing DP checklists allow the assessment of important aspects of DP in organizations, they do not provide a means for evaluating the current capability level. This alone allows performing a gap analysis that can help organizations to make investments in order to fill the gaps.

3. A CAPABILITY-BASED REFERENCE ARCHITECTURE FOR DIGITAL PRESERVATION

A reference architecture can be defined as a way of documenting good architectural practices in order to address a commonly occurring problem through the consolidation of a specific body of knowledge with the purpose of making it available for future reuse [4]. Thus, a reference architecture for DP provides a way of capturing the knowledge of the DP domain, so that it can be instantiated in concrete architectures for real system implementations.

In recent years several DP reference models and frameworks have been developed providing terminology, building blocks, and other types of knowledge derived from an in-depth analysis of the domain. Although being widely accepted, these reference models are not aligned among themselves and often overlap with established references and models from other fields, such as IT Governance or Risk Management. Moreover, those frameworks

Table 1. Reference Architecture Capabilities

Capability		Description
GRC Capabilities	GC1. Governance	The ability to manage and develop the services, processes and technology solutions that realize and support DP capabilities. This includes engaging with the designated communities in order to ensure that their needs are fulfilled is also an important aspect. The ability to negotiate formal succession plans to ensure that contents do not get lost is another important aspect.
	GC2. Risk	The ability to manage and control strategic and operational risks to DP and opportunities to ensure that DP-critical operations are assured, including the sustainability of those operations and disaster recovery.
	GC3. Compliance	The ability to verify the compliance of DP operations and report deviations, if existing. Certification is also an important aspect of this capability and it consists in the ability to obtain and maintain DP certification status.
Business Capabilities	BC1. Acquire Content	The ability to offer services for transferring content from producers into the organization's systems. This includes services for reaching agreement with producers about the terms and conditions of transfer.
	BC2. Secure Bitstreams	The ability to preserve bitstreams for a specified amount of time (Bitstream preservation).
	BC3. Preserve Content	The ability to maintain content authentic and understandable to the defined user community over time and assure its provenance. (Logical preservation).
	BC4. Disseminate Content	The ability to offer services for delivering content contained in the organization's systems to the user community or another external system. This includes services for reaching agreement about the terms and conditions of transfer.
Support Capabilities	SC1. Manage Data	The ability to manage and deliver data management services, i.e. to collect, verify, organize, store and retrieve data (including metadata) needed to support the preservation business according to relevant standards.
	SC2. Manage Infrastructure	The ability to ensure continuous availability and operation of the physical, hardware, and software assets necessary to support the preservation.
	SC3. Manage HR	The ability to continuously maintain staff which is sufficient, qualified and committed to performing the tasks required by the organization.
	SC4. Manage Finances	The ability to plan, control and steer financial plans and operations of the organization's systems to ensure business continuity and sustainability.

are not always aligned with best practices, resulting in specifications that are not easy to use or that are not reusable at all. A reference architecture following best practices in the field of enterprise architecture would fit the purpose of making that knowledge available in a way that would facilitate its reuse.

In order to create a DP reference architecture that infused domain knowledge, the TOGAF Architecture Development Method (ADM) [12] was used for developing an architecture vision accommodating DP capabilities. For that, the main reference models of the domain were surveyed and integrated, providing a means of effectively addressing the issues of DP, while providing a bridge for the development of concrete DP-capable architectures. Following the ADM, the stakeholders of the domain and their concerns were identified along with the drivers and goals. This resulted in a set of general DP capabilities derived from the context, in a process that is documented in [13].

A capability is not a business function, but an ability realized by a combination of elements such as actors, business functions and business processes, and technology, and it must be related with at least one goal. This reference architecture for DP defines a set of capabilities that can be divided in three groups, which are also described in an increased level of detail in Table 1:

Governance, Risk and Compliance (GRC) Capabilities - Governance capabilities are required to manage the scope, context and compliance of the information systems in order to ensure fulfillment of the mandate, continued trust of the external stakeholders and sustainable operation of the systems.

