Data Governance in Engineering and Science Projects

Filipe Manuel Lemos Ferreira

Thesis to obtain the Master of Science Degree in

Information Systems and Computer Engineering

Supervisor: Prof. José Luís Brinquete Borbinha

Examination Committee

Chairperson: Prof. José Manuel Nunes Salvador Tribolet

Supervisor: Prof. José Luís Brinquete Borbinha

Member of the Committee: Prof. Artur Miguel Pereira Alves Caetano

June 2014

Acknowledgments

I owe my deepest gratitude to Professor José Borbinha. This dissertation would not have been possible without his guidance and expertise and it was an honor to have him as my advisor. I am also thankful to Ricardo Vieira, Raquel Bairrão and Ana Teresa Freitas for the support given in my thesis.

I am also thankful to Miguel Coimbra and José Barateiro for their participation in interviews and revision of my results, thus adding value to this dissertation.

I am indebted to my family for supporting me throughout all my studies at university. I would like to deeply thank and dedicate this dissertation to my mother Adélia Ferreira and my grandmother Maria Guardado for all their love, support and dedication.

I would also like to thank the scholarship given by INESC-ID and European Committee (through the European project TIMBUS).

Finally, I would like to show my gratitude to my colleagues and friends that directly or indirectly helped me during this dissertation.

Abstract

Engineering and science projects are increasingly data driven, meaning data sets need to be created, stored, disseminated and reused for future use. This raises Data Management concerns and challenges. To address these concerns, funding agencies are currently requiring the creation of Data Management Plans. However, we claim Data Management Plans don't cover all important concerns, since project stakeholders responsible for developing these documents are unable to proper assess if their plan is representative of good data management practices.

To address this issue, we claim actual principles for Data Management Plans can be improved, using Risk Management guidelines and techniques taken from ISO 31000 and ISO 31010. Therefore, we propose a method suited for engineering and science projects that enables the creation of a Risk Management Plan. This new document intends to support all the decisions and measures described in the Data Management Plan. Additionally to the proposed method, a set of stakeholder roles and responsibilities, as well as a set of skills are proposed.

The motivation requirements and evaluation of our work are represented by two scenarios, namely the MetaGen-FRAME project, representing a science scenario and a case of a Civil Engineering Laboratory (LNEC), symbolizing an engineering scenario. To support the analysis of these cases, HoliRisk, a Risk Management tool developed in INESC-ID, is used. As a result, our method proposal constitutes an approach to Data Governance for engineering and science projects.

Keywords: Data Management, Data Management Plan, Risk Management, Risk Management Plan, Data Governance

Resumo

Projectos de engenharia e ciência são cada vez mais orientados a dados onde estes necessitam de ser criados, armazenados, disseminados e reutilizados para uso futuro. Este facto levanta desafios de gestão de dados. Para adereçar estes desafios, organizações financiadoras actualmente requerem a criação de planos de gestão de dados. No entanto, defendemos que os planos de gestão de dados não adereçam algumas preocupações importantes, visto os stakeholders responsáveis por desenvolver estes documentos não são capazes de avaliar se estes representam boas práticas de gestão de dados.

Para adereçar este problema, acreditamos que os princípios de planos de gestão de dados podem ser melhorados usando directrizes e técnicas de gestão de risco retiradas das normas ISO 31000 e ISO 31010. Portanto, nós propomos um método indicado para projectos de engenharia e ciência que permite a criação de um plano de gestão de risco. Este tenciona suportar as decisões e medidas descritas no plano de gestão de dados. Adicionalmente ao método, um conjunto de cargos e responsabilidades de stakeholders e um conjunto de habilidades são propostas.

Os requisitos de motivação e avaliação do nosso trabalho estão representados por dois cenários, nomeadamente o projecto MetaGen-FRAME, representando um cenário de ciência e o caso do laboratório de engenharia civil (LNEC), simbolizando um cenário de engenharia. Para suportar a análise destes casos, é usada a ferramenta de gestão de risco HoliRisk, desenvolvida no INESC-ID. Como resultado, o nosso método proposto representa uma abordagem à governação de dados para projectos de engenharia e ciência.

Palavras-chave: Gestão de Dados, Plano de Gestão de Dados, Gestão de Risco, Plano de Gestão de Risco, Governação de Dados

Table of Contents

List of Figures

List of Tables

List of Acronyms

RM	Risk Management
RMP	Risk Management Plan
DM	Data Management
DMP	Data Management Plan
LNEC	Laboratório Nacional de Engenharia Civil
DP	Digital Preservation
DG	Data Governance

1. Introduction

Data in engineering and science projects frequently is created and transformed according to different methods, tools and schemas, from multiple provenances, and reused in not always expected scenarios. These data sets need to be governed so they can be successfully created, stored, preserved and reused, becoming the organization or project's most valuable resource.

Data Management (DM) addresses the organization, protection and dissemination of data, leading to the development and execution of architectures, policies, practices and procedures that properly manage data's lifecycle (Fernandes, Bakhshandeh, & Borbinha, 2012). With this in mind, to guarantee the governance of data sets in engineering and science projects, DM guidelines become an important requirement.

1.1. Motivation

Science and Engineering projects are increasingly data intensive, high collaborative and highly computational at large scale (Crowston & Qin, 2011). This means data sets are required in order to execute the experiments, workflows and business processes, as well as to support decision making. Data then, became the most valuable resource in these types of projects, not only on a daily basis, but also for future uses. As data is generated, stakeholders are more and more concerned about how that data is created, managed, reused and preserved. This leads to an increase of DM concerns along with demands for efficient and effective DM practices and procedures to support the main activities of (Darlington, Ball, Howard, Culley, & McMahon, 2010):

- Making existing research data fit for a future known research activity (data repurposing);
- Managing existing data such that it will be available for a future unknown research activity (supporting data reuse);
- Using research data for a research purpose or activity other than that for which it was intended (data reuse).

These DM challenges and practices for engineering and science projects are currently being addressed through the definition of Data management plans (DMP) that describe how data is created, collected, stored, managed and disseminated during the project (Fernandes, Bakhshandeh, & Borbinha, 2012), being these DMP requested by the funding agencies of research projects, reflecting the requirements and main concerns of the funding agencies, which in turn leads to the creation of different DMP guidelines. With this in mind, DM challenges in a world driven by data and the current response for these challenges, namely the DMP, are the main motivation that drives our work.

Since DM and DMP concerns conceptually also represent data governance (DG) concerns, principles for a structured and complete DG also become part of the motivation.

The development of this dissertation was motivated by the DataStorm[1] project. DataStorm focus on creating a mass of scientists and engineers for addressing the design, implementation and operation of the new wave of large-scale data-intensive software systems. In order to achieve this goal, first a series of research challenges must be addressed, being one of them, try to understand and improve the proper data lifecycle management processes of projects and organizations that use and require large-scale data sets. Contributing for this objective is where our work becomes relevant.

Our work falls into two distinct projects that helped us demonstrate and validate our proposal. First we used the MetaGen-FRAME project developed by Miguel Coimbra from the KDBIO group. This project is suited for our work, since it represents a good example of a scientific project, where large data sets are accessed remotely and created, existing the interest of preserving data for future scenarios. This project allows us to understand DM concerns and the data lifecycle in a scientific project.

The second case study used was the one of the Portuguese National Civil Engineering Laboratory (LNEC). This case gave us the scope of an organization, being a case that gave understanding of DM concerns and data lifecycles in the field of engineering. This is a suited case since it also creates, stores and shares large data sets, with multiple entities, being the digital preservation (DP) of these data sets a main concern in LNEC. These data sets are not only relevant for day-to-day activities but also for future concerns and decisions. The DP concerns of LNEC have also been analyzed in the scope of the TIMBUS[2] project (Redlich, et al., 2012).

1.2. Problem Statement

Even though DMP try to impose DM good practices, the current guidelines that help create this document are typically restrict to describing the elements that should be included, not applying as much effort as they should into justifying the solutions or explaining how these solutions are decided and what techniques supported their decision. Also, funding agencies that request a DMP usually state that the document will be assessed and will serve as criteria for grant acceptance. However, project stakeholders responsible for creating the document have no way of assessing if their DMP is exemplary of good practices. Since in engineering and science projects low awareness of data management problems is still common, especially in small projects, that hinders the identification of DM solutions required to create a DMP (Crowston & Qin, 2011).

Other aspects that are often overlooked in DM procedures and decisions concerning engineering and science projects, are the potential vulnerabilities, threats and risks that can be associated and endanger the assets (objects of value to the organization or project), like for example data.

One of the main goals of DM is data protection. Viewing data as an asset, risk management (RM), which directs and controls an organization, as its assets, with regard to risk (ISO Guide 73, 2009), also intends to protect data from potential risks and threats. From this point of view, we can see that DM

[1] http://dmir.inesc-id.pt/project/DataStorm
[2] http://timbusproject.net/

and RM share objectives. DM activities can be related to risk control measures for risks associated with data and together, these activities or controls can form risk policies.

With this in mind, it becomes clear that DM concerns are related to RM concerns and that both conceptually can be seen as DG concerns. DMP are currently used to address DM concerns, although not all DG concerns are addressed by DMP, namely the ones related to RM. Consequently, we believe DMP are not enough to address DG issues in engineering and science projects.

In summary, science and engineering projects need a more concrete guidance on how to analyze, identify, assess and control DM problems (Crowston & Qin, 2011).

With this in mind, our work tries to understand how to achieve a more concrete guidance for DM concerns through the improvement of DMP guidelines and good practices in order to promote a more solid DG.

1.3. Contributions

Considering the problem stated before, the contributions of our work were the following:

- **Definition of a DG approach for engineering and science projects**: Our method proposal manages to become a DG approach, through the combine usage of DM and RM principles. In other words, our proposal by unifying DM and RM principles and techniques, and in practice DMP and RMP efforts, completely covers all relevant DG concerns, namely RM concerns not covered by the single DMP, promoting a more solid DG for engineering and science projects.
- **Improvement of data management plans through their complementation with risk management plans**: We improved DMP by using risk management good practices through a RMP created using a method that we propose. In our work we also recommend the structure and content of the RMP so the corresponding DMP and its concerns can be properly supported, since RMP justifies the measures, procedures and controls that are presented in the DMP through the detection of risks, events, vulnerabilities controls and polices.
- **Development of a generic risk management method for the creation of risk management plans concerning engineering and science projects**: In our work we propose a RM method that enables the creation of RMP for engineering and science projects. This method is based on ISO 31000 guidelines and ISO 31010 risk assessment techniques. Finally this method allows the usage of RMP to complement the respective DMP by relating the RM analysis results with the typical requirements and sections of DMP.
- **Development of a set of skills suited for a risk expert**: We believe that our method should be implemented by a risk expert. In order to do this, we suggest a set of skills that we recommend any risk expert should master. These skills are mainly DM and RM skills.
- **Presentation of a set of generic roles and responsibilities**: Generic roles and responsibilities suited for engineering and science projects are presented. These roles are also relevant by contributing to a better DG for these domains.

- **Testing and improving a risk management tool – HoliRisk**: During this dissertation, we assumed a role of beta tester for a RM tool named HoliRisk. This tool is currently being developed, for which our work contributed through the detection of several bugs and other conceptual errors, as well as the indication of several suggestions. In Appendix C a more specific list of contributions and suggestions is presented.

1.4. Research method

Our work was conducted according to the Design Science Research Methodology (DSRM) (Peffers, Tuunanen, Rothenberger, & Chatterjee, 2007) presented in Figure 1.1. This is a methodology that involves six different steps, namely:

- **Problem identification and motivation**: Definition of the specific research problem justifying the value of the proposed solution. This is shown in sections 1.1 and 1.2.
- **Definition of the objectives for a solution**: Infer the objectives of the proposed solution from the problem definition and knowledge of what is possible and feasible. This is shown in detail in section 3.1.
- **Design and development**: Creation of the artifact. Such artifacts are potentially constructs, models, methods, or instantiations. For this dissertation, the artifact is a method. This is shown in chapter 3.
- **Demonstration**: Demonstrate the use of the artifact to solve one or more instances of the problem. This is shown in chapter 4.
- **Evaluation**: Observe and measure how well the artifact supports a solution to the problem. This is shown in sub-section 1.4 and chapter 4.
- **Communication**: Communicate the problem and its importance, the artifact, its utility and novelty, the rigor of its design, and its effectiveness to researchers and other relevant audiences such as practicing professionals, when appropriate. This is shown in section 1.4 for each of the three iterations presented.

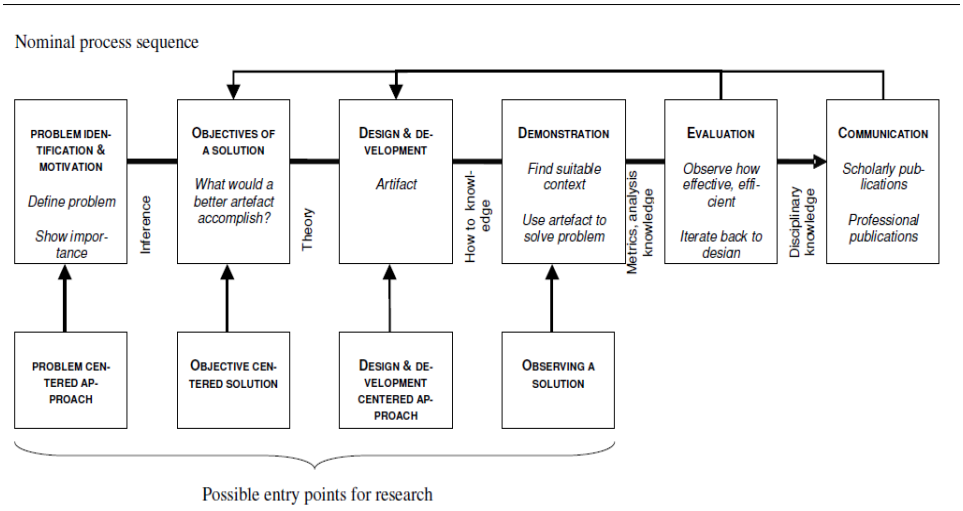

Figure 1.1: The DSRM process model

As it is shown in Figure 1.1 this is an iterative process, and in the scope of this dissertation, several iterations were necessary to produce the final method proposed in section 3.3.

