Risk Assessment in Digital Preservation of e-Science Data and Processes

Sara Canteiro
INESC-ID
Rua Alves Redol, 9
Lisbon, Portugal
s.canteiro@gmail.com

José Barateiro
INESC-ID, LNEC
Rua Alves Redol, 9
Lisbon, Portugal
jbarateiro@lnec.pt

ABSTRACT
Risk is a constant in every area and at all levels of any organization, whether in a general context or in a specific activity, project or function. Risk Management comprises a set of coordinated activities to direct and control an organization with regard to risk. Risk Assessment is considered the most important phase of Risk Management, which consists in identifying, analyzing and evaluating risks. Digital preservation's main concern is to keep information accessible and understandable over a long period of time, through means of digital objects; therefore, it is an area that needs a thorough Risk Management and, especially, a thorough Risk Assessment. In fact, the digital preservation process can be seen as Risk Management activities to protect digital information from inherent threats and vulnerabilities. The digital preservation problem can be even more complex in the context of e-Science, which is progressively being considered as a reference method for experimental scientific discovery, and whose data and processes need to be handled and preserved. As such, this paper analyzes the applicability of Risk Assessment techniques, in the context of digital preservation and, more concretely, in the preservation of e-Science data and processes, in order to develop a Risk Assessment method that can be applied while managing the life-cycle of digital information.

Categories and Subject Descriptors
H.3.7 [**Digital Libraries**]: System Issues

General Terms
Management, Measurement.

Keywords
Risk Management, Risk Assessment, Digital Preservation, e-Science.

1. INTRODUCTION
Risk can be seen as the effect of uncertainty on objectives [2]; it is usually quantified as the combination of the probability of occurrence of an event and its consequences. Risk is everywhere and in everything we do, therefore, it is thoroughly necessary to rely on Risk Management (RM) to help us perceive and control risks. RM is constantly evolving and follows specific processes that can be applied to several contexts. Generic standards [1], [2], [3] can point us in the right direction when dealing with risk. However, one must keep in mind that, even though these standards can guide us in the right direction, they cannot give us an universal approach to RM, since every case is unique and has a different background.

Digital preservation (DP) is a blooming concern. Projects are being developed worldwide towards reaching the goal of maintaining digital objects (and the information they contain) accessible and understandable to users for long periods of time, and all the while making sure that both the integrity and the authenticity of these objects are upheld. To reach that, careful planning must be put in practice, clear objectives on which information to preserve and what level of protection it needs must be considered and the characteristics of the preservation environment must be established.

The achievement of DP objectives is a process, since there are numerous threats and vulnerabilities that can affect the ultimate objective of digitally preserve objects. Moreover, it also encloses several challenges to the preservation process itself, so, it needs a firm and trustworthy way to assess and treat the involved risks.

These risks increase when considering data and processes in the e-Science (or enhanced science) context. E-Science represents an alliance between science and IT; it is a collaborative and data-intensive approach, which comprises, besides the data itself, the technological infrastructure to support such huge amounts of information [9]. This is a growing area, and a growing reference on how to make scientific discoveries as well. It is collaborative science, and, consequently, deals with both large and complex raw data sets and information collections. As such, obtained data and employed processes must be digitally preserved for future reference, and this information's life-cycle must be thoroughly managed. Thus, the need for a comprehensive and methodological way to assess risks in this type of initiatives is a critical concern.

The worked presented in this paper was developed with the purpose of achieving a methodological way to assess risks in DP and, specifically, in the DP of e-Science data and processes. It went through understanding which risk assessment techniques are adequate in this context, and how they can be used and combined in order to reach a thorough method to apply known risk assessment techniques to this particular domain. The resulting risk assessment method can be, in the future, combined with DP techniques, meant to treat the assessed risks.

This paper is structured as follows. Section 2 outlines related approaches and standards in the areas of RM, DP and e-Science data and processes. Section 3 limits the problem addressed in this paper, while Section 4 presents the proposed approach to assess risks in the digital preservation of e-Science data and processes.

iPRES2011, Nov. 1–4, 2011, Singapore.

Finally, Section 5 lists the main conclusions of the presented research work.

2. RELATED WORK

The major areas of RM, DP and e-Science converge in the work presented in this paper. We discuss the main approaches and standards adopted in each area to provide an overview of their body of knowledge.

2.1 Risk Management

On a daily basis, we are presented with challenges, there is always a certain degree of uncertainty and even a previously established system, process, activity or operation can be exposed to new and emerging threats and vulnerabilities that could compromise our objectives. This is the very definition of risk (see Figure 1), the effect of uncertainty on previously set of objectives, combining the probability of an event's occurrence and the consequences it may cause.

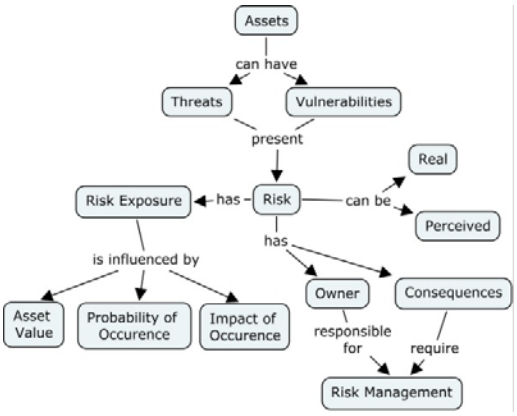

Figure 1 – Need for Risk Management

RM, which can be defined as a set of coordinated activities to direct and control an organization with regard to risk [1], and whose main goal is to define prevention and control mechanisms to address the risks attached to specific activities and valuable assets [4], should therefore be considered as an essential part of every organization and every project it may take on. RM should be iterative, not only applied while developing a project but also while operating and maintaining the resulting product [5], making sure changes that emerging risks are properly addressed.

Several standards exist in the scope of RM. Probably the most relevant of these standards is the ISO 31000:2009 [1], a set of principles and guidelines that can be used by "any public, private or community enterprise, association, group or individual" [1] when dealing with risk. It has two supporting standards as well: the ISO/IEC 31010:2009 [3], a standard guide describing systematic techniques for risk assessment; and the ISO Guide 73:2009 [2], a guide containing definitions for vocabulary terms related to RM.