Business Capabilities - Business capabilities are required to execute a specified course of action, to achieve specific strategic goals and objectives.

Support Capabilities - Support capabilities are required for ensuring the continuous availability and operation of the infrastructure necessary to support the organization, including physical assets, hardware, and software.

Table 2. Goals and Capabilities

ID	Goals	Capabilities
G1	Acquire Content…	BC1;
G2	Deliver…	BC4;
G3	…preserve provenance…	BC2, BC3, SC1;
G4	…preserve objects…	BC2, BC3;
G5	React to changes…	GC1, GC2, BC3, SC2;
G6	…sustainability…	GC1, GC2, GC3, SC2, SC3, SC4;
G7	Build trust…	GC1, GC2, GC3;
G8	Maximize efficiency…	GC1, GC2, SC1, SC2, SC3, SC4;

The reference architecture also defines general goals for DP. Eight goals were derived from the various references collected: (*i*) G1. **Acquire content** from producers in accordance to the mandate, following agreed rules; (*ii*) G2. **Deliver** authentic, complete, usable and understandable objects to designated user community; (*iii*) G3. Faithfully **preserve provenance** of all objects and deliver accurate provenance information to the users upon request; (*iv*) G4. Authentically **preserve objects** and their dependencies for the specified time horizon, keeping their integrity and protecting them from threats; (*v*) G5. **React to changes** in the environment timely

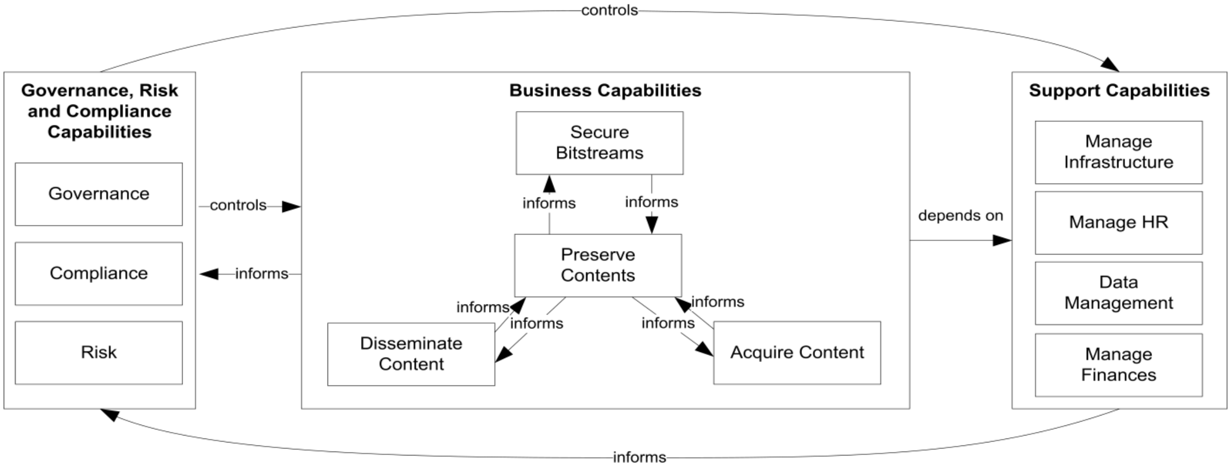


Figure 2. Capability Relationship Diagram

in order to keep objects accessible and understandable; (*vi*) G6. Ensure organization's **sustainability**: mandate, technical, financial, operational, communities; (*vii*) G7. Build **trust** in the depositors, the designated community and other stakeholders; and (*viii*) G8. **Maximize efficiency** in all operations. Table 2 provides a consolidated overview of all goals and the related capabilities considered here.