In the first iteration, our motivation was simply the concern of DP in science projects, being this one of the typical DM concerns. Within this motivation, we proposed to improve the DM concern in science projects. For that we proposed to improve the DMP concept, since we considered these didn't address DP as a RM technique, which lead us to believe DMP weren't enough for performing DM and guaranteeing DP. The solution objectives were then the improvement of DMP and to achieve that, RM principles were taken into consideration, being used to change the development of DMP. The artifact was then a method for the development of a DMP which included RM principles and guidelines. To achieve this, DMP typical sections were identified and, for each one, there was an exercise to understand, in which sections RM principles could be used. For demonstration and evaluation purposes, the MetaGen-FRAME project was analyzed, where several risks and controls, considering DP, were identified for the project's workflow. Still for evaluation purposes, several advantages were achieved, comparing with the utilization of standard DMP guidelines. These advantages were the following:

- **Better understanding of the DM problem**: the identification of risks clarifies the value of DM for particular stakeholders thereby providing the rationales for developing DM solutions.
- **Identification and evaluation of alternative DM solutions**: by identifying controls for specific risks alternative DM solutions can be identified systematically. Moreover, thanks to the assessment of risk levels it is possible to evaluate alternative solutions and to choose the preferred one (e.g. if a risk control is capable of mitigating more severe risks).

Finally for communication, a paper was accepted at the InForum 2013 conference (Ferreira, et al., 2013a). The general hypothesis of this iteration's solution is presented in Figure 1.2.

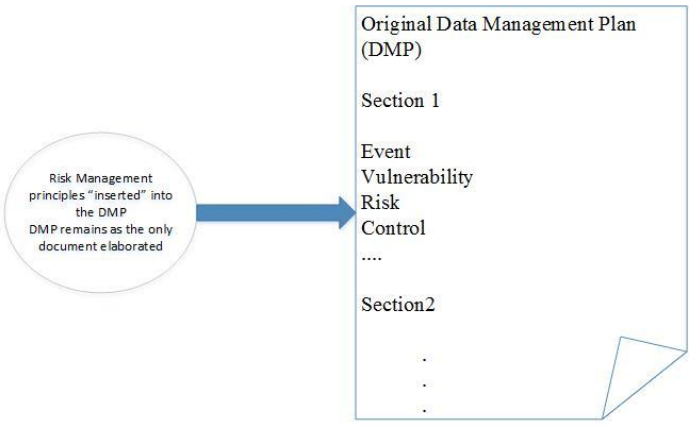

Figure 1.2: First iteration where a single DMP complemented with RM principles was suggested

In the second iteration, DP was no longer seen as the main motivation, but as one of many DM concerns related to science projects, which lead to another iteration, beginning in the first step (defining the problem and motivation), where DM concerns as well as DMP in general became the motivation behind our work.

Our hypothesis also changed, since we understood that creating a single DMP with RM improvements, wasn't the best solution, because librarians and archivists could not adhere to this solution since they don't possess the RM skills required for the new DMP creation method.

With this in mind, we introduced the concept of RMP into our hypothesis, where we proposed the creation of a RMP (using RM principles and guidelines) and a DMP (developed using current guidelines). The RMP would then complement the original DMP (it's important to note an assumption made in this solution which is the pre-existence of the DMP, in other words, we concern ourselves with the development of the RMP and assume the DMP is already developed for a certain scenario). This means that RMP results and content is used to support and justify all the DM decisions, processes and solutions presented in the DMP, being this achieved through the identification of typical sections or requirements of the DMP and then associate RMP results, like risks, events, vulnerabilities and controls, to each of the identified sections.

The new artifact became a generic method for RMP creation on science projects, being this method guided by ISO31000 guidelines and ISO31010 risk assessment techniques (see Appendix B). For demonstration and evaluation purposes, another iteration on the MetaGen-FRAME project was performed. Still for evaluation purposes, the following advantages were found in the application of the new method and in the enlargement of the scope:

- **Finding of a larger number of results**: In the previous iteration, DP was the motivation behind the research which restricted the number of results found. With the scope enlargement to DM, more results were detected.
- **Detection of irrelevant risks**: By associating risks with DMP sections it is possible to determine if the risk is relevant for DM. A risk that does not suit any of the DMP sections is probably out of the DM context and can be discarded.
- **Justification of DMP**: If a control described in the DMP mitigates an identified risk, the risk documents a rationale for describing the control.
- **Solution more friendly for librarians and archivists**: By changing the DMP guidelines and requiring RM knowledge to create the DMP, libraries and archivists would have difficulties and probably would reject the solution. By developing two documents, libraries can continue developing DMP with the same guidelines, being the RMP developed by risk experts. These experts could be outsourced if it becomes more convenient.

Finally for communication another paper was published (Ferreira, et al., 2014) for the IDDC14 conference. This paper was then accepted in the IJDC journal as a general article (the publishing process is not yet completed). A technical report was also developed (Ferreira, et al., 2013b). The generic hypothesis of the second iteration's solution is presented in Figure 1.3.

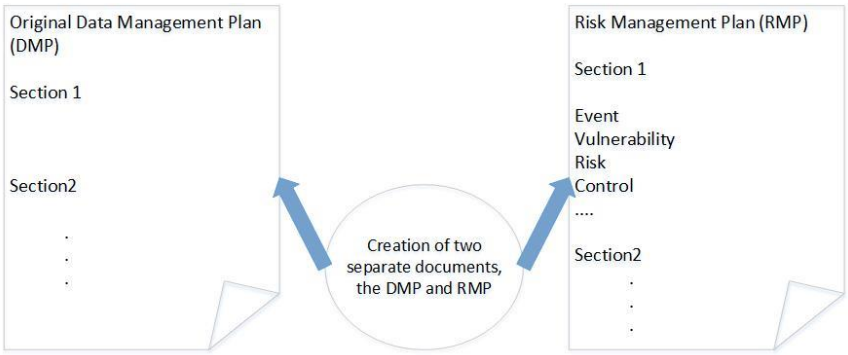

Figure 1.3: Second iteration where the combine usage of two documents, the DMP and RMP, was proposed

For the third iteration, the motivation was expanded to the DM concerns in the engineering field, becoming then the DM concerns related to engineering and science projects. Since both DM and RM, and in practice DMP and RMP, represent different ways to govern assets, like data, to address this concern, we also included DG principles and guidelines into the motivation.

DG refers to the specification of decision rights and accountabilities encouraging desirable behavior in the valuation, creation, storage, use, archival and deletion of data (Oracle, 2011).

The DG principles that we took into consideration in our work, were the following (Council, 2010), (Griffin, 2010), (Khatri & Brown, 2010):

- **Recognize and manage data risks**: engineering and science projects should establish a sound system of data risk oversight and management and internal control, where data controls and policies should be created. Communication and motorization should be permanent.
- **Responsible data usage**: engineering and science projects should establish a code of conduct to take the necessary actions due to legal obligations and confidentiality or ethical issues.
- **Process transparency**: processes should exhibit transparency, being clear to all participants how and when data-related decisions and controls were introduced into the processes.
- **Define data owners and accountabilities for decision-making**: accountabilities should be defined for cross-functional data-related decisions, processes, and controls. Who as decision power and ownership over data becomes clear.
- **Data policies and procedures**: these processes must comprise all data generated, stored and disseminated during the research project. In other words, data policies and procedures should be clearly defined and implemented during all the stages of data management, according to the stakeholder needs and decisions.
- **Data quality**: it is crucial all data used is trustworthy, meaning the source must be known and trusted. Data quality must be assured so good decisions can be made.

- **Data value recognition**: successful quantification of data value, considering the related processes, as well as the time and resources consumed into creating, storing and sharing the same data.
- **Data life cycle**: knowing all the process that move data is crucial, since by understanding how data is used, and how long it must be retained, the research team can develop patterns for optimizing storage media minimizing costs.

The former principles were chosen taking into account the main DM problem addressed in this dissertation and the RM set of principles used to tackle the former problem, being all the former principles suited for engineering and science projects. With this in mind, we understood that our solution proposal represents an approach to DG, by covering all DG concerns that the DMP couldn't cover alone. To address the new engineering motivation, the method proposed was revised, so it could be more generic and so applicable to both engineering and science projects. A new set of skills and a set of roles, with the respective responsibilities, were also suggested. The new skills are important for any risk expert that intends to use our method. We also identified an opportunity for archivists and librarians, since if they master these skills, they can assume, in a near future, the role of risk experts, avoiding the duplication of resources and efforts. The set of roles and responsibilities dictates who is responsible for what in every task performed in the execution of our method. Besides the addition of the new set of skills and roles, a risk registry composed by generic assets, events, vulnerabilities, risks and controls was also identified in order to support the given method. Finally, still concerning the artifact, the composition and structure of RMP, which represents the final output of our method, was also suggested. For demonstration and evaluation, one more scenario was analyzed, namely the LNEC case, which gave us the engineering perspective. Still for evaluation purposes, we applied the method in a case of another field of interest, namely engineering, concluding that our artifact is generic and suited enough to tackle both fields. Also some advantages were obtained:

- **Method validated for two fields**: DM in engineering and science projects can be improved by our method (other fields of application are also possible).
- **A new set of skills**: Relevant skills are presented for a risk expert allowing the application of our method. This represents an opportunity for librarians and archivists, which can in the future, assume this role.
- **Development of RMP improved**: By presenting a risk registry (which can be used to form check-lists) the process of risk assessment becomes more easy, quick and standardized. More vulnerabilities, events and risks were also identified.
- **Set of generic roles and responsibilities**: The new set of roles, and the corresponding responsibilities, helps govern data sets in engineering and science projects. Scope definition becomes easier.
- **RMP structure defined**: The structure and content of the RMP created using the method suggested is presented, standardizing the correct structure for these documents, considering engineering and science projects.

- **Development of a DG approach for engineering and science projects**: The artifact proposed in this dissertation, composed by the method, generic roles and responsibilities, a set of skills and a risk registry, promotes a more solid DG for engineering and science projects, by addressing DG concerns that the DMP did not cover.

Since the LNEC case was previously analyzed considering RM for TIMBUS project (Redlich, et al., 2012), it's possible to compare both approaches, where several similarities and differences were detected (see section 4.2.2).

For communication, an abstract was accepted for the LIBER2014 conference and this dissertation was written as well. One other paper was submitted and is currently being reviewed, for the New Review of Information Networking journal. The generic hypothesis of the third iteration's solution is presented in Figure 1.4.

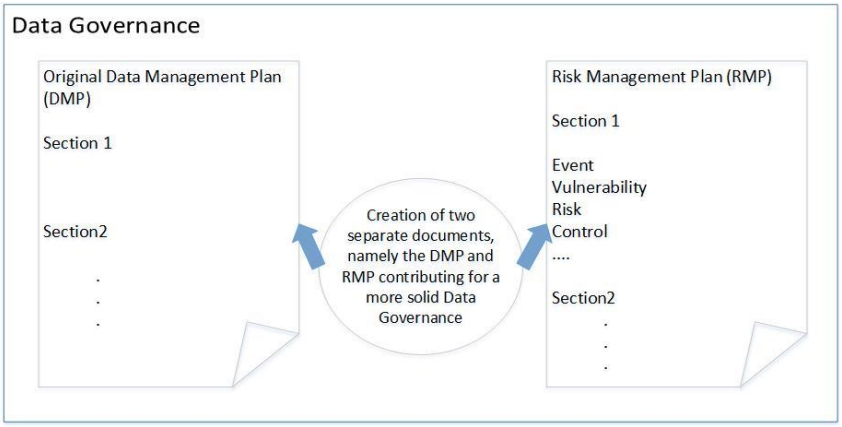

Figure 1.4: Third iteration where the usage of both DMP and RMP promotes a solid DG

1.5. Document structure

This dissertation is composed by five chapters. In chapter 1, the motivation, problem statement, contributions and research method are presented. In chapter 2 the related work is detailed, namely DM in engineering and science projects, DMP, RM, RMP and the potential relations between RM and DMP. In chapter 3 the solution hypothesis and proposal are presented. In chapter 4 we demonstrate of the application of our artifact with two scenarios, the MetaGen-FRAME project and the LNEC case. Finally in chapter 5, some conclusions are presented, some lessons learned are detailed and suggestions for future work are given.

2. Related work

This chapter presents the state-of-the-art related to the several topics relevant to the problem of this dissertation. We start by giving an overview of DM in engineering and science projects. Then, DMP related work. Next we cover the RM and RMP related work. Finally we present a relation between RMP and DMP.

2.1. Data Management in Engineering and Science Projects

Engineering and science projects are increasingly carried out through distributed global collaborations enabled by the Internet and middleware computer resources, requiring access to very large data collections and high performance visualization back to the individual users (Jankowski, 2007), (Nguyen, Guellec, Féru, Maillé, & Yannou, 2007). DM is an integral part of modern research, assuming a vital part in engineering and science projects. DM involves data backups, cooperative work, version control, metadata management, data security, and archiving. Managing data allows researchers to work more efficiently, produce higher quality data, achieve greater exposure for their research, and protect data from being lost or misused (Fernandes, Bakhshandeh, & Borbinha, 2012). DM involves organizing, protecting, and sharing data, (Fernandes, Bakhshandeh, & Borbinha, 2012), which becomes essential in some typical features for engineering and science projects (Jankowski, 2007), (Brennan, 2011), namely:

- International collaborations.
- Increasing use of high-speed interconnected computers, applying middleware architecture.
- Creation, storage, visualization and dissemination of large quantities of data or data sets (data that is growing exponentially).
- Development of Internet-based tools and procedures.
- Diverse source systems.
- Variety of data formats.

Engineering and science projects are normally based on business processes or workflows, which typically sit on top of a middleware layer. They are the means by which the operational or scientific teams can model, design, execute, debug, reconfigure and rerun their analysis and visualization pipelines (Brennan, 2011), (Braga, 2007), (Vermaaten, Lavoie, & Caplan, 2012). They typically involve several steps, access terabytes of data and generate similar amounts of intermediate and final products (Darlington, Ball, Howard, Culley, & McMahon, 2010), (Deelman & Chervenak, 2008). Business processes and workflows have a lifecycle which consists of an analysis phase, a planning phase, where the resources needed are selected, an execution part, where the computations happen, and the result, metadata, and provenance storing phase (Brennan, 2011), (Deelman & Chervenak, 2008). In a typical process or workflow, the data lifecycle includes the following transformations

(Brennan, 2011), (Deelman & Chervenak, 2008), (Darlington, Ball, Howard, Culley, & McMahon, 2010):

- **Data discovery**: setting up the data processing pipeline, generation of derived data, archiving of derived data and its provenance.
- **Data analysis**: can be a collaborative process.
- **Data processing**: execution of tasks or steps that takes data as input and generates new data.
- **Derived data and provenance archival**: storage of data, correspondent metadata and provenance information allowing interpretation and sharing.