Even though there are other prominent standards in this arena, like COSO ERM [10], AIRMIC, ALARM, IRM (AAIRM) [12], M_o_R [11], ISO/DIS 21500 [15], ISO 28000:2007 [16], Value-at-Risk [14], IT Governance Institute's Risk IT Framework [13], and OCTAVE [17], among others, the ISO 31000:2009 is the internationally recognized RM standard; thus, the work presented in this paper is mainly directed by the principles, concepts and guidelines provided in this standard family.

In order to guarantee a successful RM, a systematic RM process (see Figure 2) should be followed, in order to realize not only what the possible risks are, but also to analyze, evaluate and treat them, as well as to establish the context and criteria against which they should be judged. This process must be constantly monitored and reviewed in order to act on possible emerging risks; stakeholders must also be constantly involved in the process.

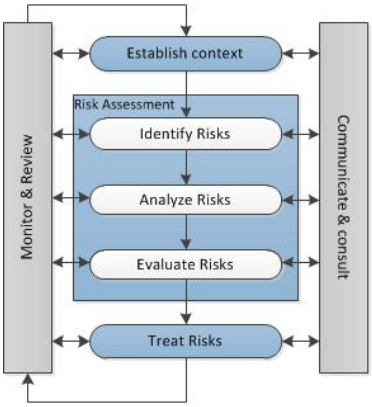

Figure 2 – Risk Management Process [1]

Perhaps the most important task of the whole RM process is risk assessment; and this is the focus of this paper. Risk assessment is not an easy task, it can be very subjective, has a strong dependency from the context where it is to be applied, and has to be a balance between science and judgment and take several psychological, social, cultural and political factors into account [6], which makes it a multidimensional problem. It can be done in either a quantitative, semi-quantitative or qualitative manner and should be as thorough as possible since, if the assessment fails, the subsequent risk treatment will also be inadequate, which may have catastrophic implications.

Assessing risks consists on identifying, analyzing and evaluating them. Risk identification involves ascertaining which events may occur that will jeopardize the normal behavior and/or development of a certain project or activity.

The goal of risk analysis is to understand the identified risks, through a multi-level analysis. There are three main views to risk analysis [3]: the consequence of the risk; the probability that the risk will occur; and the level of risk (combination of its consequences and probability).

The final stage of risk assessment is risk evaluation, where all the information gathered on the previous stages is used, along with the list of criteria produced when establishing the context, to prioritize risks and decide whether or not treatment is necessary.

Several methods and techniques can be used by Risk Assessment. The ISO/IEC 31010:2009 [3] standard surveys 31 techniques to perform Risk Assessment, and shows how they can be applied to each step of the Risk Assessment process as follows: (*i*) risk identification; (*ii*) risk analysis – consequence analysis; (*iii*) risk analysis – qualitative, semi-quantitative or quantitative probability estimation; (*iv*) risk analysis – assessing the effectiveness of any existing controls; (*v*) risk analysis – estimating the level of risk; and (*vi*) risk evaluation.

2.2 Digital Preservation

The main goal of DP is to provide long term preservation and accessibility of digital objects, while maintaining their authenticity and integrity [4].

Throughout time, important information, knowledge and data arise in a digital form, which must not be lost and should, therefore, be preserved for future use (see Figure 3). However, DP poses some serious problems, since not only the original content needs to be maintained, but one must be able to provide evidence that it is authentic, correct and has not been changed.

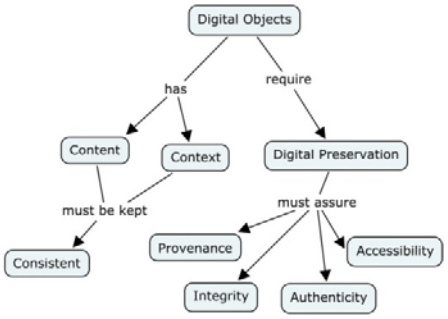

Figure 3 – Preservation Needs

DP aims at preserving digital objects for the long term, making sure the needs of future users are satisfied [7], allowing not only the ingestion and preservation of data, but also its dissemination, making it available to those whom it might concern. Since each type of digital object has its own specific set of requirements, this poses a great challenge, demanding an accurate planning of DP activities.

A common DP environment encompasses all the information entities, the control processes for those entities and the technological infrastructure to support the environment. However, the development of this environment is not a simple chore; not every repository is trustworthy enough to keep such sensitive items and preserve them for the long term, controlling the threats and vulnerabilities involved.

Such a repository must be reliable so as to keep the digital objects intact, accessible and authentic; it must also be flexible, scalable and heterogeneous, as to respond and adjust to emerging changes.

These concerns and requirements should all be taken in consideration while planning the DP process; there needs to be constant monitoring and auditing of this planning process, to make sure the DP plan is adequate to the established goals and requirements, and to make it possible to react to changes whenever they occur. Such monitoring and audit should also be a part of the DP process itself as to keep existing threats and vulnerabilities under control and to discover emerging ones as well, making sure we can timely and adequately react to every new change and challenge.

DP is very challenging to plan and undertake; it has many variables and perspectives to take in consideration. Hand to hand with the challenges come threats and vulnerabilities.

Even though everything is exposed to threats, and everything has vulnerabilities, when it comes to DP, this exposure may be especially dangerous, since we are dealing with information that can be a very sensitive, valuable and powerful asset. This is why these vulnerabilities (see Table 1) and threats (see Table 2) must be assessed from the very planning of the DP venture.

To help in this process, and even though there is not a standard way to approach DP, there are some standards and references that provide principles and guidelines for several steps of the process. The most prominent initiative addressing DP through RM is DRAMBORA [8], which is based on a generic RM process to propose a methodology for self-assessment, encouraging organizations to establish a comprehensive self-awareness of their objectives, activities and assets before identifying, assessing and managing the risks implicit within their organization.