The categorization of these capabilities of course is partly context-dependent: in a concrete business environment, DP will generally be seen as a part of IT Governance and thus of Governance. Since it is our core focus of modeling, DP is highlighted and as such presented separately from more general aspects of IT Governance. Upon realization in a business environment, DP (and Data Management) will likely be realized as part of IT Governance, and will thus be submerged into it.

Capabilities do not exist in isolation and will have mutual dependencies. A model of their relationships and the specification of the relations existing between capabilities enable operationalization of these capabilities and an assessment of the influences exerted within capabilities in concrete scenarios. Table 3 describes the different types of relations that may exist between capabilities.

Table 3. Relations between Capabilities

Name	Description
influence	A directed relation between two capabilities
controls	An influence that determines the range of possible behavior
informs	An influence that does not exert full control, but constitutes a flow of information (that may drive or constrain the range of possible behavior in a non-exclusive way)
depends on	A relation that constitutes a dependency: The using capability is unable to act without relying on capabilities offered by the used capability. This implies a reverse "informs" relationship.

Figure 2 depicts the relations existing between capabilities. At the top level, GRC capabilities exert control over Business capabilities and Support capabilities, since they set out the scope and goals for business, and represent the regulators that constrain business. Business capabilities inform the GRC capabilities, in particular: (*i*) Governance, to provide information about the operations and the status of the organization's systems, to assess opportunities and potential and be aware of operational constraints, and to determine the adequacy of means to achieve ends; (*ii*) Compliance, to enable auditing of compliance to regulations; and (*iii*) Risk, to provide information about the adequacy of preservation actions to face threats endangering the preserved contents. Support capabilities inform GRC capabilities since GRC needs information to successfully govern support capabilities. Business capabilities also have a dependency relationship with Support capabilities, since the former relies on the later. Although other relation types may exist between top-level capabilities, only the most prevalent are depicted on the diagram.

As for the relationships between Business capabilities, the Acquire Content capability informs the Preserve Contents capability, since the former constitutes a system boundary and thus the point where the organization gets control of content and the properties of acquired content are of interest for preservation. The same relationship is also true in the opposite direction since the limits of operational preservation may constrain the range of contents that can be accepted responsibly. The Disseminate Content informs the Preserve Contents since Dissemination requirements may drive and/or constrain preservation. Again, the same relationship is also true in the opposite direction since the limits of operational preservation may constrain the options for dissemination. The Secure Bitstreams capability informs the Preserve Contents capability since the way the bitstreams are physically secured may drive or constrain preservation (i.e. probabilities for bit corruption). The same relationship is also true in the opposite direction since effects of preservation may drive or constrain the way the bitstreams are physically secured (i.e. file sizes). For a detailed discussion on the existing relationships, please refer to [12].

4. ASSESSING DIGITAL PRESERVATION CAPABILITIES

With the detailed description of capabilities provided, it becomes possible to assess concrete scenarios for the existence of capabilities, since the breakdown provided allows easier assessment of the organization, making the bridge into the business architecture. An organization should map the stakeholders and their concerns in the ones provided in the reference architecture [13]. Based on that, the preservation drivers and goals are determined, also based on the ones provided by this reference architecture, but also checking at all times for possible constraints that might affect the architecture work. That process shall provide a clear vision of the current DP capabilities and the ones effectively missing. The next following section provides a method to be used together with a checklist. After the assessment, the development and deployment of capabilities in concrete scenarios becomes possible through the development of architecture viewpoints, following the TOGAF ADM Business Architecture phase.

This section describes a checklist-based method for assessing an organization for its DP capabilities.

4.1 Checklist Assessment Method

The Checklist Assessment Method comprises five steps, as shown in Figure 3. It requires a companion checklist document, described in the following subsection. The first three phases deal with setting the organizational context. The two last steps respectively deal with the application of the checklist for determining which DP capabilities are currently deployed in the organization and their current level of effectiveness.