There are several DM challenges typical of engineering and science projects (Beagrie, 2006), (Deelman & Chervenak, 2008), (Vermaaten, Lavoie, & Caplan, 2012), (Brennan, 2011); (Nguyen, Guellec, Féru, Maillé, & Yannou, 2007); (Darlington, Ball, Howard, Culley, & McMahon, 2010). These challenges are synthetized in Table 2.1:

- **Business process or workflow's traceability**: Operational teams need to know how each process or workflow has been run and what it produced. It's also necessary to keep track of dependencies among tasks. There's the need to identify the input parameters allowing the rerun of the process or workflow. There's also the need to know which data sets were produced (data's provenance).
- **Data reuse**: When there's collaboration between systems that are controlled by different entities, to assure that data can be reused, the need for a common and unique means of data's representation and communication arises.
- **Data Dependencies**: Earlier tasks in the workflow may produce intermediate data that is consumed by tasks that run later. The challenge includes finding available resources whose capabilities match the requirements of the workflow or process.
- **Intentional attacks**: Middleware systems can be the target of unauthorized access or malicious usage were data can be stolen or modified. These attacks can be originated from internal or external sources.
- **Financial changes**: Lack of financial resources to store and maintain all the data or bankruptcy of the organization responsible for managing the data.
- **Organizational changes**: Political or management changes in the organizations responsible for data preservation can compromise data storage, dissemination and reuse for future needs.
- **Web content erosion**: Sometimes data is collected through web documents and web references. This data also needs to be stored, consulted and shared for future reuse, but this raises the challenge of small life span of web documents and references due to failed or broken links.
- **Data sharing**: Data sets created are meant to be shared and reused, which raises several challenges:
    - **Renderability of data**: Data must be used and reused in a way that is able to retain its characteristics and content for a proper understanding.

    - **Data confidentiality**: It's difficult to share personal and confidential data without compromising the subject's privacy.
    - **Intellectual property**: When sharing data, some individuals can take the credit for the work and data created by others.
    - **Data management**: Sharing data in a distributed environment hardens its physical management causing difficulties identifying the data set's location.
- **Data set's size**: Applications can generate teraByte or petaByte-size data sets, and so, lack of storage size can result in data loss. This challenge can lead to catastrophic consequences. Digital objects can grow into a size that becomes very expensive or even impossible to move, increasing the difficulty of remote access.
- **Data obsolescence**: It's necessary to define what data needs to be saved stored, and also when data becomes outdated. It's vital to address how or who should delete this data.
- **Infrastructure obsolescence**: The software and hardware supporting the workflows and processes can become outdated. Different applications can have their own metadata catalogs. Legacy infrastructures may not have metadata support.
- **Infrastructure faults**: Hardware, software or network faults can compromise the storage, communication or analysis of data and the execution of process or workflows. During execution, data is transferred making it vulnerable in case of failures and server timeouts, leading to corrupted or lost data.
- **Natural disasters**: Data must survive natural disasters like earthquakes, floods or fires.
- **Human errors**: Data can be unintentionally deleted or modified, or the modifications made to a digital object can go unregistered.

Type of challenge	Challenge
Data	Data reuse
	Data Dependencies
	Data sharing (Renderability of data, Data confidentiality, Intellectual property, Data management)
	Data set's size
	Data obsolescence
	Web content erosion
Infrastructure	Infrastructure obsolescence
	Infrastructure faults
Disasters	Natural disasters
	Human errors
Attacks	Intentional attacks (security)
Workflow	Scientific workflow's traceability
Management	Financial changes
	Organizational changes

Table 2.1: Engineering and science challenges

An overview of engineering and science is given in Figure 2.1.

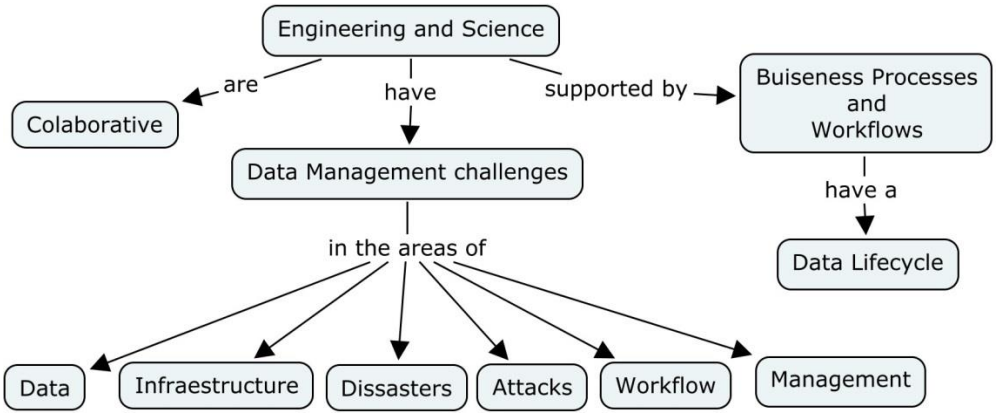

Figure 2.1: Engineering and Science overview

2.2. Data Management Plans

Researches are concerned about several issues, namely identify the nature of the research products (characteristics of the data, samples, physical collections and other materials), share the data and products (experimental materials resulting from research that will be available to others), provide access to data (the ways that researchers will be able to obtain the data and products) and archive the data (the long term storage and organization of the data and research products) (Fernandes, Bakhshandeh, & Borbinha, 2012).

National and international research funders have been increasingly requiring research projects to provide a DMP. A DMP can be defined as a document that describes what data will be created, collected, stored, managed and disseminated during a project (Fernandes, Bakhshandeh, & Borbinha, 2012). DMP represents an opportunity to demonstrate the awareness of good practice and reassure funders that the proposal is in line with their data policy.

Each DMP addresses several requirements, which the most common are (Fernandes, Bakhshandeh, & Borbinha, 2012):

- Which data will be generated during research.
- Metadata, standards and quality assurance measures.
- Plans for sharing data.
- Ethical and legal issues or restrictions on data sharing.
- Copyright and intellectual property rights of data.
- Data storage and backup measures.
- DM roles and responsibilities.
- Costing or resources needed.

Each DMP has two functions (Fernandes, Bakhshandeh, & Borbinha, 2012). On one hand, it acts as a guide to researchers on reusing existing data, repurposing their own data and supporting data reuse. On the other hand, its purpose is to act as a record of how the data have been reused and re-

purposed and how data reuse has been supported. Further, this document provides information about the location, accessibility and ownership of the data associated with an organization or project promoting the second use. A DMP must be a living document, and therefore should be reviewed and updated regularly.

In what concerns DMP guidelines, each University or research collaboration team addresses its own issues to consider in DMP, accordingly to the research that they perform. Several examples of these guidelines are provided.

2.2.1. Digital Curation Center Guideline

The Digital Curation Centre is a world-leading center of expertise in digital information curation with a focus on building capacity, capability and skills for research DM. The Digital Curation Centre has developed a guideline that describes the structure of a DMP as follows (Fernandes, Bakhshandeh, & Borbinha, 2012):

- **Introduction and context**: This section records administrative details, which tie the plan to a particular project.
- **Data types, formats, standards and capture methods**: Why and how the data was created, what represents, and whether it's likely to be compatible with other datasets.
- **Ethical and privacy issues**: Anonymisation of data, referral to departmental or institutional ethics committees or formal consent agreements.
- **Access, data sharing and reuse**: If researchers should share or withhold their data, and how they share it.
- **Short-term storage and data management**: Short-term storage policies made against unintended loss of portable equipment.
- **Deposit and long-term preservation**: Relevant repositories used as an appropriate place of long-term deposit.
- **Resourcing**: Appropriate funds should be allocated to DM. Namely: DM time allocations, project management of technical aspects, training requirements, storage and backup, contributions of non-project staff, etc.
- **Adherence and review**: Everyone should adhere to the DMP. It must be established a communication's plan. The DMP should be kept up-to-date via regularly scheduled review.
- **Statement of agreement**: Possible formalization of the DMP with a statement of agreement.
- **Annexes**.

2.2.2. Australian National University Guidelines

The Australian National University has also developed a guideline to write a DMP, which must be as follows (Fernandes, Bakhshandeh, & Borbinha, 2012):

- **Project description**: Describe the research project.

- **Survey of existing data**: See if there is existing data that could replace or augment the data that is going to be created.
- **Data to be created**: List of all the data that will be created.
- **Data quality and organization methods** (optional): Data quality and organizations methods can be useful if resources are required for IT infrastructure, software, or training.
- **Data administration issues**.
- **Funding and legislative requirements**: List any relevant policies (such as data archiving).
- **Data owners and stakeholders**: List of the owners and stakeholders of the data.
- **Access and security**: List who will have access to the research data and what access permissions they will have for specific data. Describe how the access permissions will be enforced and what IT security practices will be used.
- **Backups**: List what data will be backed up and the backup schedule.
- **Data sharing and archiving**: Lists archiving and sharing policies.
- **File formats, standards, and conventions**: List what formats, standards, and conventions are applied to each data item.
- **Sharing**: List what data will be made available for each researcher to use.
- **Archival and disposal**: Estimate the amount of storage space required for archiving and which archive is used Describe the procedure for the elimination of data.
- **Responsibilities**: List who will be responsible for ensuring each item in the DMP, as well as, who is responsible for reviewing and modifying it.
- **Budget**: Project's DM cost estimation. Includes the time involved in documenting, writing metadata, and archiving, as well as any equipment purchased (such as fileservers, backup media, software, etc.) used for DM.

2.2.3. National Science Foundation Guidelines

All proposals submitted to the National Science Foundation must include a DMP. This foundation has created a series of guidelines relating DMP in several departments. According to the scope of Bioengineering, the guidelines that are relevant are the ones of Engineering and Biology.

National Science Foundation Engineering Guidelines

A DMP in the Engineering area, according to the National Science Foundation, must have the following composition (Fernandes, Bakhshandeh, & Borbinha, 2012):

- **Roles and responsibilities**: Describe who owns the data. Also describes the personnel who will collect and manage the data.
- **Products of Research**: Describes all types of data collected and the amount of data in terms of numbers and disk space. Describes how the data is collected and the instruments that are used to do so.
- **Period of Retention**: Indicates how long the data will be preserved.

- **Data Formats and Metadata**: Describe the data's format; Structural standards that are applied in making data and metadata available.
- Which file formats will be used and why?
- What naming conventions will be used?
- How will the directories be organized?
- How data quality is assured?
- What form will the metadata describing the data take?
- Which metadata standards will be used and why?
- What contextual details (metadata) are needed to make the data captured meaningful?
- **Data Dissemination and Policies**: Describes who is allowed to use the data, how, and if it's possible to disseminate it.
- **Data Storage and Preservation**: Describe both short-term and long-term strategy for storing, archiving and preserving data. Describe any backup systems and versioning that might be used. Indicates how to protect web accessible data from malicious deletion or corruption. Finally indicates if the data be will be stored after the end of the project and where.

National Science Foundation Biology Guidelines

A DMP in the Biology area, according to the National Science Foundation, must follow a similar composition than the former, simply with a different organization, a bigger focus on data dissemination policies and sharing and aimed for a different type of data (Fernandes, Bakhshandeh, & Borbinha, 2012):

- **Products of Research**.
- **Data Storage and Preservation**.
- **Data Formats and Metadata**.
- **Data Dissemination and Policies**: Describes who is allowed to use the data, how, and if it's possible to disseminate it. The following questions must be addressed.
- How and when will the data be available?
- Will any permission restrictions need to be placed on data?
- Who will hold the intellectual property rights to the data and how might this affect data access?
- Will the findings be published?
- **Roles and responsibilities**.

Despite the differences, there are common issues for all these guidelines, allowing the definition of typical sections. Table 2.2 shows these typical sections, as well as a comparative analysis of DMP guidelines defined in four different cases that correspond to the funding agencies of Australian National University (ANU) and National Science Foundation (NSF, from the guidelines for Engineering and Biology) and also the guidelines of the Digital Curation Center (DCC).

Sections	Organizations			
	DCC	ANU	NSF (Eng)	NSF (Bio)
Ethics and privacy	X	-	-	X
Resourcing (Budget)	X	X	-	-
Legal Requirements	-	X	-	X
Access and Sharing	X	X	X	X
Archiving and Preservation	X	X	X	X
Stakeholders / Responsibilities	-	X	X	X
Data Formats and Metadata	X	X	X	X
Data Quality Assurance	-	X	X	-

Table 2.2: Comparison of DMP Guidelines (present(x), not present (-))

- **Ethics and privacy**: A description of how data consents will be handled and how security measures will be taken to ensure data confidentiality and authenticity. A number of ethical requirements applied particularly where the research involves animal subjects. These include the purpose and nature of the research, the nature of consent obtained and what data needs to be safeguarded and destroyed after its use. This raises risks that need to be attended, and a RMP is useful in achieving that goal. In (Kaye, Boddington, Vries, Hawkins, & Melham, 2010) there are several examples of ethical risks of the use of genomes.
- **Resourcing (Budget)**: A description of how costs are assigned to data management. This raises several risks (Redlich, et al., 2012) that can be assessed and mitigated with the help of RM principles and a RMP.
- **Legal Requirements**: A list of all relevant requirements regarding data management. Several risks are related (Redlich, et al., 2012) which RM principles and a RMP can help assess and mitigate.
- **Access and Sharing**: A description of how data will be shared and accessed. Necessary requirements for a successful dissemination of the research results throughout the scientific community, which involves licensing measures; Dissemination incurs a series of risks that must be dealt with a cost-benefit analysis to determine if the value outweighs the liabilities. A RMP would be appropriate to perform such a task. There are several examples of risks associated with data sharing (Bimholtz & Bietz, 2003), (Schmitt & Burchinal, 2011).
- **Archiving and Preservation**: A description of taken procedures to ensure long-term archival and preservation of data. Necessary requirements for storing data, in a short or long term as well as maintaining the data secure from intentional or non-intentional attacks. RM would be helpful in achieving these goals through a RMP; There are several examples of risks associated with data storing and security (Schmitt & Burchinal, 2011).
- **Stakeholders/Responsibilities**: How responsibility will be assigned to DM procedures. Several risks related to stakeholders, as well as their roles and responsibilities (Redlich, et al., 2012) can be assessed and threated with RM principles through a RMP.
- **Data Formats and Metadata**: A description of how file formats and metadata will be applied and maintained, describing the file formats and conventions used, as well as, the metadata form that is applied in the project. RM would be helpful in achieving these goals through a RMP. In (Day, 2004) there are several examples of issues and risks associated with metadata;

- **Data Quality Assurance**: A description of all taken procedures to ensure data quality during the project. Several risks are related (Ferreira, et al., 2013b) (Redlich, et al., 2012) being RM principles and RMP recommended.

Figure 2.2 gives an overview of DM and DMP concepts presented earlier.