Table 1 – Digital Preservation Vulnerabilities [4]

	Vulnerability	Description
Process	Software faults	bugs that can cause abnormal behavior or even software failure
	Software obsolescence	software becomes obsolete and unable to run or communicate with other components
Data	Media faults	irreversible bit errors (bit-rot) or irrecoverable loss of bulk data (disk crashes or loss of offline media)
	Media obsolescence	representation formats become obsolete and cannot be rendered
Infrastructure	Hardware faults	transient recoverable failures (power loss) or irrecoverable failures (burnt-out power supply unit)
	Hardware obsolescence	hardware becomes obsolete and unable to communicate with other components
	Communication errors	occur while transferring data, these errors might be detected but might also, in some cases such as check-sum errors, go by undetected
	Network services failures	such as DNS and persistent URL errors

Table 2 – Digital Preservation Threats [4]

	Threat	Description
Disasters	Natural disasters	such as earthquakes, floods and fires
	Human operator error	can include both recoverable and irrecoverable errors, such as data deletion; might also involve hardware or software components
Attacks	Internal attacks	malicious users, with privileged access to the organization or physical location of components, may cause: data orcomponent destruction or modification; denial of service; theft
	External attacks	similar to the internal attacks but done over public networks connections; may also encompass attacks such as viruses and worms
Management	Economic failures	budgets are not very stable when it comes to digital preservation, funding may become insufficient over time
	Organizational failures	such as political changes, incompetent management or other unpredictable reason; may lead to changes in what concerns digital preservation requirements, constraints, priorities, …
Legislation	Legislative changes	current processes for digital preservation or preserved data may not obey to the new or revised legislation
	Legal requirements	current processes for digital preservation, preservation environment, repository, and preserved data must obey to the current legislation; if not, legal punishments and fines may take place

2.3 e-Science data and processes

E-Science, which goes through several stages, (see Figure 4) takes science to a new paradigm, a collaborative one, which relies very much on data intensive computing and on community access to distributed data [9].

This new science paradigm comes with a whole new set of challenges, which derive mostly from the colossal amounts of data involved and the ability to share one's scientific information (whether raw data captured from sensors, instruments and/or simulations, or data analysis) and to view and use information shared by other scientists.

Many of the captured scientific data can be unrepeatable (it can be too costly to retake an experiment or even impossible due to external conditions and events); which would make losing that data a potential catastrophe, not only making it impossible to use that same data for further studies, but also any other data derived from it, since it would not be possible to attest to its provenance and authenticity.

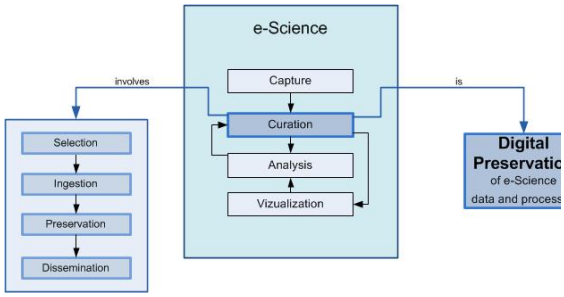

Figure 4 – e-Science activities

The sharing and collaboration aspect of e-Science poses several major issues; one of them is intellectual property. In such an environment, there are those who generate the original data, those who analyze it (possibly generating other resulting data as well), those who use it for research, etc., making it imperative to know where the data came from and who is responsible for it.

Since different analysis methods, workflows and processes can lead to different results and data and, if these methods and processes are not maintained and properly related to the corresponding data, that can lead to potentially mislead research and even misinformed decision making. Along with these workflows, processes and methods, logbooks regarding each experiment (if they are kept) must be duly related to the corresponding information as well.

When considering a digital repository containing e-Science information, one of the main issues is the quality of that information; one expects it to be correct, reliable and trustworthy enough to be useful in research and for further studies and analysis [9].

Though all general DP needs and requirements are maintained in this context, it poses even more demands and requires even more care, since the information might be the target of further exploitation and developments, and is not only meant to be read and consulted in the future.

3. PROBLEM CONTEXT

The information resulting from e-Science processes and workflows has a long life-cycle, which needs a very careful management, in order to assure the properties as well as the content of the information in question.

The DP arena developed several knowledge and best practices, but those concepts have been mainly applied to the cultural heritage sector. The e-Science domain imposes new requirements and raises several challenges on the way this problem should be addressed. In fact, while DP is the main driver of cultural heritage organizations, it must be addressed as an issue (among several other requirements) of the overall e-Science environment, where RM can be seen as a powerful approach to address the potential threats affecting the achievement of DP.

When digitally preserving e-Science information, most of the technological requirements are the same as general digital preservation ones. However, these scenarios come along with the necessity of standard formats and representation, to guarantee future understandability, and make sure the preserved information can be read and used by others in future studies, which also entails the preservation of processes along with the data objects. Also to make it possible for the preserved information to be used in future studies, there is the need to keep a more thorough context than a simple hardware and software one; it is necessary to keep experiments contexts (input parameters, etc.) for them to be able to be reproduced or validated.

While technological requirements of digital preservation are mostly maintained when dealing with complex e-Science scenarios, when it comes to the trustworthiness of the information, the requirements are more specific and require even more attention.

Before any data is ingested, there is the need to make a methodical selection, including a thorough validation of this data to assure no "bad" information, which might potentially taint studies and analysis, is preserved.

The need for authenticity assurance grows even larger when dealing with scientific information, it is absolutely imperative to be sure that a digital object corresponds to the information provided by the original owner, so as to make sure that no information contained in the repository is illegitimate and that digitally preserved data and processes actually correspond to those captured and/or used by scientists. For similar reasons, it is also strictly necessary to attest to the information's integrity for as long as it is preserved, guaranteeing no changes have been made to the informational content.

This need for integrity assurance is all the more pressing when dealing with this type of information, since ingested scientific data should never be subject to change. If the preserved information is used, and changes/additions are made, another version of that information must be ingested and appropriately related to the original one, in order for it to be able to be verified or even reused in the future. No scientific information, regardless of following developments, should be lost or written over, not even in case of discovered errors, bugs, etc., since it might be needed for future consultation or use.