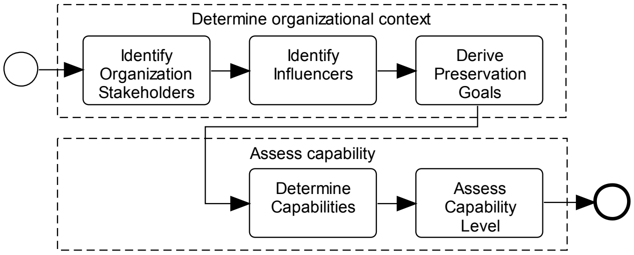

Determine organizational context
Identify Organization Stakeholders
Identify Influencers
Derive Preservation Goals
Assess capability
Determine Capabilities
Assess Capability Level

Figure 3. Checklist Assessment Method

1. Identify Stakeholders: This first step deals with the identification of the stakeholders in the organization referring to the stakeholders defined in the reference architecture [13]. Since stakeholders in the organization might not be named as the ones described, they can be mapped to one or more stakeholders of the organization. For that identification, it is essential that the key questions and concerns of each stakeholder are taken into account.

2. Identify Influencers: After the identification of the stakeholders, it will be possible to identify the influencers according to their concerns. For that, the list of influencers present in the reference architecture [13] should be used. Note that both drivers (which motivate the organization to set its goals) and constraints (which might constrain the deployment of means and the achievement of goals) should be identified.

3. Derive Preservation Goals: The drivers derived in the previous step can then be used for deriving specific preservation goals for the organization. Those goals should be based on the generic goals provided in the reference architecture [13].

4. Determine Capabilities: Then, according to the defined goals and their relationship to the capabilities, the capabilities needed to achieve the goals for the specific case should be determined, using for that purpose the checklist described in the next subsection.

5. Assess Capability Level: Using the checklist, the capability level of a given organization in certain period of time can be verified. The checklist is divided into three main sections, one for each top-level capability (GRC, Business and Support). Then these sections are divided into their constituent sub-capabilities. With results given by the checklist, a gap analysis can be performed to check the current level of capability, compare it with the organization goals or compare between different points in time.

4.2 The Assessment Checklist

Table 4 depicts an excerpt of the capability assessment checklist. The compliance criteria are based on references of the area of DP, especially on TRAC, which were reworked in order to be aligned with the capability approach followed in this work, thus loosing the repository-orientation. In other words, mentions to the concept of repository where removed and when possible, repository-specific criteria were reworked and generalized in order to widen the scope of application to all types of information systems. When the adaptation was not possible, the requirements where still accommodated in the checklist, although with a note stating the conditions to which the criteria apply.

Table 4. Excerpt of the Capability Assessment Checklist

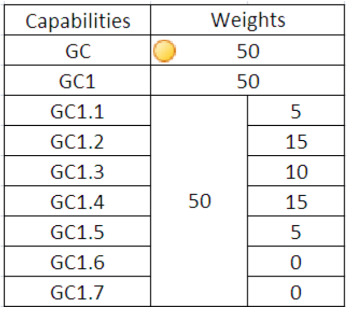
No.	Criteria	Y/N
GC	GRC Capabilities	
GC1	Governance	
GC1.1	The organization has a documented history of the changes to its operations, procedures, software, and hardware.	
GC1.2	The organization has issued a statement that reflects its commitment to the long-term retention, management and access to digital information that is accessible to its community of users.	
GC1.3	The organization has defined the potential community(ies) of users and associated knowledge base(s) required for understanding information.	
GC1.4	The organization has publicly accessible definitions and policies in place to dictate how its preservation requirements will be met.	
GC1.5	The organization has policies and procedures to ensure that feedback from producers of information and users is sought and addressed over time.	

The idea behind the checklist is that any organization of any domain and with any type of information systems deployed can be able to apply it and check its current DP capabilities.

This checklist is available as a spreadsheet, allowing two methods for calculating the compliance level: automatic, which is a linear method; and custom in which we can define the weights for each criterion.