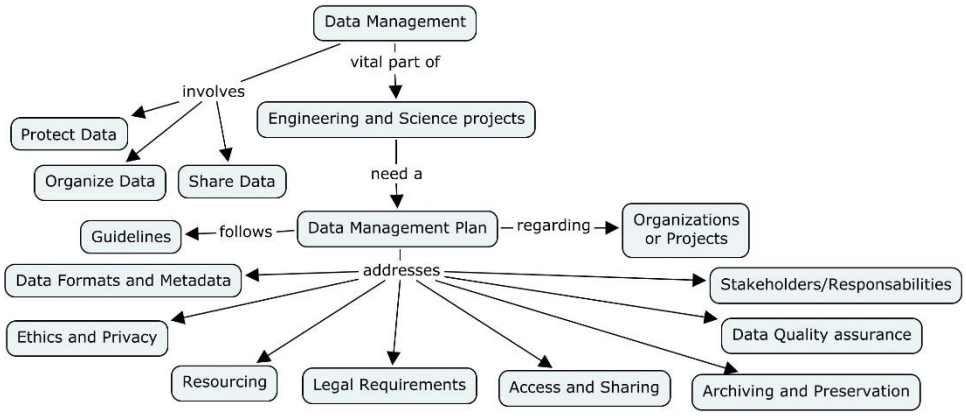

Figure 2.2: Overview for the DM and DMP main concepts

2.3. Risk Management

Internal and external factors and other influences make it uncertain whether and when certain objectives will be achieved. The effect this uncertainty has on objectives and assets (which represent anything of value) is “risk” (ISO Guide 73, 2009), (ISO FDIS 31000, 2009), (Canteiro, 2011). A risk results in an arrangement between the likelihood and consequence of an event/threat occurring, which can be positive representing an opportunity, or negative, representing a possible vulnerability that an asset can have which can be explored by a potential threat. That’s why the identification and treatment of risks must be done as early as possible so they can be successfully managed (Boehm, 1991). Each stakeholder has its risk perception, representing he’s personal view on a risk (ISO Guide 73, 2009). Risks must have a risk owner, which is a person or entity with the accountability and authority to manage those same risks, since they are identified until they are treated or controlled (ISO Guide 73, 2009), (ISO FDIS 31000, 2009).

A set of coordinated activities must be created to direct and control several possible risks (ISO Guide 73, 2009), (ISO FDIS 31000, 2009). Therefore, the RM process aids decision making by taking account of uncertainty and the possibility of future events or circumstances (intended or unintended) and their effects on agreed objectives (ISO IEC 31010, 2009).

RM includes the application of methods for communicating and consulting through continual and iterative processes that share information with stakeholders (ISO Guide 73, 2009), (ISO FDIS 31000, 2009), establishes the context for identifying, analysing, evaluating and treating risk associated with any activity, process, function or product, monitors risks and reports the results appropriately (ISO IEC 31010, 2009). RM not only tries to diminish the event/threat likelihood but also tries to enhance the risks that represent opportunities if the value these risks bring to an organization or project is bigger

than the liabilities. The key contribution of RM is to create a focus around the organization and project's critical success factors, providing techniques to deal with those same factors (Boehm, 1991).

2.3.1. Risk Management Standards and Guidelines

There are standards, methodologies and tools in what concerns RM, depending on the market sector, type of business or organizational activities (Ramirez, 2008). The three main standards are:

- Risk management – Vocabulary. ISO Guide 73. Geneva, Switzerland : ISO (ISO Guide 73, 2009).
- Risk Management - Principles and guidelines. ISO FDIS 31000. Geneva, Switzerland : ISO (ISO FDIS 31000, 2009).
- Risk management - Risk assessment techniques. ISO IEC 31010. Geneva, Switzerland : ISO (ISO IEC 31010, 2009).

There are other relevant standards that need to be mentioned in the RM area spread throughout several fields. These are represented in Table 2.3:

Area	Standard	Description
Security	ISO/IEC 27005 (ISO IEC 27005, 2011)	Guidelines for information technology and security techniques that guide on how to perform an information security risk management
	Risk IT Framework (IT Governance Institute, 2009)	Set of guiding principles that help enterprises identify, govern and effectively manage IT risk
	Operationally Critical Threat, Asset, and Vulnerability Evaluation (OCTAVE) (Caralli, Stevens, Young, & Wilson, 2007)	Helps organizations understand and address their information security risks, focusing on the organizational assets through a self-directed approach
	NIST, Risk Management Guide for Information Technology Systems (NIST, 2012)	Helps organizations understand and address their information security risks
Banking	Value-at-Risk (VaR) (Holton, 2003)	Widely used as a risk measure on specific portfolios of financial assets
DP	Digital Repository Audit Method Based on Risk Assessment (DRAMBORA) (McHugh, Ruusalepp, Ross, & Hofman, 2007)	Proposes a methodology for self-assessment
Enterprise Risk Management frameworks	AIRMIC, ALARM, IRM (AAIRM) (Association of Insurance and Risk Managers (AIRMIC, 2002)	Provide references and guidelines to RM
	Management of Risk (M_o_R) (Commerce, 2007)	Shows how to achieve a good RM practice, which elements and approaches are required which are the critical processes, and which are the reviewing mechanisms that should be used
	Committee of Sponsoring Organizations of the Treadway Commission Enterprise Risk Management (COSO ERM) (Tread, 2004)	Provides guidelines to manage risk enterprise-wide and to define the key enterprise RM principles, concepts and components

Table 2.3: RM Standards

2.3.2. Relevant Risk Management Principles

RM sometimes doesn't work, or produces bad results, because it doesn't respect several principles or because it established the wrong RM policy, which can result in the definition of a wrong RM framework. Therefore, the correct framework and policy must be defined and the properly standard RM process must be met, composed by a proper risk assessment, followed by a proper risk treatment both applied with the correct context and criteria, so RM can detect and treat successfully all risks.

RM can be effective and generate positive results only when several principles are respected. These principles are represented in Figure 2.3:

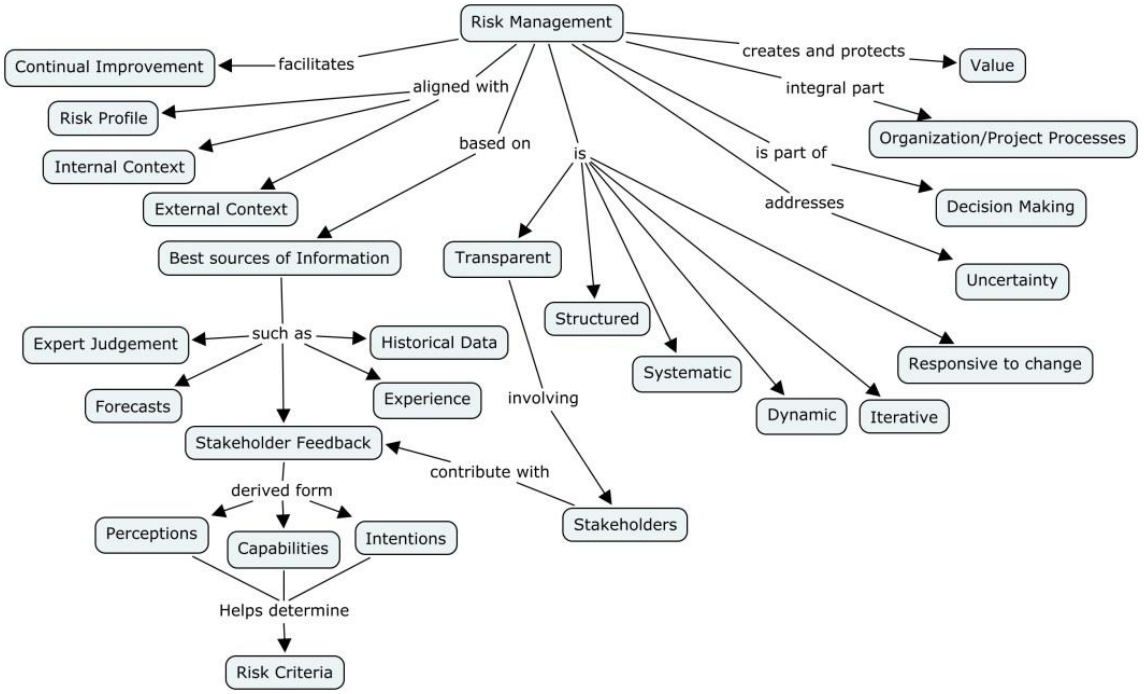

Figure 2.3: Conceptual map with the main RM principles

2.3.3. Risk Management – according to the ISO 31000

ISO 31000 represents a main standard in RM, defining generic RM framework and RM process, being these presented in the next sub-sections.

The Risk Management Framework proposed by the ISO 31000

The success of a RMP depends on the effectiveness of the RM framework used. To facilitate the creation of an effective framework it´s been created a standard method (ISO FDIS 31000, 2009), that can be adapted to each organization's or project's needs and situations. This framework ensures that all information gathered about risk, derived from the RM process is adequately reported and used as a basis for decision making and accountability, thus helping integrate RM into the overall management system.

This framework is composed by several necessary phases that relate in an iterative way that is represented in Figure 2.4.

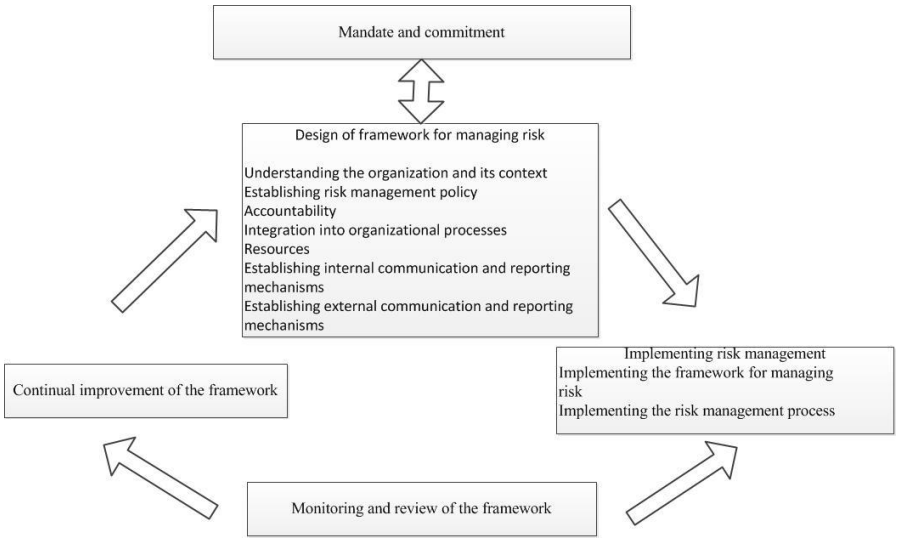

Figure 2.4: ISO 31000 RM Framework

To ensure the on-going effectiveness of RM, a strong commitment by the management must exist, obtained through a strategic and rigorous plan. Management must define and endorse the RM policy, ensure that the culture and RM policy are aligned, determine RM performance indicators, align RM objectives with the general objectives and strategies and ensure legal and regulatory compliance. Accountabilities and responsibilities must be assigned, the necessary resources must be allocated to RM, the benefits of RM must be communicated to all stakeholders, and the framework used must remain appropriate. RMP must be created in order to define the strategy that must be applied according to the goals and risk criteria.

This framework intends to assist in integrating RM into the overall management system. Therefore, there's the need to adapt all the components of this framework to the specific cases and needs. To achieve this, it's necessary to undergo a design process. First it's necessary to understand the context, where an evaluation must be made to understand both the external and internal contexts of the organization or project. To evaluate the external context it's important to analyse the social and cultural, legal, regulatory, financial, technological, economic, natural and competitive environment, whether international, national, regional or local; key drivers and trends having impact on the objectives; and relationships with, and perceptions and values of, external stakeholders. To evaluate the internal context it's important to analyse governance, structure, roles and accountabilities, policies, objectives, the strategies that are in place, capabilities, information systems, information flows, decision making processes (formal and informal), relationships with, and perceptions of, internal stakeholders. The culture present; standards, guidelines, models adopted, and the form and extent of contractual relationships must also be analysed.

After this, a risk management policy must be established stating the objectives for RM and it needs to be communicated appropriately. This policy addresses the several factors, namely the organization's or projects rationale for managing risk. Manages links between the general objectives and policies and the RM policy and determines responsibilities for managing risks. Indicates the commitment needed to

make the necessary resources available to RM, how to measure the RM performance and praises the commitment needed to review and improve the RM policy and method in response to changes.

Accountabilities (clear understanding of who are the individuals with the competence, accountability and authority to manage each risk) are also necessary. To achieve this, first it's necessary to identify the risk owners that have the accountability to manage risks, who's accountable for the development, implementation and maintenance of the framework for managing risk and other responsibilities of people for the RM process. It is also necessary to establishing performance measurement ensuring appropriate levels of recognition.

An integration into the organization/project processes is needed, since RM cannot be a stand-alone activity. It must be embedded and become part of all practices and processes being relevant, effective and efficient. There should be a RMP to ensure that the RM policy is indeed implemented.

The appropriate resources for RM should be allocated, namely people, skills, experience and competence. The definition of the risk processes must also be addressed, as well as the methods and tools to be used for managing risk. All the processes and procedures must be documented. A complete understanding and knowledge of the information and knowledge management systems must exist.

Internal communication and reporting mechanisms must be established. These mechanisms must ensure that key components of the RM framework are communicated appropriately there must exist internal reporting on the framework. Relevant information derived from the application of RM must be available at appropriate times and there must exist processes for consultation with internal stakeholders. External stakeholders must be engaged ensuring an effective exchange of information. External reporting has to comply with legal, regulatory, and governance requirements and feedback on communication and consultation must be frequently provided. Communication must be used as a tool to build confidence.

In order to implement correctly RM, it is necessary to correctly implement a RM framework and process. The appropriate timing and strategy for implementing the RM framework must be defined. The RM policy and process must be applied. Decision making, including the development and setting of objectives, must be aligned with the outcomes of RM processes. Training sessions are useful. The implementation of RM must ensure that the RM process is successfully applied through a RMP at all relevant levels, functions, practices and processes.

RM framework must be continually monitored and reviewed, so that RM is always effective and continues to support performance. RM performance must be measured against indicators. Periodically measure progress against, and deviation from, the RMP. Periodically review whether the RM framework, policy and plan are still appropriate. Report on risk progress and how well the RM policy is being followed. Based on results of monitoring and reviews, decisions must be made on how the RM framework, policy and plan can be improved. These decisions should lead to improvements in the management of risk and its RM culture.

Risk Management Process proposed by the ISO 31000

The totality of objectives and criteria which will be used to measure risks, must be defined, and risks must be detected to, later on, asses, treat or control them.

To achieve this, it was created a standard RM process by (ISO FDIS 31000, 2009), (Figure 2.5) that helps in what regards RM, although this process must be customized to fit the needs and demands. This process is composed by several steps.