It is necessary that the preserved information is absolutely correct, maintaining these properties, in order for data to be able to be used in further studies, analysis, and experiments or for processes and workflows to be reproduced, for example to confirm results and replicate experiments.

However, some of this information may not be supposed to be accessible for the general public, being restricted to certain entities or communities. Thus, it is necessary that some degree of confidentiality is maintained.

Long-term provenance is imperative to be kept, in order to guarantee not only the ability to identify who is responsible for the information but also intellectual property rights which are obviously important when it comes to scientific discoveries. These properties must be kept not only for captured data, but also for corrections (new versions) made to those data, and data analysis processes, workflows, and results, which may lead to

scientific breakthroughs and must, therefore, be associated with their rightful owners.

These analysis processes and workflows need also to be associated with the original data, as well as posterior results and, in case they are kept, logbooks, each with their own provenance assured, in order to guarantee intellectual property rights of each are maintained along the scientific information's life-cycle.

And this is a very long life-cycle: data and analysis results and processes are not only kept for consultation but can also be the subject of further analysis or studies and, even though the original information is never changed, new and associated information will keep rising.

For the digital preservation of e-Science data and processes to be successful, it is necessary to guarantee that these requirements and needs are met, which makes it imperative to manage possible risks in the most effective and possible way.

However, the use of RM methods in DP is still immature, and there is a lack of guidance to bring and apply the established RM concepts to the DP arena. In fact, despite DRAMBORA [8], a standard way to apply RM to DP does not exist; which would be an added value to the process of preservation, since it could provide specific methods to identify, analyze, evaluate and treat the risks presented in this process, which is becoming more vital with each passing day.

One of the most important phases of RM is Risk Assessment, which consists on identifying, analyzing and evaluating potential risks. Risk Assessment is completely vital to RM in general and DP in particular, since, if the assessment of risks fails, the subsequent treatment will most likely be inadequate, causing the failure of the whole RM process. As such, and, since it is a very complex and extensive area on its own, risk assessment is the main focus of this paper, leaving the treatment of risks as future work.

Since science has always and will always play such a big and important role, a thorough Risk Assessment of e-Science digital repositories is essential. This was one of the main drivers of this work.

Thus, we propose a method to guide Risk Assessment in DP of e-Science data and processes. Its main focus lies on the management of the information's long life-cycle, and it is meant to provide a way to, given a specific scenario in this particular domain, be able to detect and quantify potential threats. This approach can be seen as a complement to generic RM processes or the DRAMBORA approach to DP. It is not an alternative, but a guide for the Risk Assessment activities in DP.

4. PROPOSED APPROACH

The proposed Risk Assessment method was developed through the comprehensive study of known risk assessment techniques (see Figure 5); this study was mostly based on [3] and is meant to complement DRAMBORA [8].

A previous separation of Risk Assessment techniques was made, dividing them into identification techniques, analysis techniques and evaluation techniques, according to which of these Risk Assessment activities they could be applied to. While all the identification techniques were studied with regards to their applicability to the DP context, both the analysis and evaluation techniques were further separated, in order to rule out those that, from the start, were not adequate to the creation of a complete Risk Assessment method. As such, the analysis and evaluation techniques were separated into representative and rating techniques, and the first ones were excluded (when applying a method in a systematic way, these techniques are too subjective, allowing for different interpretations and, consequently, possibly different results when applying this method to the same scenario, thus compromising the correctness of the method itself).

Afterwards, the rating techniques (which can be qualitative, semi-quantitative, and quantitative), along with all of the risk identification techniques, were subjected to a primary and general analysis in order to discard those techniques that, from the start, were not adequate to the scenario at hand. An example of such a technique is the Environmental Risk Assessment, whose scope (people, animals and plants) is completely divergent from the one of this work.

From that point, all the remaining techniques were studied and analyzed in detail, in order to establish their capability of correctly identifying risks (whether known or new), analyzing and evaluating them in each of the different DP of e-science data and processes activities. This was accomplished by verifying the compliance of these techniques with a list of objectives, needs and requirements imposed by the context at hand.

After the individual analysis of each technique was done, a more global study took place. Techniques were compared in order to ascertain the most suitable ones, to be applied in each of the DP of e-science data and processes stages and activities, among the existing possibilities; dependencies between techniques were studied to understand which of these it made sense to combine in each activity of the risk assessment. Even though other techniques may be used, and each case is always a different case, the techniques found in the next subsections are the ones that we recommend to be used in this type of scenario.