Each capability group is measured from 0% to 100% of compliance. Then each sub-capability has a maximum percentage which in the custom evaluation method can be defined. For instance, if we want the Governance capability (GC1) to weight 50% of the Governance Capability (GC) group, then we can add the weights 32% for the Risk capability (GC2) and 18% for the Compliance capability (GC3) (Note that the total amount for GC, GC1+GC2+GC3, has to be 100%). If we want to define custom weights for the GC1 criteria, for example, GC1 has a maximum of 50%, so we want GC1.1 to weight 5%, GC1.2 to weight 15%, GC1.3 to weight 10%, GC1.4 15%, GC5 5% and the others 0%. Finally, we want GC2 and GC3 to be calculated evenly between the criteria. Figure 4 depicts the customization of GC1. The compliance levels can also be adapted using the table pictured in Table 5.

In order produce a gap analysis with the results achieved, the organization's compliance level target for each capability must be provided in the 'questionnaire' spreadsheet, as an organization might set its own goals concerning the deployment of capabilities due to a variety of reasons (e.g., cost, schedule, etc.) This is pictured in Figure 5.

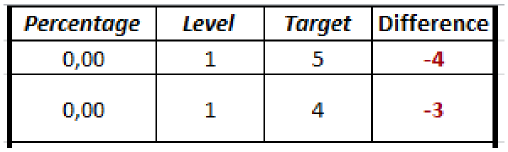


Capabilities	Weights	
GC	● 50	
GC1	50	
GC1.1	50	5
GC1.2		15
GC1.3		10
GC1.4		15
GC1.5		5
GC1.6		0
GC1.7		0

Figure 4. Assigning Weights to Capabilities

Table 5. Compliance Levels Configuration

Levels		
Levels	Percentage	
	Min.	Max.
1	0	25
2	26	45
3	46	65
4	66	80
5	81	100



Percentage	*Level*	*Target*	Difference
0,00	1	5	-4
0,00	1	4	-3

Figure 5. Gap Analysis Configuration

After filling the questionnaire, results can be observed by the means of spider graphs. Figure 6 depicts the compliance levels of a fictional company, the organization XYZ, determined using the companion checklist. In the top-left we can see the global compliance level regarding the three main capabilities depicted in this document. The additional graphs depict the compliance levels for each of the three top-level capabilities. There are three lines in each of these figures: one for organization's target which is the compliance level that the organization wants to achieve, another line for the first compliance check (start) which is the result achieved by the organization on the first compliance check, and finally, another line for the actual compliance level which should be refreshed through time in each compliance check. The main goal here is for the stakeholders to check periodically if their concerns are being correctly addressed through time.

5. ASSESSMENT APPLIED TO TWO E-SCIENCE INSTITUTIONS

e-Science concerns the set of techniques, services, personnel and organizations involved in collaborative and networked science. It includes technology but also human social structures and new large scale processes of making science.

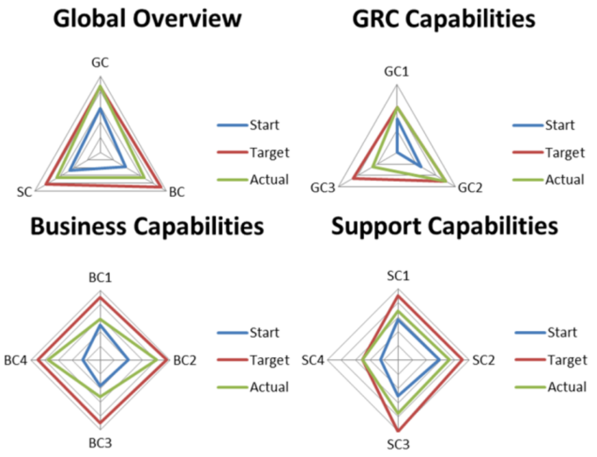


Figure 6. Compliance Graphs

DP is recognised as a required property for future science, to assure communication over time, so that scientific information that is understood today can be reused in the future to produce new knowledge [16].