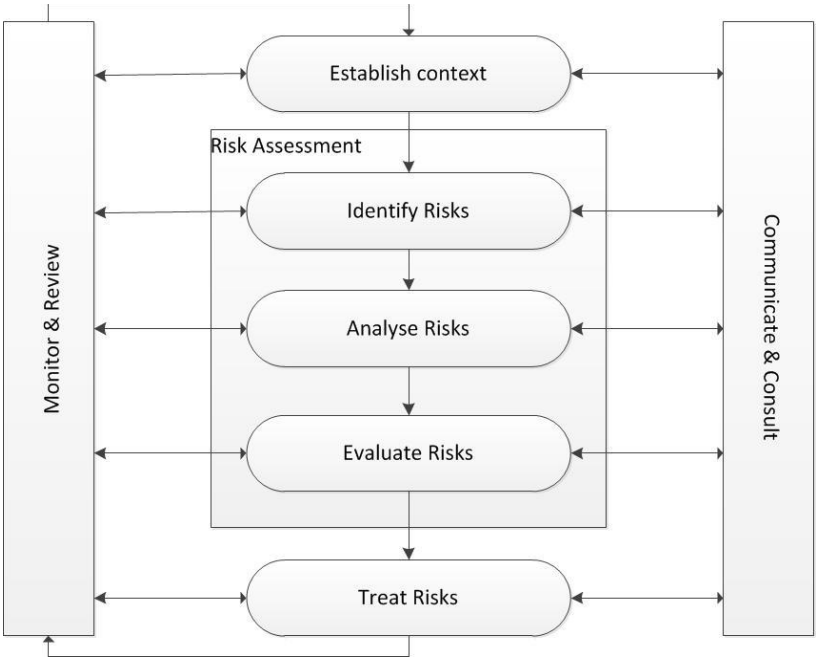

Figure 2.5: The RM process proposed by the ISO 31000

First a context must be established, managing to articulate its strategic objectives and define the external and internal parameters to be taken into account when managing risk, setting the scope and risk criteria for the remaining process, determining which consequences are acceptable in a specific context (Barateiro, Antunes, Freitas, & Borbinha, 2010). This step includes the establishing of an external context which refers to the external environment in which one seeks to achieve its objectives and the establishing of the internal context, which refers to the internal environment, being composed by anything that can influence the way in which the risk will be managed. Risk criteria should also be defined, since it's used to evaluate the significance of risk. The criteria must be based in values, objectives and resources. Risk criteria must be consistent with the RM policy. When defining risk criteria, it must be considered the nature and types of causes and consequences that can occur and how they will be measured.

The next step within the RM process is risk assessment, which can be quantitative, semi-quantitative or qualitative. Risk assessment´s output will define, not only which are the risks that exist within an organization or project, but also, the ones that should be addressed, and so, in this step, risks are identified, analysed, and evaluated. In an assessment of risk, there must be taken into consideration the strategic plans and objectives, as well as, the longer-term perspective on market trends, customer needs, competitors etc. What seems important today may not be in five years. Similarly, although

some longer-range risks may not seem important today, these risks could threaten the organization's or project's survival if left unmanaged (Institute of Management Accountants (IMA), 2007), so in risk assessment the time factor is very important and cannot be neglected. Risk assessment is a hard task, because, it is subjective, depending on people's aspirations, fears, needs, judgments and cultures, and so, there must exist a balance between social, political and scientific factors, representing a multidimensional problem (Slovic, 2001). Social sciences should also be considered and used to improve quantitative estimates of risk probabilities and consequences, leading to an integration of human behaviours into the thinking of technological systems (Jankowski, 2007).

Due to information being assumed as a link between both the natural science paradigm and the social science paradigm, evaluation techniques of both paradigms must be incorporated when attempting to manage risk to information (NIST, 2012). There are several techniques that can be applied in risk assessment (see Table 2.4 – made according to (ISO IEC 31010, 2009)) to identify or analyse risks. The descriptions of these techniques are presented in Table B.1 in Appendix B.

		Applicability		
Technique Category	**Technique**	Identification	Analysis	Evaluation
Lock-up methods	Check-lists	xx	-	-
	PHA	xx	-	-
Supporting methods	Structured Interview brainstorming	xx	-	-
	Delphi Technique	xx	-	-
	SWIFT	xx	xx	Xx
	HRA	xx	xx	X
Scenario analysis	Root cause analysis	-	xx	Xx
	Scenario analysis	xx	x	X
	Toxicological risk assessment	xx	xx	Xx
	Fault tree analysis	x	x	X
	Event tree analysis	x	x	-
	Cause/consequence analysis	x	xx	X
	Cause-and-effect analysis	xx	x	-
	Decision tree analysis	-	xx	X
	CBA	x	x	X
Function analysis	FMEA/FMECA	xx	xx	Xx
	Reliability centered maintenance	xx	xx	Xx
	HAZOP	xx	x	X
	HACCP	xx	x	Xx
Controls Assessment	LOPA	x	x	-
	Bow tie analysis	-	xx	X
Statistical methods	Markov analysis	xx	x	-
	Bayesian analysis	-	x	Xx
	FN curves	x	xx	Xx
	Risk indices	x	xx	Xx
	Consequence/likelihood Matrix	xx	xx	X

Table 2.4: Risk Assessment techniques; (xx) – Strongly applicable, (x) – Applicable, (-) – Not applicable

Risk identification, being the first step of risk assessment, is the process of finding, recognizing and recording risks. Its purpose is to identify sources and causes of risk, areas of impacts and events that might affect the achievement of the objectives as well as the respective causes and potential consequences. It is important to identify all the risks because a risk that is not identified at this stage will not be included in further analysis. Risks should be included whether or not their source is under the control of the organization or project, even though the risk source or cause may not be evident, so once a risk is identified, controls should be designated. The aim of this step is to generate a list of risks (threats - events that affect normal behaviour or vulnerabilities - weaknesses (potential points of failure) in the environment (Barateiro, Antunes, Freitas, & Borbinha, 2010)).

The next step is risk analysis, which provides an input to risk evaluation and to decisions on whether risks need to be treated or controlled and on the most appropriate risk treatment strategies and methods. Risk analysis considers the causes and sources of risks, their positive and negative consequences and the likelihoods of the respective events or threats that can trigger a risk. Existing controls must also be taken into account.

The level of the risk or severity must be determined, resulting in the combination of the likelihood and consequence values. The level of a risk should reflect the type of risk, the information available and the purpose for which the risk assessment output is to be used. Risk levels must be consistent with the risk criteria. It is important to consider the interdependence of different risks and their sources. A continuous and efficient communication with decision makers and stakeholders is needed. Risk analysis can be qualitative, semi-quantitative or quantitative, or a combination of these.

Finally we proceed to risk evaluation, which helps decision making, based on the outcomes of risk analysis, where, depending on the risk level, it's decided, not only if a certain risk is tolerable or if it should be treated, but also the priority for treatment implementation, whether an activity should be undertaken, and which of a number of paths should be followed. The decision about whether and how to treat or control a risk may depend on the costs and benefits of taking the risk and the costs and benefits of implementing improved controls. Risk evaluation can be quantitative, semi-quantitative or qualitative.

A common approach is to divide risks according to their level (risk level) and priority (which is represented through a risk matrix expressed in Figure 2.6):

- **Level V – Risks negligible (grey)**: risks with a very small priority, that are negligible or so small that no treatment measures are needed.
- **Level IV - Low level of risk (Green)**: risks with a low risk level and priority.
- **Level III - Medium level of risk (Blue)**: risks with medium priority and risk level meaning that costs and benefits are taken into account and opportunities balanced against potential consequences, representing risks.
- **Level II - High level of risk (Yellow)**: risks with a high risk level meaning that risk treatment is essential whatever its cost and the benefits obtained representing risks with high priority.

- **Level I – Unacceptable level of risk (Red)**: risks with an intolerable risk level meaning that risk treatment is imperative but also very difficult representing risks with the highest priority.

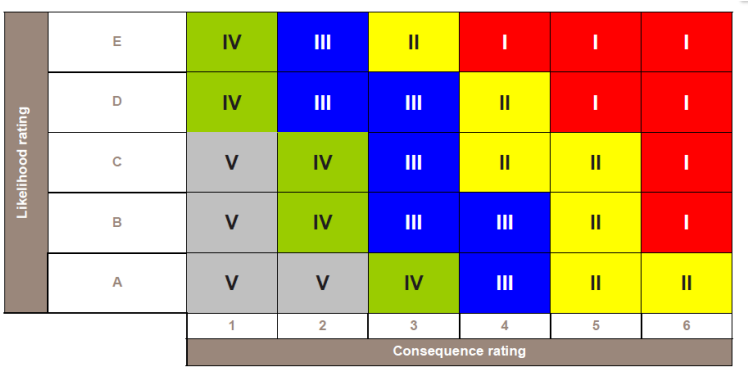

Figure 2.6: Risk Matrix according to ISO 31010

After risk assessment, it comes the risk treatment phase. In this phase it was already decided which risks must be treated and the ones who only need control. Risk treatment involves selecting one or more options for modifying risks, and implementing those options. Once implemented, treatments provide or modify the controls which can introduce risks, namely the failure or ineffectiveness of the risk treatment measures, in which case, there is the need to change the options of treatment applied to a certain risk. Because of this, risk treatment is a cyclic process, where monitoring is very important.

In order to select the most appropriate risk treatment option, there is the need to balance the costs and efforts of implementation against the benefits derived. Treatment options can be considered and applied either individually or in combination, in which case the treatment plan should clearly identify the priority order in which individual risk treatments should be implemented. When selecting risk treatment options, stakeholders values and perceptions must be considered. Though equally effective, some risk treatments can be more acceptable to some stakeholders than to others.

The risk remaining after the implementation of new or enhanced controls is the residual risk. Practically no IT system is risk free, and not all implemented controls can eliminate the risk they are intended to address or reduce the risk level to zero. Residual risks need to be classified as either acceptable or unacceptable (NIST, 2012). If the residual risk has not been reduced to an acceptable level, the RM cycle must be repeated to identify a way of lowering it.

Communication and consultation with external and internal stakeholders should take place during all stages of the RM process. Effective external and internal communication and consultation ensures that those accountable for implementing the RM process and stakeholders understand the basis on which decisions are made. Communication and consultation with stakeholders is important as they make judgments about risk based on their perceptions. These perceptions vary and as their views have a significant impact on the decisions made, the stakeholders' perceptions should be identified, recorded, and taken into account in decision making.

Both monitoring and review (also called control), has to be planned and defined in an early stage of the RM process, and it must become part of this process, where regular checking or surveillance must

be applied periodically or ad hoc. Though this activity, is ensured an effective and efficient control by improving RM, in the detection of changes in the contexts or criteria used, leading to possible revisions of risk treatments and priorities, as well as identifying new risks. All the results generated by this activity should be recorded.

2.4. Risk Management Plan

A RMP specifies the approach, the management components and resources to be applied to the management of risk in a particular product, process, project, part or whole of an organization. Management components typically include procedures, practices, assignment of responsibilities, sequence and timing of activities (ISO Guide 73, 2009).

The RMP is considered a living document which must be updated as needed to ensure that previously identified risks are managed effectively and new risks are quickly identified and managed. RMP depends on the RM framework used. A typical RMP comprises the following steps that were explained in the previous sections:

- Establishment of context.
- Risk Assessment composed by.
    - Risk Identification.
    - Risk Analysis.
    - Risk Evaluation.
- Risk Treatment.
- Risk Control and Monitor.

An overview of RM and RMP can be presented in Figure 2.7.

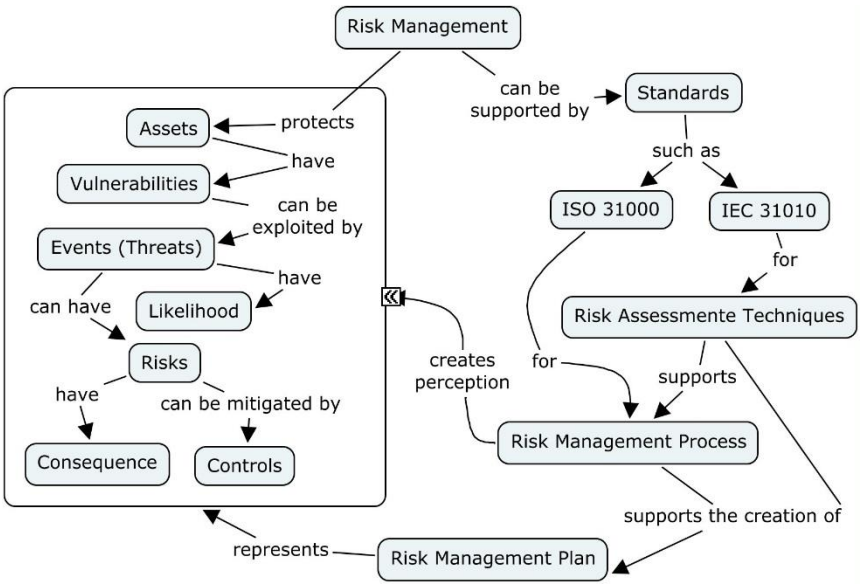

Figure 2.7: RM and RMP overview

2.5. Potential relations between Risk Management and Data Management Plan

As it was stated before, we defend that DMP can be improved, becoming a more complete concept, leading to a better DM and better decision making. To improve DMP we suggest the usage of RM principles and guidelines, though the creation of a new document called RMP. Considering the RM related work already presented, we believe the RMP should be created using ISO 31000 guidelines as a point of reference, since this standard is generic. As a result, we believe DMP can be supported and justified by the RMP, where the results of RMP can be associated with the typical sections of DMP. With this proposal, a better DM for engineering and science projects can be achieved.

To understand why we believe RM principles, and in practice a RMP, are useful and appropriate to complement DMP, we must analyze the objectives, requirements and current guidelines for DMP, so it becomes clear what usual concerns these documents comprise and their relation to RM and RMP.

DMP tries to help understand how researches can share their data by detecting and removing what might limit or prohibit the dissemination of that data. This objective is shared by RM and can be achieved through its application.

DMP assigns roles and responsibilities across partners. This can be supported by risk owner and stakeholder definition, not only for communication purposes but also to consider each of the stakeholder's risk perception in decision making.

DMP is a living document just like a RMP implying also the monitoring phase of ISO 31000 risk process.

For the DMP sections presented Table 2.2, it has been proved the concerns they address raise risks. Despite this fact, in any of these guidelines and sections, RM principles or techniques aren't used to help address the respective issues. This fact lead us to believe DMP can be improved by RM, through the consideration of its guidelines and techniques by helping detect risks and controls that can be associated with each section's concerns, being these risks and controls otherwise ignored, since the correct "tools" that help finding them are not used.

To conclude, we think RM and RMP can complement DMP because both areas and documents share objectives, namely the protection of data, seeing data as an asset, also DM activities can be seen as risk control measures for risks associated with data and together, these activities/controls can form risk policies. Finally DMP sections have related risks that should be addressed by RM good practices and techniques.