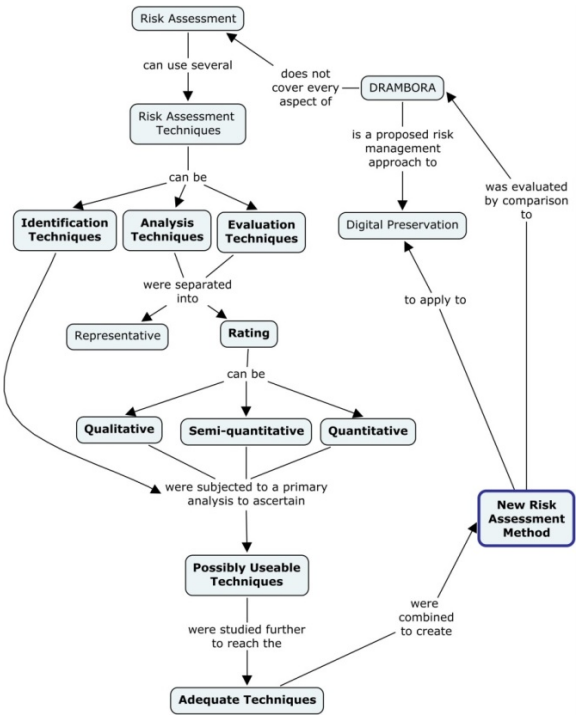

Figure 5 – Developed work

Table 3 – Risk identification techniques

	Context				Types of Risk					Recommended for risk identification
Risk Identification Technique	Problem context & scope	Feasible	Systematic	Comprehensive	Known	New	Human	System	Process	
Check-lists	√	√	√	×	√	×	√	√	√	Yes
PHA	×	–	–	–	–	–	–	–	–	No
Brainstorming	√	√	√	√	√	√	√	√	√	Yes
Interviews	√	√	√	√	√	√	√	√	√	Yes
Delphi Technique	√	×	√	√	√	√	√	√	√	No
SWIFT	√	√	√	√	√	√	√	√	√	Yes
Environmental Risk Assessment	×	–	–	–	–	–	–	–	–	No
Scenario Analysis	√	√	×	×	√	√	√	√	√	No
BIA	×	–	–	–	–	–	–	–	–	No
FTA	√	√	√	×	√	√	√	√	√	No
ETA	√	√	√	×	√	√	√	√	√	No
Cause-Consequence Analysis	√	√	√	√	√	√	√	√	√	Yes
Cause & Effect Analysis	√	√	×	×	√	√	√	√	√	No
CBA	×	–	–	–	–	–	–	–	–	No
MCDA	√	√	×	×	√	√	×	×	×	No
HAZOP	√	√	√	√	√	√	√	√	√	Yes
HACCP	√	√	√	×	√	√	×	×	√	Yes
FMEA	√	√	√	√	√	√	√	√	√	Yes
RCM	√	√	√	×	√	√	×	√	×	Yes
HRA	√	√	√	×	√	√	√	×	×	Yes
SA/SCA	√	√	√	√	√	√	×	√	×	No
LOPA	×	–	–	–	–	–	–	–	–	No
Markov Analysis	×	–	–	–	–	–	–	–	–	No
FN Curves	√	√	×	×	√	√	√	√	√	No
Risk Indices	√	×	×	√	√	√	√	√	√	No
Consequence / Probability Matrix	√	√	√	×	√	×	√	√	√	No

4.1 Risk Identification Techniques

A summary of the carried out analysis, regarding the risk identification techniques and their applicability to the problem context, can be found in

Table 3. The columns of this table have the following purposes:

- 1st column: shows whether or not the technique is applicable to the context at hand as well as to the project's scope;
- 2nd column: shows whether or not it is feasible/realistic the use of that technique in the context at hand (having in mind the possible constraints regarding resources, time, etc.);
- 3rd column: indicates if it can be applied in a systematic manner;
- 4th column: specifies whether the technique is comprehensive when it comes to the potential risks;
- 5th, 6th, 7th, 8th and 9th columns: regard the types of risk which can be identified through the use of that technique (known or new; of human, process, or system nature);
- 10th column: states whether or not it was recommended to be used in the scenario at hand for risk identification purposes.

After the study of all the risk identification techniques, these, in this order, are the ones that we propose to be applied to the DP of e-Science data and processes:

- **Check-lists**, as a preliminary technique, to provide a starting point to the identification of risks, and guarantee no known/common risks to digital preservation are overlooked;
- **Brainstorming**, using a formal process, to have a group of knowledgeable stakeholders gather a list of both known and new risks regarding the scenario at hand in a systematic manner;
- **Interviews**, to target specific stakeholders with the aim to identify "concern-related" risks, and provide further details on risks potentially related to those identified by check-lists and brainstorming;
- **Structured "what-if" technique (SWIFT)**, to be used when change is eminent, particularly taking into consideration the selection, preservation and dissemination stages of the curation process, where change can be more influential, to identify potential risks arising from that change;
- **Failure Mode and Effect Analysis (FMEA)**, to identify design objective deviations and associated risks, potential causes, and consequences, regarding both the curation process and the digital repository itself, while making sure digital preservation's objectives, needs, and requirements have not been neglected;
- **Reliability Centered Maintenance (RCM)**, used along with FMEA, resorting to a specific approach to the latter, in order to identify preventive measures and policies that should be put in place to protect the digital repository, especially regarding the ingestion, preservation, and dissemination phases of the curation process, which are the ones which rely on the repository;
- **Human Risk Assessment (HRA)**, to assess possible human impact on every stage of the curation process;

Table 4 – Risk analysis techniques analysis summary

	Context				Considers			Properties		Recommended for risk analysis
Risk Analysis Technique	Problem context & scope	Feasible	Systematic	Comprehensive	Probability	Consequence	Level of risk	Objective	Possibly Quantitative	
SWIFT	√	√	√	×	√	√	√	√	√	No
RCA	√	√	×	×	√	√	√	√	×	No
Environmental Risk Assessment	×	–	–	–	–	–	–	–	–	No
Scenario Analysis	√	√	×	×	√	√	√	×	√	No
BIA	×	–	–	–	–	–	–	–	–	No
FTA	√	√	√	×	√	×	√	√	√	No
ETA	√	√	√	×	√	√	√	√	√	No
Cause-Consequence Analysis	√	√	√	√	√	√	√	√	√	Yes
Cause & Effect Analysis	√	√	×	×	×	√	×	√	×	No
Decision Tree	√	√	√	√	√	√	√	×	√	Yes
CBA	×	–	–	–	–	–	–	–	–	No
MCDA	√	√	×	×	√	√	√	×	√	No
HAZOP	√	√	√	√	√	√	√	×	×	No
HACCP	√	√	×	×	×	√	×	√	√	No
FMECA	√	√	√	√	√	√	√	√	√	Yes
RCM	√	√	√	×	√	√	√	√	√	Yes
HRA	√	√	√	×	√	√	√	√	√	Yes
LOPA	×	–	–	–	–	–	–	–	–	No
Bow-tie Analysis	√	√	×	×	√	√	√	×	√	No
Markov Analysis	×	–	–	–	–	–	–	–	–	No
Bayesian Analysis	×	–	–	–	–	–	–	–	–	No
FN Curves	√	√	×	×	√	√	√	×	×	No
Risk Indices	√	×	×	√	√	√	√	√	√	No
Consequence / Probability Matrix	√	√	√	×	√	√	√	√	√	used in FMECA

- **Cause-consequence analysis**, to make sure possible underlying and/or consequent risks relating to the previously identified risks are not neglected. This technique can also be used to understand which risks are related among each other.