To conduct a systematic assessment of the preservation capabilities of research organizations, the checklist assessment method was applied to two selected institutions with preservation scenarios dealing with e-Science data: a Civil Engineering (structure monitoring data) and high-energy physics (experimental data). A meeting was scheduled with both groups of stakeholders in which the issues surrounding DP in e-Science scenarios were described along with the reference architecture. After that, the stakeholders were asked to apply the checklist assessment method.

5.1 High Energy Physics Institution

The institution approached is responsible for several studies in the fields of high energy physics (HEP). It is also engaged in large scale experiments occurring in the context of international collaborations. Due to the special characteristics of each of those experiments and the associated costs, it is highly unlikely that the data obtained in that context can be fully reproduced in a new experiment. This fact presents a motivation for preserving this data, since with the development of new theoretical models it might highly relevant to perform a reanalysis of the produced data. The preservation of this data is a challenging task due to the fact that researchers of participating institutions perform local analysis of that data, using for that purpose specific data software which might make use of custom modules developed by the researcher himself, data analysis tools, simulation software, and other scripts. Each of the steps in the analysis might produce different types of intermediate data, each one stored in a determined format.

Table 6 depicts an excerpt of the checklist that was filled by a HEP stakeholder for the Risk capability. The "x" indicates that the criterion is being fulfilled, and the "0" indicates otherwise. We see that two criteria are not met by the organization.

The overall results of the assessment for the high energy physics scenario can be observed in Figure 7. Since this is in fact a first assessment, only the Start and Target lines are displayed. The global overview indicates that Support capabilities are at the level 4 out of 5 of compliance, while Governance and Business capabilities are at level 2 out of 5 of compliance. Through the observation of the GRC capabilities graph, it is possible to see that the governance and compliance capabilities are at a very low level. The Business capability graph indicates that the Preserve Contents capability is almost non-existent, while the Secure Bitstreams capability is at the level 4 out of 5. Finally the Support capabilities graph shows that the Manage Data and the Manage HR capabilities need improvement.

Table 6. Risk Capability Assessment for the HEP Institution

GC2	Risk	
GC2.1	The organization has ongoing commitment to analyze and report on risk and benefit (including assets, licenses, and liabilities).	x
GC2.2	The organization has a documented change management process that identifies changes to critical processes that potentially affect the organization and manages the underlying risk.	0
GC2.3	The organization has a process for testing and managing the risk of critical changes to the system.	x
GC2.4	The organization has a process to react to the availability of new software security updates based on a risk-benefit assessment.	x
GC2.5	The organization maintains a systematic analysis of such factors as data, systems, personnel, physical plant, and security needs.	x
GC2.6	The organization has implemented controls to adequately address each of the defined security needs.	x
GC2.7	The organization has suitable written disaster preparedness and recovery plan(s), including at least one off-site backup of all preserved information together with an off-site copy of the recovery plan(s).	0

It is possible to conclude from the analysis that the knowledge about the implications of DP was somewhat lacking: The organization has a strong capability level for securing bitstreams, the capability of performing the logical preservation of objects is at a very low-level. This is also noticeable in the fact that the capabilities concerning the governance and compliance of preservation are also very low, which indicates that top-level management is not aware of the need to perform effective preservation of the scientific production.

5.2 Civil Engineering Institution

The civil engineering institution approached is responsible for the monitoring of large civil engineering structures to ensure their structural safety, which is achieved through the usage of automatic and manual data acquisition means for real-time monitoring and automatically trigger alarms, when needed. The collected data is then transformed and stored in an information system where it can be later accessed and analyzed. The motivation for preserving this data comes from different aspects such as the fact that it is unique and cannot be produced again, legal and contractual compliance issues are involved, and that its future reuse is highly desirable since new research on the behavior of structures can be performed. The preservation of this data raises several challenges due to the fact that a large variety of sensors are used, making use of different representations for organizing data, and that a large variety of data transformation algorithms can be applied to data.