3. Solution Hypothesis and Proposal

This chapter reports to the definition of the objectives for a solution and the design and development stages of the research method. In order to solve the stated problem, a structured method was developed for RMP creation, used for engineering and science projects with DM concerns. Besides the method, first we present a RM conceptual model used for our method, being this model a synthesis of the most relevant RM concepts presented in section 2.3. Some skills (required by the risk expert) and generic responsibilities/roles are also presented. A RM tool used for support to the presented method is presented. A risk registry is also suggested, being useful for the creation of check-lists used in the beginning of the risk analysis. All these elements combined represent our proposal for complementing DMP and promoting a solid DG in engineering and science projects.

3.1. Solution Hypothesis

To solve the problem presented in section 1.2, and considering the related work presented in the areas of interest, we try to explore the hypothesis of improving DM concerns in engineering and science projects by improving the DMP concept with RM good practices and guidelines, which in practice means using a RMP and a DMP to promote DG. This translates into several more concrete hypotheses:

- **Promote a DG approach for engineering and science projects**: Since DM and RM, as well as DMP and RMP, address DG concerns and DMP don't cover all the former, we propose a new approach to DG through the joint utilization of methods and techniques belonging to both areas, namely the usage of both DMP and RMP.
- **Identify the typical sections of a DMP relevant for engineering and science projects**: Since our work is based on the assumption that the DMP already exists, with no DMP being actually developed during this dissertation, it becomes necessary to generalize DMP requirements in form of generic sections.
- **Create a method for engineering and science projects that allows the creation of RMP**: This method represents the artifact produced in this dissertation and it should be guided by ISO 31000 guidelines and ISO 31010 risk assessment techniques. As a final output it should produce a RMP.
- **The RMP created should support and justify the decisions, processes and controls implemented by the DMP**: The RMP, besides comprising all the processes, results and techniques used or obtained in a RM analysis, it should also support or justify the DMP. To achieve this, it should complement the DMP, which means the results of the RMP should be associated with the typical DMP sections defined in the second hypothesis.

3.2. Risk Management Conceptual Model

As stated before, risk standards, references and terminologies vary with the market sector. Due to the vast scope of engineering and science we needed to ensure that the concepts were sufficiently

generic so that they apply to any engineering and science projects. At the same time we wanted to ease understandability by using the minimal set of concepts relevant to our goal. The set of risk management concepts proposed to apply our process (Barateiro, 2012) are illustrated in Figure 3.1, and are based on ISO 31000.

Based on those concepts, risk is an effect of uncertainty and is expressed by the combination of the likelihood of an event and its consequences when exploiting a vulnerability of an asset (Barateiro, 2012). Asset is defined has something (e.g. process, data, hardware, software, people) that has value to the project. A risk is expressed by a risk severity (or risk level) that is a combination of its consequence with the likelihood of the event triggering the risk. Finally controls are defined as actions that can be taken to mitigate risks. Controls can reduce the exposure of a vulnerability, reduce the likelihood of an event, reduce the risk consequence, transfer the risk and accept the risk. A risk policy represents a set of controls that were applied to mitigate the risks in a specific context.

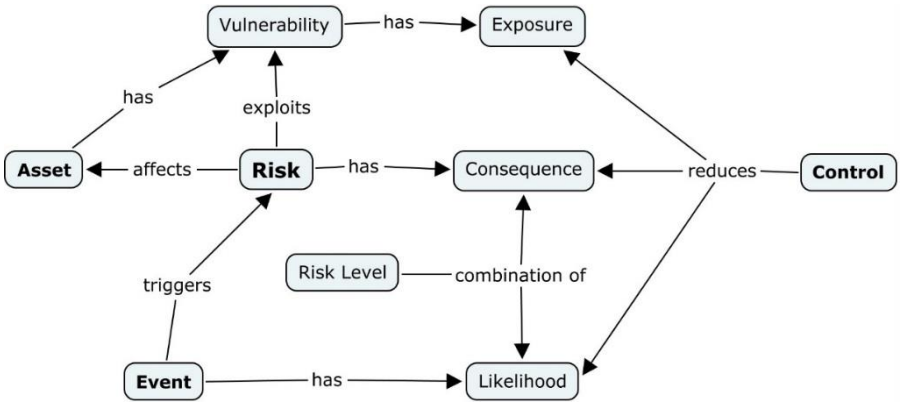

Figure 3.1: RM conceptual model overview

3.3. Risk Management Method

The proposed method is illustrated in Figure 3.2 and intends to produce a RMP as final output. The structure of this method represents the structure of the final RMP created. Two examples of this structure and application of this method are given in sections 4.1.2 and 4.2.1.The process is based on ISO 31000 being generic enough and suited for engineering and science projects. The main differences are:

- **The assessment process is simplified**: Due to the necessary rigor in science and engineering, projects are usually based on well-defined workflows or business processes. Therefore, to ease the assessment process we recommend that stakeholders analyze individually every task of the project. Using this approach, it is also possible to validate the completeness of the assessment since we assure that all steps were covered.
- **The results of the risk assessment are associated with the requirements of DMP**: As stated above, different organizations have different requirements for DMP. Those requirements are typically translated to a set of sections that need to be provided. By associating the results of the risk assessment with those sections we ensure traceability

between the DMP and the results of the risk assessment easing the definition of the document. With this association the DMP is complemented by the RMP that justifies and supports the problems/risks defined in the DMP.

- **Risk treatment supported by the association of risk assessment results with DMP requirements**, being the average risk level/severity of each section calculated and used to determine the more priority risk treatment measures.
- **The results of the risk treatment are also associated with the requirements of DMP**: By associating the results of the risk treatment with DMP sections we ensure traceability between the DMP and the results of the risk treatment justifying treatment or control measures defined in the DMP.

The method, and consequently the RMP starts by establishing the context, i.e., defining the internal and external context of the project. External context may include a description of the regulatory environment of the project or any other element that might affect DM. Internal context involves defining all the elements of the project, i.e. its objectives, resources, data, processes, systems, among others that may be relevant to consider. We recommend that DMP contends are used to help define the scope of the project, due to their descriptive and less technical nature (It is assumed that the DMP is already created, being is creation out of the scope of this dissertation). The inputs of this phase are the documents or expert opinions that help define the context. The output is a well-defined and structured scope of analysis.

In the RMP, after the context is determined, the project or system analyzed should be divided into tasks or smaller components. This is done, to ease the analysis process by diminishing the scope of risk assessment. Therefore, risk assessment should be performed individually for each task of the project. Using the input from the previous phase, the systems or processes are then divided, being these smaller components the output of this phase.

The next content on the RMP is risk assessment, which is composed by three different steps: (1) risk identification where we identify all relevant assets, vulnerabilities, events and risks; (2) risk analysis where we estimate the asset value, the vulnerability exposure, the event likelihood, the risk consequence and ultimately the risk severity; and (3) risk evaluation, where we evaluate the information produced in the previous steps to check against risk criteria, deciding whether a specific risk is acceptable or tolerable. If the assessed information is insufficient to decide whether the risk is acceptable then the risk assessment is considered unsatisfactory and should be repeated. Using as input the outputs of both previous stages, the final output of risk assessment is constituted by lists of assets, vulnerabilities, events and risks, being the former ranked according to their severity. In this method, no criteria is suggested, being that decision made with the project experts. Risk assessment is performed using a set of techniques presented in section 3.4.

In the next step of the RMP, risks are associated with the DMP requirements to identify the level of risks for each DMP section. The output of this stage is the correspondence between the risks and the DMP sections, being the average risk severity of each section calculated.

The result of these steps eases the prioritization for risk treatment (next step of RMP) where controls are identified and associated to DMP requirements.

If the risk treatment results are satisfactory, i.e. the controls are sufficient to lower the overall risk level to an acceptable value, then a series of conclusions are drawn up, using the results or outputs of all risk assessment and risk treatment phases. These conclusions will support risk report to all stakeholders and represent the final component of the RMP.

Finally, just as recommended by ISO 31000, all steps of the method should be communicated through the stakeholders of the project for consultation and validation (which is supported by the output of risk report phase, namely a set of conclusions). Also, the method should be regarded as a continuous process where results from different steps are constantly being monitored and reviewed if necessary.

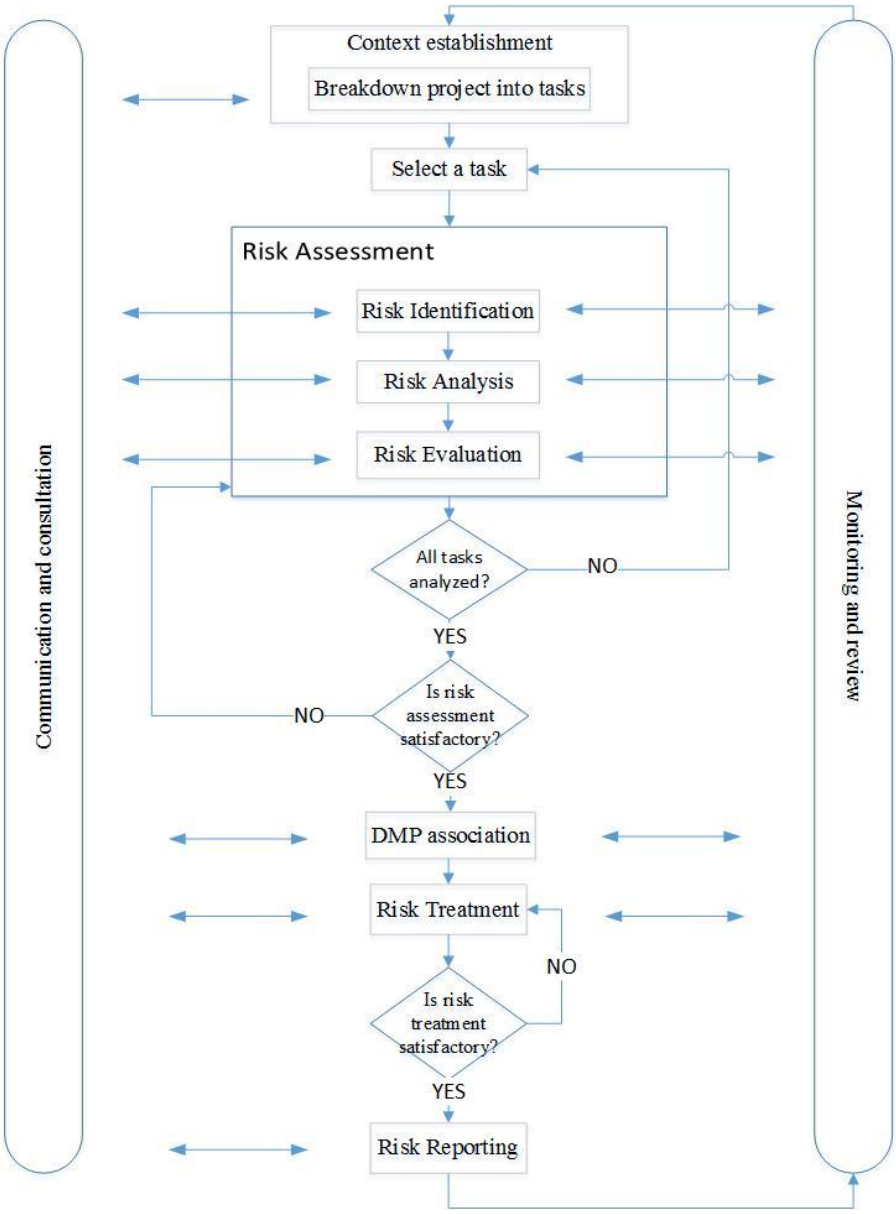

Figure 3.2: Method for RMP creation based on ISO 31000 guidelines

3.4. Risk Management Techniques

As described above, ISO 31010 (ISO IEC 31010, 2009) identifies a set of risk management techniques that can be applied to the steps of risk identification, analysis, evaluation and treatment. For each technique, the standard defines the step(s) of the process for where the technique is more suitable. However, as stated in the standard, techniques should be selected according to several factors such as: (1) the objective of the assessment, (2) the needs of the stakeholders, (3) the type of risks being analyzed, (4) the potential magnitude of the consequences (e.g. the initial perception of consequence may indicate how thorough the risk assessment technique needs to be), (5) the skills of the stakeholders, (6) the availability of resources (e.g. some techniques may require information that is not available), (7) the need for future changes (e.g. some techniques are more suitable for change than others), or (8) any regulatory and contractual requirement (e.g. an obligation to use specific techniques) (ISO IEC 31010, 2009).

In (Canteiro, 2011), the risk techniques identified in ISO 31010 were analyzed regarding its suitability for preservation of e-Science data and processes. The authors analyzed each technique in terms of applicability, feasibility, and understandability for the context in hand. Using those results as a starting point and considering our own requirements we recommend the following techniques:

- **Check-lists**: Since for specific domains it is possible to identify common threats and vulnerabilities a check-list is recommended as a simple method to be used as a starting point for risk identification. A list of typical vulnerabilities and events for science and engineering projects can be found at (Barateiro, Antunes, Freitas, & Borbinha, 2010).
- **Brainstorming**: Due to vast range of science and engineering, projects may include groups of stakeholders with different skills and knowledge. Assembling those stakeholders for brainstorming promotes imaginative thinking which favors the use of this technique for risk identification. Also, through the use of brainstorming we assure that identified risks are fully comprehended and agreed by stakeholders.
- **Hazard and operability studies (HAZOP) technique**: This technique is based on the use of guidewords which question how a certain task or system might fail, being considered unwanted outcomes and deviations from the intended outcomes, as well as, possible causes and failures modes. Since several data management threats occur due to failures, and can produce unwanted outcomes, this technique is recommended for risk identification, analysis and evaluation.
- **Structured "what-if" (SWIFT) technique**: This is a systematic technique that utilizes a set of "prompt" words or phrases to stimulate the assessment and treatment of risks. The technique proposes a workshop between stakeholders where "what-if" type phrases are used to stimulate the exploration of potential scenarios, their likelihood and the consequences. Participants should also be encouraged to rank and prioritize risks and identify possible controls. Therefore the technique is recommended for all steps of risk assessment and risk

treatment. The technique is considered a similar high-level approach to HAZOP with the benefits of the brainstorming technique.

- **Consequence/probability matrix**: As the name suggests this technique is a matrix that combines qualitative and semi-quantitative values of risk consequence and event likelihood to produce a level of risk/severity. The technique is mostly used to rank and prioritize risks making it a suitable technique for risk evaluation.

It is also important to note that the techniques selected were also chosen taken into consideration their ability to individually analyze tasks or smaller components as suggested in our method.