4.2 Risk Analysis Techniques

When it comes to risk analysis, a summary of the undertaken study regarding their applicability can be found in

Table 4. The columns of this table have the following purposes:

- First 4 columns: the same as those in Table 3;
- 5th, 6th, 7th and 8th columns: refer to whether or not probabilities, consequences, and/or the level of risk are considered by each risk analysis technique;
- 8th column: indicates if the technique can be objective, not giving room for different interpretations in the same situation;
- 9th column: states if the technique in question can be used quantitatively;
- 10th column: states whether or not it was recommended to be used in the scenario at hand for risk analysis purposes.

After the study of the available techniques, the ones that we propose to be used in the context of DP of e-Science data and processes, and the order in which they should be applied are:

- **Decision tree**, to be used considering the selection stage of the DP process, which is where most decisions are made, in order to estimate, for each path coming from a certain decision/event, the value/cost of its outcome, to provide means to later choose the best from the available set of options;
- **Failure Mode Effect and Consequence Analysis (FMECA)**, resorting to the use of a **consequence/probability matrix**, to calculate each risk's criticality, in order to both provide the means to later prioritize risks and serve as input to cause-consequence analysis;
- **Reliability Centered Maintenance (RCM)**, used along with FMECA, resorting to a specific approach to the latter, to estimate the frequency of each failure that may occur especially in the ingestion, preservation, and dissemination phases of the DP process, in case maintenance is not performed;
- **Human Risk Assessment (HRA)**, to calculate probabilities and possible consequences of human error in the DP process and provide input to cause-consequence analysis;
- **Cause-consequence analysis**, to analyze the possible causal and consequent risks of each of the identified risks, and calculate their probabilities and possible consequences.

4.3 Risk Evaluation Techniques

Regarding the study of risk evaluation techniques (a summary of this analysis can be found inTable 5, where the columns have the same meaning as the corresponding ones in

Table 4), the ones that we propose as most suitable to be used in the DP of e-Science data and processes are (in this order):

- **Decision tree**, to be used considering the selection stage of the DP process, which is where most decisions are made, and choose the best from the available set of options, taking into account the previously made analysis;

Table 5 – Risk evaluation techniques analysis summary

		Context				Properties		Recommended for risk evaluation
	Risk Evaluation Technique	Problem context & scope	Feasible	Systematic	Comprehensive	Objective	Possibly Quantitative	
6	SWIFT	√	√	√	×	√	√	No
7	RCA	√	√	×	×	√	×	No
8	Environmental Risk Assessment	×	–	–	–	–	–	No
9	Scenario Analysis	√	√	×	×	×	√	No
10	BIA	×	–	–	–	–	–	No
11	FTA	√	√	√	×	√	√	No
13	Cause-Consequence Analysis	√	√	√	√	√	√	Yes
15	Decision Tree	√	√	√	√	×	×	Yes
16	CBA	×	–	–	–	–	–	No
17	MCDA	√	√	×	×	×	√	No
18	HAZOP	√	√	√	√	×	×	No
19	HACCP	√	√	×	×	√	√	Yes
20	FMECA	√	√	√	√	√	√	Yes
21	RCM	√	√	√	×	√	√	Yes
22	HRA	√	√	√	×	√	√	Yes
25	Bow-tie Analysis	√	√	×	×	×	√	No
27	Monte Carlo Simulation	×	–	–	–	–	–	No
28	Bayesian Analysis	×	–	–	–	–	–	No
29	FN Curves	√	√	×	×	×	×	No
30	Risk Indices	√	×	×	√	√	√	No
31	Consequence / Probability Matrix	√	√	√	×	√	√	used in FMECA

- **Human Risk Assessment (HRA)**, to be used according to the previously made analysis, by realizing which errors or task failures have higher contribution to risk, so as to establish risk priorities and decide whether or not a risk should be treated;
- **Failure Mode Effect and Consequence Analysis (FMECA)**, resorting to the use of a **consequence/probability matrix**, to prioritize the previously analyzed risks, and decide whether or not they should be treated based on this prioritization;
- **Reliability Centered Maintenance (RCM)**, used along with FMECA, resorting to a specific approach to the latter, to prioritize risks according to the previously estimated frequency of each in case maintenance is not performed;
- **Cause-consequence analysis**, by using the previously analyzed fault trees, present in this analysis, in order to prioritize risks and decide on their treatment based on their estimated probabilities and consequences;

4.4 The Proposed Method

Finally, the chosen techniques and combinations were all put together to create a Risk Assessment method for the DP of e-Science data and processes. This method can be found in Figure 6.

This is a cyclic method, intended to be used as a guide, which allows for the overall assessment of risk in this domain, focused on providing a tool to help in the management of this information's life-cycle, supplying the means to identify, analyze and evaluate risks throughout this life-cycle.

The proposed method follows the risk assessment phase of the RM process (see Figure 2).

It starts by identifying risks, where the proposed techniques can be used either separately or together (if used together they should follow the proposed order) and result in a preliminary set of documents encompassing a list of identified risks and some attributes of these risks, as well as documents resulting from the used techniques, such as diagrams, tables and figures.

When risk identification is done, risk analysis takes place and, again, the proposed techniques can be used either separately or together and, if used together, they should follow the proposed order; risk analysis results in an intermediate set of documents, including a more complete risk list, with some more attributes, and documents resulting from the used techniques, including diagrams, tables, figures and necessary calculations.

These documents will then be the input to the risk evaluation stage, which, as the previous two stages, can be done through the use of the proposed techniques (either separately or together), and uses the given inputs to decide whether or not the identified and analyzed risks should be treated; this results in a final risk list.

Thus, as an output, besides the intermediate documents containing the figures and results from each of the three risk assessment activities, this method provides a document containing a list of risks, encompassing, for each risk, a set of attributes to describe it.

These attributes encompass: the risk's nature (whether it's a system risk, process risk, human-related, etc.); the phase(s) of the DP process where the risk may arise; the techniques used to assess the risk; the risk owner (the one responsible for it, from the moment it is identified); affected stakeholders; related risks; probability of occurrence; risk's consequence (only regarding the potential loss of digital objects); the resulting level of risk (combination between the probability and consequence); the date its assessment was completed; the risk's priority.