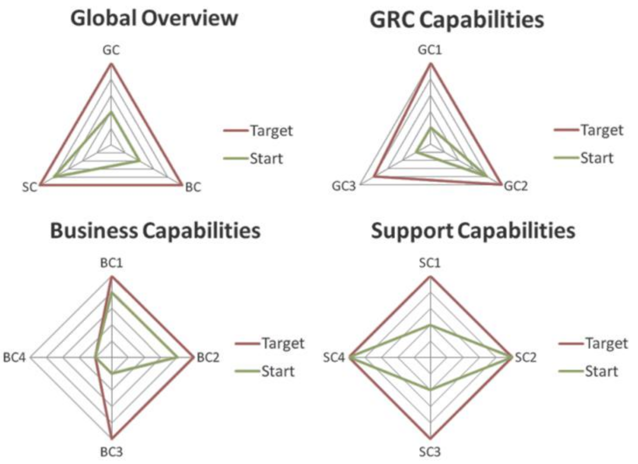


Figure 7. Compliance Assessment for the High Energy Physics Institutions

Table 7. Secure Bitstreams capability assessment for the civil engineering institution

BC2	Secure Bitstreams	
BC2.1	The organization provides an independent mechanism for audit of the integrity of all the data.	x
BC2.2	The organization implements/responds to strategies for the secure storage of objects and storage media migration in order to perform bitstream preservation of digital objects.	x
BC2.3	The organization actively monitors integrity of digital objects.	x
BC2.4	The organization reports to its administration all incidents of data corruption or loss, and steps taken to repair/replace corrupt or lost data.	x
BC2.5	The organization has effective mechanisms to detect bit corruption or loss.	0

Only operational stakeholders were available for applying the checklist assessment, which limited the assessment to the business capabilities. Table 7 depicts an excerpt of the checklist filled by a civil engineering stakeholder for the Secure Bitstreams capability. Only one of the criterions was not being filled.

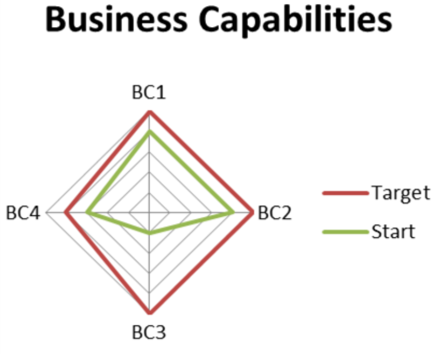


Figure 8. Assessment of Business Capabilities in the Civil Engineering Scenario

Figure 8 depicts the results of the assessment of business capabilities. The assessment determined that the Preserve

Contents capability is almost non-existent, while the Disseminate Content capability needs improvement. From the analysis of the results, it can be concluded again that the knowledge about what sets DP apart from bitstream preservation is very low, since despite having high bitstream preservation capabilities, the capabilities concerning logical preservation are very low. This might be the potential reason for also having low content dissemination capabilities.

6. CONCLUSIONS AND OUTLOOK

This article presented and evaluated a checklist-based method for capability assessment in digital preservation. The method presented is based on a capability-based reference architecture for DP that aims to provide guidance in the assessment and integration of DP capabilities into the information systems of organizations. For that purpose a checklist aimed to be used together with the method was described. The checklist provides sets of criteria for each DP capability which then can be used for evaluating the current level of the DP capabilities of an organization and the gap between current and desired capability, and in that way determining which strategic options can be taken in order to improve capability levels. It was implemented in a way that it can be configured by the stakeholders, allowing changing the weights of the criteria according to the concerns of the stakeholders of the organization being assessed.

The implemented checklist was then applied to two institutions dealing with the need for preserve e-Science data: a High Energy Physics institution and a Civil Engineering institution. From the results of the application, we can conclude that the knowledge of the implications of the logical preservation of data is not well known, despite the existence of bitstream preservation capabilities. This is a commonly observed phenomenon, since many organizations are moving step-by-step from physically securing bitstreams to ensuring continued access to the encoded information. The state of capabilities is also reflected on the level of the governance and compliance capabilities which indicates that the issue is mainly seen as a technological issue, disregarding all the policy aspects that are so important to DP.