3.5. Skills and Responsibilities

Despite the diversity in engineering and science projects it is possible to identify generic roles that are typically present in any project. Additionally risk management good practices also recommend specific roles that should be present in any risk assessment. The roles that should be present when applying the method and their responsibilities within it are presented in Table 3.1 being some based on (Hillson & Simon, 2007) and others resulted from the interactions of the interviewers from both case analysis (see section 4):

Role	Description	Responsibility
Project Sponsor	Typically assumed by the project's funding agency	Should be informed of all existing risks and controls
Group Leader	Normally represented by the project PI (principal investigator). This role is more relevant for scientific projects	This stakeholder is responsible for risk communication to all relevant stakeholders. It also has a major role in decision making, together with the project sponsor
Project Manager	Typically assumed by the researcher that is responsible for coordinating the project	Responsible for defining the project context and for risk communication to all relevant stakeholders. Accountable for all decision making
Risk Expert	Person with knowledge of principles, processes and techniques to identify, analyse, evaluate and treat any risk	Should be consulted in all steps of the risk management process.
Risk Owner	Person in charge of controlling and monitoring a specific risk	Should be informed of all decisions regarding that specific risk. It is responsible to communicate any issue to the project manager
Operational/Scientific Team	Persons in charge of executing the project	Responsible for the implementation of risks controls
DM expert	Person in charge of executing the project's DMP	Responsible for creating the DMP and assisting in the association of its sections with RM results

Table 3.1: Roles and responsibilities in engineering and science projects relevant for the purpose of this work

Considering the former roles and responsibilities present in Table 3.1, we propose a generic RACI matrix (see Table 3.2), where these roles are and responsibilities are related with a series of tasks that are performed in the execution of the proposed method (Ferreira, et al., 2013a), (Hillson & Simon, 2007). This matrix was reviewed and supported by the interviewees of both cases.

Method steps	Tasks	Project Sponsor	Principal Investigator (PI)	Project Manager	Operational/ Scientific Staff	Risk Expert	Risk Owner	DM expert
Context establishment	Insurance of RM resources	R	C	A/R		C		
	Establish boundaries	I	C	A/R	C	I		
	Define risk ranges		I	A		R	C	
Risk assessment	Identify assets, vulnerabilities, events and risks	I	I	A	C	R	I	I
	Analyse risks	I	I	A		R	I	
	Evaluate risks	I	I	A		R	I	
DMP association	Determine all relevant sections			A		R		R
	Associate RM results to DMP sections			A		R	I	C
Risk treatment	Controls creation			I	R	C	A	C
	Control acceptance	I	I	A/R	I	C	I	C
Risk report	RM conclusions conception	I	I	A		R	I	
	RMP acceptance	I	I	A/R		C	C	I
Monitoring and review	Controls monitoring			I	R	C	A	
	Control enforcing			I	R	C	A	
Communication and consultation	Sharing risk results and conclusions	I		A	C	R	C	I
Tasks transversal to every method step	Decision making	I	R	A/R				
	RMP creation			A		R	C	I

Table 3.2: Proposed RACI chart (R=Responsibility, A=Accountable, C=Consulted, I=Informed))

As with data management, librarians and archivists can play an important role in the method due to their skills in data creation, preservation, and access. Although the role also demands risk management skills, this raises an opportunity for learning new skills and assume themselves, in the future, as risks experts. In order to assume the role of risk experts, the following set of skills is proposed:

- **Data Management**: know DM principles, techniques, standards and the project's data life cycle.
- **Security**: A good background on breaches that threatens data is fundamental to assess risks and mitigate them.
- **Metadata**: Knowing how to produce, collect, manage and secure metadata.
- **Advocacy, copyright and intellectual property rights**: Data dissemination is important, so copyright or property infringement brings risks threatening that goal.

- **Technical skills**: Relevant to determine technical risks and controls related to the technology and infrastructures in use.
- **Data value**: Know how to assess the value of the data objects worth protecting.
- **RM skills**: Knowledge of principles, processes and techniques to identify, analyse, evaluate and treat any risk surrounding data.
- **Engineering or science field focus**: Knowledge of the field in question is mandatory.

3.6. HoliRisk: a Risk Management Tool

To register the risk information produced by a risk management process organizations typically use desktop applications (e.g. spreadsheets), or they build custom tools to fill their needs. While the latter option is normally unavailable for small projects (due to lack of resources), the former option presents some disadvantages especially regarding sharing and reuse of risk information. Also tools lack the capacity of managing risks using a consistent and holistic view of the organization or project (Barateiro, 2012).

HoliRisk[3] is a web-tool developed by INESC-ID in the context of the European project TIMBUS. The tool is based on ISO 31000 guidelines and it has two main goals: (1) to support risk assessment by storing risk related data or risk registries in a structured way; (2) to support risk reporting by providing graphical aids and risk information representation mechanisms to support decision making, communication and consultation by presenting risk information using a consistent and holistic view through personalized risk reporters (see Figure 3.3).

Figure 3.3: HoliRisk risk reporter dashboard

[3] The tool is publicly available at http://bd1.inesc-id.pt/riskReporter/

3.7. Risk Registry

A knowledge base, useful for the beginning of any RM analysis is presented. This is composed by some known generic vulnerabilities, events, risks and controls. These elements were obtained by studying the related work (Barateiro, Antunes, Freitas, & Borbinha, 2010), (ISO IEC 27005, 2011), (ISO IEC 27001, 2013) and by experience obtained by the analysis of both cases (Ferreira, et al., 2013b), (Redlich, et al., 2012). The assets are presented in Figure 3.4. Vulnerabilities are presented in Figure 3.5. The events are presented in Figure 3.6. The generic risks are showed in Figure 3.7. Finally the generic controls are presented in Figure 3.8.

Name	Value
Data	very-high
Documentation (metadata)	very-high
Hardware	medium
Personal/staff	medium
Software	medium

Figure 3.4: Generic assets

Name	Exposure
Change of laws	high
Confidential data belonging to human or protected species	medium
Data fragmentation	high
Data standards and formats need to be updated	high
Development teams are composed mainly by scientists and biologists, lacking personal with DM skills	medium
Economic/organizational breakdowns	low
High error rate in data	high
Human dependency	high
Lack of a standard for metadata/documentation representation	high
Lack of data preservation policy	low
Lack of suited descriptors for each technique used in a task	high
Security breaches	
Too large data sets in size or quantity	very-high
Unreliable hardware	medium
Unreliable software	medium
Vital knowledge remains in certain personal	medium
Workflow/tasks need to be preserved	very-high

Figure 3.5: Generic vulnerabilities

Name	Likelihood
Abandonment of a stakeholder	medium
Creation and utilization of new technologies and techniques that increase substantially the quantity of data and metadata generated	high
Data used inappropriately	very-high
Employee with vital knowledge leaves	medium
Errors on search, access and delivery of preserved data	high
Financial, legislative or organizational changes	very-high
Hacker attack	medium
Human errors	high
Infrastruture failiure	high
Infrastruture maintenance	high
Law changes overlooked	medium
Natural disasters	very-low
Non successful extrapolation of the data, metadata and documentation's meaning	high
Preservation community demands new requirements that cannot be met by the preservation solution	high
Software failiure	medium
The research team uses data sets or techniques created by other teams in other experiments and claim their ownership	medium
Tool discontinuation or lack of support	medium
Unauthorized individuals, entities or processes access and disclose classified information	medium
Violation of data protection regulations	high

Figure 3.6: Generic events

Name	Consequence
Accidental deletion of data	very-high
Accidental or deliberate system failure	very-high
Computational servers alteration	high
Copyright infringement	high
Data base alteration	very-high
Difficulties sharing the information and the workflow's execution details in future scenarios	very-high
Hardware obsolescence and faults	very-high
Inapt, incompreensible or incomplete data, metadata and documentation	very-high
Insertion of wrong input values, influencing the results	very-high
Lack of DM and computer science skilled personal in the research teams	high
Lack of financial requirements	very-high
Lacking data for the success of the experiment	very-high
Loss of data traceability	very-high
Loss of expert knowledge	high
Loss of information due to communication failures	very-high
Loss of metadata	very-high
Non-compliance with general legal obligations	high
Preservation Strategy/Plan cannot be Met	very-high
Stakeholders lack of involvement	medium
Tools obsolescence and faults	medium
Unavailability of storge capacity to responde to data growth	very-high
Violation of Intellectual property	high

Figure 3.7: Generic risks

Name	Type
Anti-fire and earthquake measures	Likelihood
Assure all data, metadata, software components or other accomplishments of any experiment are protected by copyleft or copyright	Likelihood
Backup system	Consequence
Create a long term storage policy (recovery management Plan)	Consequence
Creation of data, metadata and documentation standards for scientific community information sharing	Likelihood
Creation of mechanisms for long-term funding	Likelihood
Creation of suited descriptors to each technique and tool used	Exposure
Development of appropriate software to accommodate new accepted techniques in the scientific community	Consequence
Encourage formation and training	Consequence
Get consent from data sources for data usage	Consequence
Have an emergency budget (Sustainability plan)	Consequence
Improve security measures	Likelihood
Include computer science and DM experts in research teams	Likelihood
Keep all the software and hardware components up to date	Likelihood
Usage of fall-back computational servers	Consequence
Usage of fall-back data bases	Consequence
Use open-source tools and recent formats	Consequence
Use several forms of documentation	Likelihood

Figure 3.8: Generic controls

These generic elements, belonging to a generic engineering and science model, were then used as baseline for the RM analysis in both case studies, where some elements were considered or excluded, depending on the requirements and characteristics of the cases analyzed.

4. Demonstration and Evaluation

This chapter corresponds to the demonstration and evaluation stages of DSRM. For demonstrating and evaluating the proposed method, two different cases were used and analyzed, as described in the next sections.

4.1. Case 1 – Metagenomics (MetaGen-FRAME Project)

The first case used, was a project called MetaGen-FRAME from the field of Metagenomics. HoliRisk is used to support the presentation of results.

4.1.1. What is Metagenomics

Metagenomics, being a field of science and belonging to the area of Bioengineering, is a recent discipline that enables the genomic study of uncultured microorganisms (Wooley, Godzik, & Friedberg, 2010). Like genomics, metagenomics is both a set of research techniques, containing several approaches and methods, and a research field that, on one hand, seeks to understand biology focusing on a certain community genes and how genes might influence each other's activities in serving collective functions, and on the other hand also recognizes the need to develop computational methods to maximize the understanding of genetic composition and community activities.

Meatagenomics, as stated earlier, is a Bioinformatics project, comprising two components, a Biology component and an Informatics component. As the main concern of this dissertation resides in achieving effective DM in engineering and science projects, the Biology component of these projects don't belong to the scope of this dissertation, meaning all characteristics typically biological from all the data, tasks, workflows, as well as all the risk results that are generated by the analysis of these projects, are not considered, being the scope concentrated in the Informatics part of these projects.

Metagenomics projects make use of large quantities of data, which need to be effectively gathered, sorted, treated and finally properly stored for future use. An effective DM of all these large data sets becomes a huge challenge for the research teams responsible for developing these projects. This dissertation tries to solve this challenge through the effective combine usage of DMP and RM principles, but first becomes necessary to fully understand the typical scientific workflows (set of typical tasks) that characterize a typical Metagenomics project.

The typical tasks that comprise the general workflow of a Metagenomics project are the following (Wooley, Godzik, & Friedberg, 2010):

- **Sample collection**: The first step of a Metagenomics project consists in retrieving the set of environmental samples to perform all the required experiences. This step also involves the effective storage of the retrieved samples in an appropriate lab for their analysis.

- **Filtering data**: After collecting the samples, there must be a kind of quality control assuring the removal of the duplicate and low quality samples.
- **Sequencing**: When the desired set of samples is filtered, each sample is divided in several fragments, and the involved DNA is isolated and extracted, being a set of sequences obtained.
- **Assembly**: After the sequences belonging to a certain genome are obtained, the reads are then assembled into progressively longer continuous sequences or contigs, and finally to the whole genome.
- **Gene Prediction**: This task intends to predict the presence of gene information on the assembled sequences or contigs obtained in the previous task. In other words, this task intends to measure the diversity of genes in a sample, knowing what species populate that same sample.
- **Functional annotation**: After assembling a metagenome and identified the species that populate the sample, there's the need to understand the functional potential of the microbial community.
- **Metabolic reconstruction**: Produce data results about the metabolic properties and networks observed in the samples.

This set of tasks represents the initial or local Metagenomics data processing pipeline. Since there are several different communities involved in this field, communication and collaboration is mandatory between them. To achieve this, all the results obtained from the local pipeline described earlier, must be integrated with the results obtained by different communities, allowing the advancement of scientific knowledge. This means, maintaining a constant communication with shared large data centers, being this communication essential for the development of some tasks described earlier.

Large quantities of metadata are also generated during the local pipeline being as much important as the data itself. All this metadata must be recorded and stored, as well as disseminated throughout the scientific community, just like the generated data, so that other communities or scientific teams can replicate the experiments performed.

4.1.2. Method Application and RMP Creation – MetaGen-FRAME Project

This section presents the application of the risk management method, and corresponding RMP creation, on a science project, namely the MetaGen-FRAME project (Coimbra, 2013). To facilitate the understandability of the application, the results of the method are presented progressively according to the steps of the proposed method (see section 3.3). These results compose the RMP content, being the structure of this document similar to the succession of steps presented. To present these results, HoliRisk is used, although with certain exceptions since there are a large number of results for some entities, reducing the view quality via HoliRisk due to a data presentation issue present in this tool. In Appendix C a series of proposed improvements to HoliRisk are presented.

Context Establishment

This project concerns itself with the design of an open method with several bioinformatics modules that studies environments composed by multiple types of bacteria. In this project, large data sets are created, stored and accessed, being one of the objectives the results dissemination.

The project's general workflow and all bioinformatics modules are presented in Figure 4.1. This workflow and modules, including their respective inputs, outputs and used tools are detailed in Appendix D.

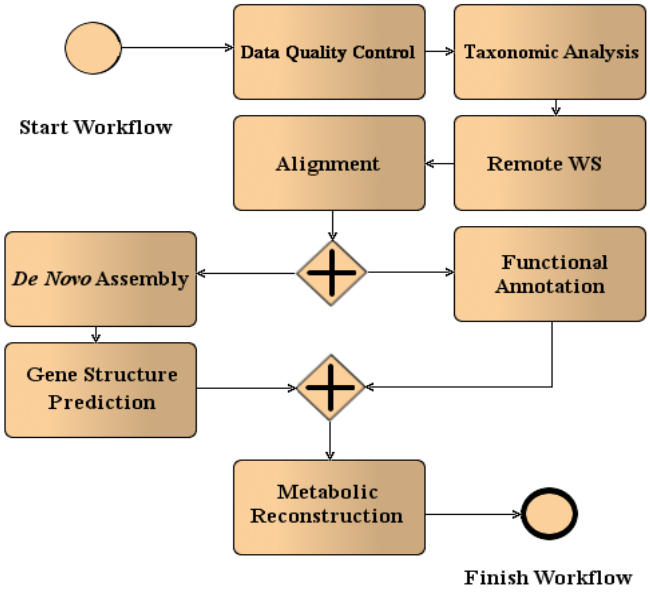

Figure 4.1: MetaGen-FRAME workflow

Beyond the requirement of data reutilization and dissemination with possible licenses related to that, this project also requires a reliable and secure data storage, protection against legal issues related with possible ethical, loss of confidentiality or copyright issues, requires a good documentation, including metadata, guarantying data's traceability, requires a good remote data access and finally requires the correct definition of every data object's owner, as well as the responsibilities of every stakeholder involved.