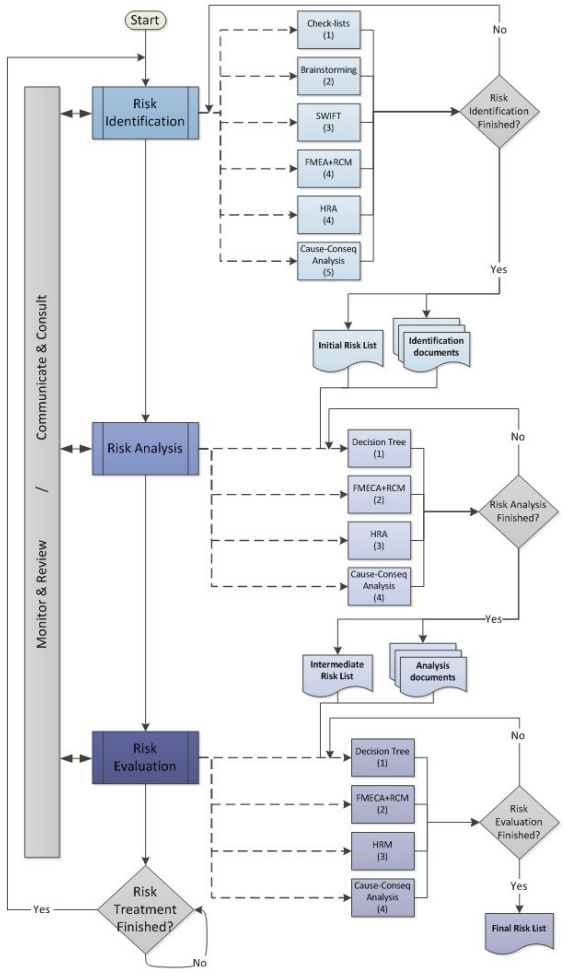

Figure 6 – Proposed Risk Assessment Method

This method provides a comprehensive way of assessing risks in scenarios of DP of e-Science information, from their identification, to their analysis and evaluation. It provides guidance when it comes to the more suitable risk assessment techniques to be used in this context, along with how they may be combined to be as complete and thorough as possible, indicating the best means to aid in identifying a broader range of risks (regarding all the elements involved in DP), analyzing and evaluating them.

4.5 Results and Evaluation

To evaluate the proposed method, a concrete e-Science scenario was used. This scenario concerns LIP[1], a scientific and technical laboratory of particle physics.

A commonly used software to simulate experiments in the high energy physics and astroparticles arena is CORSIKA[2], which is a modular program and requires that each different simulation follows a specific process (see Figure 7).

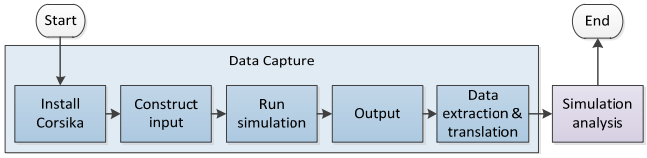

Figure 7 – CORSIKA simulation and analysis process

Several factors may influence this process's outcome, for instance:

- The decisions made can influence the entire process as well as the outcome;
- Changes in the CORSIKA software version may affect the simulation's output;
- Different parameters in CORSIKA installation may affect the simulation's output;
- Different options in the input file may affect the simulation's output;
- Different translations (possible due to the ambiguous manual) may originate different data;
- Different analysis may originate different data.

Some of these simulations can be too costly to reproduce (some run for a long time and have large outputs) and the generated program and outputs must be kept for as long as possible, in order to be able to use the data and verify conditions and results. Hence, the need for digital preservation arises and, along with it, a whole new set of risks.

Through the use of the proposed method, it is possible to assess risks in this particular scenario, in a comprehensive way, by identifying risks which are not as commonly found in known digital preservation risks (as those identified in DRAMBORA reports [8]), analyzing and evaluating them.

This specific case did not call for the use of all of the proposed risk assessment techniques, since it was a fairly simple scenario, to be considered prior to any preservation effort, strictly on a theoretical basis, for the time being. Thus, a simple technique could be used and have a thorough result all the same.

As such, the following examples of possible risks were identified through brainstorming and analyzed and evaluated through FMECA (these risks' probabilities, consequences and levels of risk are represented in Figure 8 by means of a consequence/probability matrix):

R1 – Loss of data translation information, a system/process risk which can arise during the preservation or dissemination stages of the DP process and affects those wanting to analyze data. Since this risk was categorized as Level III, it should be treated as soon as possible.

R2 – Loss of relationship information between preserved analysis processes/workflows and the original data, a system/process risk which can arise during the preservation or dissemination stages of the DP process and affects future results confirmation. Since this risk was categorized as Level II, it should be monitored to see if the risk escalates, in which case treatment might be needed.

R3 – Loss of CORSIKA input parameters for a given simulation, a system/process risk which can arise during the preservation or dissemination stages of the DP process and affects future simulation recreation. Since this risk was categorized as Level II, it should be monitored.

R4 – Selection of incorrect information due to erroneous data validation, a human/process risk which can arise during the

[1] http://www.lip.pt

[2] http://www-ik.fzk.de/corsika/

selection stage and influence all the future stages of the DP process and affects every analysis, study or consultation made based on that information. Since this risk was categorized as Level IV, it should be treated immediately.

R5 – Loss of CORSIKA software version information, regarding a given simulation, a system/process risk which can arise during the preservation or dissemination stages of the DP process and affects future simulation recreation. Since this risk was categorized as Level II, it should be monitored.

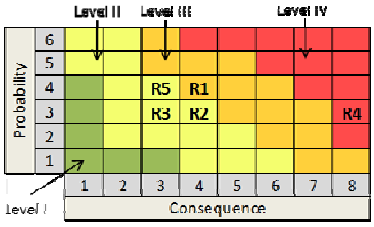

Figure 8 – Levels of risk

Even though these risks can be associated with known needs and requirements of DP (see section 2.2), their specificity prevents them from being listed in commonly used general DP risk checklists.