The application of the checklist to the two institutions was considered valuable by the involved stakeholders, as it raised awareness of the different aspects involved in the preservation of data. Additionally, the resulting assessment provided an overall picture of the current DP capabilities. Nonetheless, despite providing hints about the possible solutions to the identified gaps, the assessment does not provide concrete and clear answers in terms of solutions to the identified issues. Due to recognizing that need, current and future work focuses on the development of techniques for the modeling and visualization of DP capability patterns so that capabilities can be designed and implemented based on a capability pattern catalog after an assessment has been performed.

7. ACKNOWLEDGMENTS

This work was supported by FCT (INESC-ID multiannual funding) through the PIDDAC Program funds and by the projects SHAMAN and TIMBUS, partially funded by the EU under the FP7 contracts 216736 and 269940.

8. REFERENCES

[1] ISO 14721:2010. Space data and information transfer systems – Open archival information system – Reference model. 2010.

[2] Becker, C., Antunes, G., Barateiro J., Vieira R. 2011. A Capability Model for Digital Preservation. In *Proceedings of the iPRES 2011 8th International Conference on Preservation of Digital Objects*. (Singapure, November 01 – 04, 2011).

[3] Cloutier, R., Muller, G., Verma, D., Nilchiani, R., Hole, E., Bone, M. *The Concept of Reference Architectures.* Wiley Periodicals, Systems Engineering 13, 1 (2010), 14-27.

[4] RLG-NARA Digital Repository Certification Task Force. 2007. *Trustworthy repositories audit and certification: Criteria and checklist.* OCLC and CRL, http://www.crl.edu./sites/default/files/attachments/pages/trac_0.pdf (accessed 18 May 2012).

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

[6] ISO 16363:2012. Space data and information transfer systems – Audit and certification of trustworthy digital repositories. 2012.

[7] Bishhoff, L., Rhodes, E. 2007. Planning for Digital Preservation: A Self-Assessment Tool. Northeast Document Conservation Center, http://nedcc.org/resources/digital/downloads/DigitalPreservationSelfAssessmentfinal.pdf (accessed 18 May 2012).

[8] Mansour, C. 2008. *IT Governance and COBIT.* Presentation. ISACA North of England Chapter, http://www.isaca.org.uk/northern/Docs/Charles%20Mansour%20Presentation.pdf, Leeds. 2008.

[9] IT Governance Institute. *COBIT 5 – A business Framework for the Governance and Management of Enterprise IT.* 2012.

[10] Center for Advanced Research on Language Acquisition. Evaluation – Process: Checklists. University on Minnesota, http://www.carla.umn.edu/assessment/vac/evaluation/p_3.html, Minneapolis, MN. 2012.

[11] The Open Group. *TOGAF Version 9.* Van Haren Publishing. 2009.

[12] SHAMAN Consortium. *SHAMAN Reference Architecture v3.0.* http://shaman-ip.eu/sites/default/files/SHAMAN-REFERENCE%20ARCHITECTURE-Final%20Version_0.pdf. 2012.

[13] Bote J., Termens M., Gelabert G. *Evaluation of Healthcare Institutions for Long-Term Preservation of Electronic Health Records.* Springer-Verlag, Centeris 2011, Part III, CCIS 221, 136-145. 2011.

[14] Ashley L., Dollar C. *A Digital Preservation Maturity Model in Action.* Presentation. PASIG, Austin, TX. 2012.

[15] Rogers B. *ISO Audit and Certification of Trustworthy Digital Repositories, Part II - IT Assessment Methodologies.* Presentation. PASIG, Austin, TX. 2012

[16] *Data's Shameful Regret* [Editorial]. Nature, 461, 145. 2009.