DRAMBORA [8] provides guidelines regarding digital repositories in general, proposing a methodology for self-assessment, encouraging organizations to establish a comprehensive self-awareness of their objectives, activities and assets before identifying, analyzing and managing the risks implicit within their organization. It has a risk management approach to digital preservation to assess and audit digital repositories.

However, since this is a general approach, to be applied to several digital repositories, it lacks the ability to extend to specific scenarios and, thus, to identify and further assess some unknown/uncommon risks that may rise in these cases.

This is especially evident when it comes to risk identification; since this is the starting point or risk assessment, it should be as thorough and comprehensive as possible. The proposed method recommends several techniques to be used in risk identification, providing a way to identify a wide range of risks instead of only those present in a general check-list.

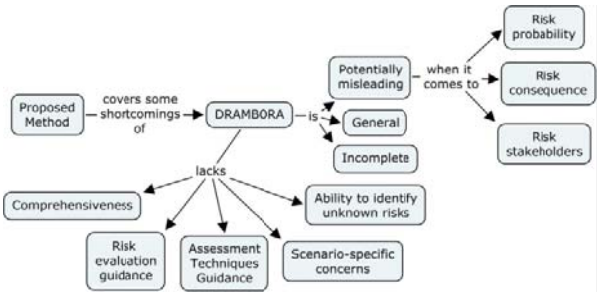

Figure 9 – Proposed method as complement of DRAMBORA

One of the biggest shortcomings in DRAMBORA is exactly this lack of guidance regarding risk assessment techniques. The suggested check-list for risk identification and the level of risk calculation for risk analysis may, in fact, be sufficient to assess some general, common repositories; however, when it comes to more specific cases, and especially when it comes to risk identification, they can be incomplete and not as comprehensive as they should be.

Moreover, the risk list example provided by DRAMBORA, which presents the results of both risk identification and analysis, may be misleading in some cases, given that it lists the same stakeholders, probabilities, and consequences to every single risk.

Another very important shortcoming of DRAMBORA, in what concerns risk assessment, is the lack of guidance when it comes to risk evaluation, not giving any basis on risk prioritization and decisions concerning whether or not to proceed to risk treatment. In fact, this methodology considers only two risk assessment tasks as part of the whole risk management process: “identify risks” and “assess risks”, being that the latter corresponds simply to risk analysis.

All this comes to show the need of a more comprehensive and scenario-specific risk assessment method, as the one proposed in Section 0. This method allows for the identification, analysis and evaluation of both known, general, and new, more particular, risks, that can only be identified through techniques which can be adapted to a given scenario; in this case, the DP of e-Science data and processes.

Thus, the proposed method can be used as a complement to DRAMBORA (see Figure 9), covering its shortcomings, and serving as a guide for each of the three risk assessment’s phases.

5. CONCLUSIONS

RM is an ever evolving area, with application in numerous areas of our lives, businesses and organizations; there is always room for innovation, for creating new and better ways to address risks. Since e-Science’s collected and processed data and used proceedings should be able to be used for future consultation and reference, they must be digitally preserved. This imposes an immense set of risks, regarding both the handling of the data itself and its preservation.

This paper proposes a Risk Assessment method to guide in such efforts, by providing a systematic and comprehensive approach to identifying, analyzing and evaluating potential risks.

The particle physics evaluation scenario, to which the proposed method was subjected, encompassed all 3 stages of Risk Assessment regarding DP. Risk monitoring, communication and treatment tasks fall outside of this work’s scope; however, since the analyzed risks can be mapped into the risk taxonomy presented in section 2.2, this work provides a decision support basis when it comes to evaluating risk treatment controls as well.

The main goal of this approach is to identify wide-ranging risks (regarding the whole DP process, infrastructure, etc.), instead of focusing the assessment strictly on a component level, in order to identify as many risks as possible, to then analyze and evaluate.

6. 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.

7. REFERENCES

[1] ISO/FDIS 31000 (2009) *Risk Management - Principles and guidelines.*

[2] ISO Guide 73. (2009) *Risk management - Vocabulary*.

[3] ISO/IEC 31010 (2009) *Risk management - Risk assessment techniques*.

[4] Barateiro, J., Antunes, G., Freitas, F., Borbinha, J. (2010) Designing digital preservation solutions: a Risk Management based approach. *The International Journal of Digital Curation, Issue 1, Vol. 5*, 2010.

[5] Boehm, B. W. (1991) Software Risk Management: Principles and Practices, *IEEE Software, Number 1, Vol. 8*, 1991.

[6] Slovic, P. (2001) The risk game. *Journal of Hazardous Materials, Issue 86*, 2001.

[7] Doyle, J., Paquet, E., Viktor, H. L. (2007) Long term digital preservation - An end user's perspective. *2nd International Conference on Digital Information Management*, 2007.

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

[9] Hey, T., Tansley, S., Toll, K. (2009) The Fourth Paradigm – Data-Intensive Scientific Discovery, *MS Research*, 2009.

[10] Committee of Sponsoring Organizations of the Treadway Commission (COSO) (2004) Enterprise Risk Management — Integrated Framework, *Jersey City, NJ: AICPA*, 2004.

[11] Office of Government Commerce (OGC) (2007) Management of Risk: Guidance for Practitioners (M_o_R), *United Kingdom*, 2007.

[12] Association of Insurance and Risk Managers (AIRMIC), ALARM (National Forum for Risk Management in the Public Sector), Institute of Risk Management (IRM) (2002) A Risk Management Standard, *London*, 2002.

[13] IT Governance Institute (2009) The Risk IT Framework.

[14] Holton, G. (2003) Value-at-Risk: Theory and Practice, *Academic Press*, 2003.

[15] ISO/DIS 21500 (2011) *Guidance on Project Management*.

[16] ISO 28000 (2007) *Specification for security management systems for the supply chain*.

[17] Caralli, R. A., Stevens, J. F., Young, L. R., Wilson, W. R. (2007) OCTAVE Allegro: Improving the Information Security Risk Assessment Process, *Software Engineering Institute at Carnegie Mellon University*, 2007.