Using the proposed RACI matrix (see Table 3.2) each role and given responsibility in this scenario was defined together with Miguel Coimbra. As project sponsor, the FCT (Fundação para a Ciência e Tecnologia) was considered. For group leader or principal investigator (PI), professor Ana Teresa Freitas was considered. For project manager, professor Luís Russo was considered. For scientific staff, Miguel Coimbra was considered. I represented the risk expert role. As risk owners, depending on the risk, they could be Miguel Coimbra or elements of the IT division of NCBI (concerning NCBI related risks) or other database consulted. Finally for DM experts, the NCBI or other database archivists are considered.

Risk Assessment

For risk assessment, we opted to use the HAZOP and SWIFT techniques (see section 3.4 for more details). As proposed by the technique, a workshop with Miguel Coimbra was conducted for brainstorming risk assessment activities. As a starting point for the discussion and risk assessment we used the list of typical assets, events, vulnerabilities and risks (see section 3.7). Certain guidewords and what if scenarios characteristic from the HAZOP and SWIFT techniques also helped with risk assessment (see Appendix E). HoliRisk is used for result presentation support. All the results are further detailed at (Ferreira, et al., 2013b).

The project assets are identified in Figure 4.2 as well as the vulnerabilities associated with them. Related to risk identification step, the vulnerabilities found for this project are presented in Figure 4.3. In Table 4.2, the project's events are presented with their respective likelihood. In Table 4.3, we present the MetaGen-FRAME risks, with their respective consequence and severity values, as well as their trigger events and related assets. The criteria used for likelihood and consequence are presented in Table 4.1 and were chosen in collaboration with Miguel Coimbra. Vulnerability exposure criteria was also chosen in collaboration with Miguel Coimbra. For calculation of risk severity, the following formula was used:

$$Severity_{risk} = Consequence_{risk} * Likelihood_{event} \quad (4.1)$$

This equation was defined in collaboration with Miguel Coimbra.

	Likelihood	Level	Consequence	
0.1	Extremely unlikely risk due to usage of very well understood technologies and tools.	Very-low	Very small chance of endangering the workflow. Almost no changes are necessary.	1
0.3	Unlikely risk due to usage of well understood technologies and tools with few problems and deficiencies.	Low	Small chance of endangering the workflow. Very few changes are necessary.	3
0.5	Somewhat likely risk due to usage of technologies with some problems or deficiencies, which take some time and effort to mitigate.	Medium	Can endanger the workflow. Some changes are necessary.	5
0.7	Likely risk due to the presence of several serious problems and deficiencies, which take a considerable time and effort to mitigate.	High	High chance of endangering the workflow. Large changes are necessary.	7
0.9	Extremely likely risk due to the presence of major problems and deficiencies, which take a major time and effort to mitigate.	Very-high	Very high chance of endangering the workflow. Major changes are necessary.	9

Table 4.1: Likelihood and consequence criteria used for MetaGen-FRAME project

For each risk we identify the event that could trigger it. As visible on Table 4.3, risks have relations between them in the sense that a risk could be trigger event of another risk. As an example, loss of data traceability by non-successful extrapolation of metadata's meaning can then lead to a non-successful representation of the output to the user via Taverna. Events can also be seen as risks and vice versa.

#	Name	Value	Vulnerabilities Number
A0	Computational Servers	high	V8 V10 V11
A1	Computers	medium	V8 V10 V11
A2	Data/Metadata	very-high	V8 V0 V2 V4 V5 V6 V7 V9 V10 V11 V12
A3	Databases	very-high	V8 V10 V11 V2
A4	Staff	medium	V1 V2
A5	Tools	high	V8 V11
A6	Web-Service	high	V8 V11
A7	Workflow/Tasks	very-high	V8 V10 V11 V12 V3 V1

Figure 4.2: MetaGen-FRAME assets

#	Name	Exposure
V0	Confidential data belonging to human or protected species	medium
V1	Development teams are composed mainly by scientists and biologists, lacking personal with DM skills	medium
V2	Economic/organizational breakdowns	low
V3	Human dependency	medium
V4	Lack of a standard for metadata/documentation representation	high
V5	Lack of data criteria defining if data is confidential or not	low
V6	Long storage policy lacking	high
V7	Preservation law changes	high
V8	Security breaches	medium
V9	Too large data sets in size or quantity	very-high
V10	Unreliable hardware	medium
V11	Unreliable software	medium
V12	Workflow/tasks inputs and outputs need to be preserved for future use	very-high

Figure 4.3: MetaGen-FRAME vulnerabilities

ID	Events	Likelihood
E0	Abandonment of a stakeholder	0,5
E1	Creation and utilization of new technologies that increase substantially the quantity of data and metadata generated	0,5
E2	Data managed by biologist with no DM experience	0,3
E3	Data used inappropriately	0,7
E4	Errors on search, access and delivery of preserved data	0,5
E5	Financial, legislative or organizational changes	0,3
E6	Hacker attack	0,3
E7	Hardware obsolesce	0,3
E8	Human errors	0,7
E9	Infrastructure failure	0,3
E10	Infrastructure maintenance	0,5
E11	Natural disasters	0,1
E12	Non successful extrapolation of the data, metadata and documentation's meaning	0,5
E13	Preservation Law change overlooked	0,3
E14	Sharing of information without consent	0,3
E15	Software failure	0,5
E16	The research team uses data sets or techniques created by other teams in other experiments and claim their ownership	0,3
E17	Tool discontinuation and lack of support	0,1

Table 4.2: MetaGen-FRAME events

ID	Risks	C	Event	Assets	Severity
R0	Copyright infringement by sharing confidential information	5	E14	A2	1.5
R1	Copyright infringement by claiming ownership of data sets produced by others	5	E16	A2	1.5
R2	Difficulties sharing the information and the workflow's execution details in other future scenarios due non successful extrapolation of data's metadata	7	E12	A2, A7	3.5
R3	Difficulties sharing the information and the workflow's execution details in other future scenarios due to errors on search, access and delivery of preserved data	7	E4	A2, A7	3.5
R4	Difficulties sharing the information and the workflow's execution details in other future scenarios due to data being used inappropriately	7	E3	A2, A7	4.9
R5	Inapt, incomprehensible or incomplete data, metadata and documentation by being managed by inexperienced personal	9	E2	A2, A3	2.7
R6	Inapt, incomprehensible or incomplete data, metadata and documentation by errors on data or metadata	9	E4	A2, A3	4.5
R7	Insertion of wrong inputs by human operator	9	E8	A2, A7	6.3
R8	Lack of DM and computer science skilled personal in database management	5	E2	A4	1.5
R9	Lack of financial requirements	7	E5	A3, A7	3.5
R10	Lacking data for the success of the experiment due to legal barriers	9	E13	A2	2.7
R11	Loss of data traceability by non-successful extrapolation of metadata's meaning	7	E12	A2	3.5
R12	Loss of data/metadata/documentation due to errors on the process of storing or sharing data/metadata/documentation	9	E4	A2	4.5
R13	Loss of data/metadata/documentation due to hacker attack	9	E7	A2	2.7

ID	Risks	C	Event	Assets	Severity
R14	Loss of data/metadata/documentation due to human errors	9	E8	A2	6.3
R15	Loss of data/metadata/documentation due to infrastructure failure	9	E9	A2	2.7
R16	Loss of data/metadata/documentation due to natural disasters	9	E11	A2	0.9
R17	Loss of data/metadata/documentation due to software failure	9	E15	A2	4.5
R18	Loss of data/metadata/documentation due to software obsolesce	9	E17	A2	0.9
R19	Loss of metadata denying the representation of the output to the user via Taverna	5	E12	A2	2.5
R20	Sharing confidential data	7	E14	A2	2.1
R21	Stakeholder lack of involvement	3	E0	A4	1.5
R22	System failure due to hacker attack	7	E6	A0, A1, A3, A5, A6	2.1
R23	System failure due to hardware failure	7	E9	A0, A1, A3, A5, A6	2.1
R24	System failure due to hardware maintenance	7	E10	A0, A1, A3, A7, A6	3.5
R25	System failure due to hardware obsolesce	7	E7	A0, A1, A3, A5, A6	2.1
R26	System failure due to human errors	7	E8	A0, A1, A3, A5, A6	4.9
R27	System failure due to natural disasters	7	E11	A0, A1, A3, A5, A6	0.7
R28	System failure due to software failure	7	E15	A0, A1, A3, A5, A6	3.5
R29	System failure due to software obsolesce	7	E17	A0, A1, A3, A5, A6	0.7
R30	Unavailability of storage capacity to respond to data growth	9	E1	A2	4.5

Table 4.3: MetaGen-FRAME risks. Consequence values are presented by the designation (C). The affected assets, trigger events and severity values are also presented

The modules presented in Figure 4.1 also represent the decomposition of the project. Therefore, risk assessment was individually performed for each task. The result of each individual assessment is presented on Table 4.4.

Task	Risks
Data quality control	R2, R3, R4, R7, R11, R12, R13, R14, R15, R16, R17, R18, R19, R21, R22, R23, R24, R25, R26, R27, R28, R29
Analysis of taxonomy	R11, R19, R22, R23, R24, R25, R26, R27, R28, R29
Remote Web-Service (WS)	R2, R3, R4, R5, R6, R8, R9, R10, R12, R13, R14, R15, R16, R17, R18, R22, R23, R24, R25, R26, R27, R28, R29, R30
Alignment	R5, R6, R11, R22, R23, R24, R25, R26, R27, R28, R29, R30
De novo assembly	R2, R3, R4, R22, R23, R24, R25, R26, R27, R28, R29
Functional Annotation	R0, R1, R2, R3, R4, R22, R23, R24, R25, R26, R27, R28, R29
Gene structure prediction	R11, R19, R22, R23, R24, R25, R26, R27, R28, R29
Metabolic Reconstruction	R2, R3, R4, R5, R6, R11, R19, R20, R21, R22, R23, R24, R25, R26, R27, R28, R29

Table 4.4: MetaGen-FRAME results of each individual task assessment

For risk evaluation we used a consequence/probability matrix generated by the HoliRisk and reproduced on Figure 4.4. The technique allows a consistent and holistic view of the risks in terms of their risk level or severity. The colors assigned to the cells of the matrix represent the risk severity.

Risks in light green (at the lower left corner of the matrix) have a very low severity and consequently are not relevant, whether risks in red (at the upper right corner of the matrix) have an extreme severity and consequently must be treated as soon as possible. Therefore it is possible to conclude that, R6 – Inapt, incomprehensible or incomplete data, metadata and documentation by errors on data or metadata; R7 – Insertion of wrong inputs from human operator; R12 - Loss of data/metadata/documentation due to errors on the process of storing or sharing data/metadata/documentation; R14 - Loss of data/metadata/documentation due to human errors; R17 - Loss of data/metadata/documentation due to software failure and R30 – Unavailability of storage capacity to respond to data growth are the more severe risks and should become a priority for risk treatment.

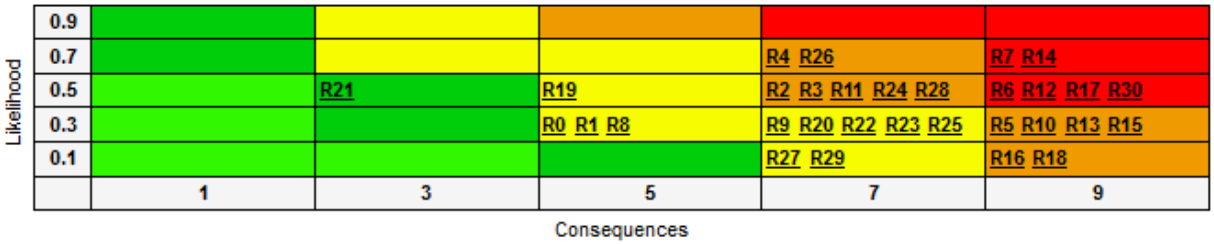

Figure 4.4: MetaGen-FRAME risk matrix

DMP Association

After validation of the risk assessment results, risks were associated with typical DMP sections (see Table 2.2 for more details). The association of the risks with the DMP sections is presented in Table 4.5, where we can see data quality assurance section has the risks with the higher average severity. The average severity was calculated using the values in Table 4.3.

Sections	Risks	Severity (Average)
Ethics and privacy	R20	2.1
Resourcing (Budget)	R9	3.5
Legal Requirements	R0, R1, R10	1.9
Access and Sharing	R2, R3, R4, R12, R13, R14, R15, R16, R17, R18	3.4
Archiving and Preservation	R12, R13, R14, R15, R16, R17, R18, R22, R23, R24, R25, R26, R27, R28, R29, R30	2.9
Stakeholders/Responsibilities	R8, R21	3
Data Formats and Metadata	R5, R6, R19	3.2
Data Quality Assurance	R7, R11	4.9

Table 4.5: MetaGen-FRAME risk results association with DMP sections

This association supports the definition of a DMP by easing the identification of the risks that need to be covered in each section of the document. As an example, the section of resourcing of MetaGen-FRAME's DMP should describe which actions are going to be taken to mitigate the risk of lacking financial requirements (R9).

Risk Treatment

For risk treatment, controls were identified using the SWIFT technique. The controls in section 3.7 also were considered. A workshop was conducted where risks were analyzed individually in order to

identify possible controls. Controls were then evaluated in terms of feasibility to verify if the MetaGen-FRAME project team could apply the controls. Some of the controls can't be applied directly by the MetaGen-FRAME team, belonging the responsibility of their implementation to other entities (risk owners). The final set of controls is presented in Table 4.6. From this set, two policies were defined (see Figure 4.5), considering the several risk owners. Despite this, some controls are shared in more than one policy, meaning they should be implemented by more than one entity/team. Each control can reduce the consequence of a risk, decrease the likelihood of trigger events or reduce vulnerability's exposure.

ID	Controls	Type	Entities
C0	Anti-fire and earthquake measures in the NCBI and computational servers	Consequence, Likelihood	R27, E11
C1	Backup system	Exposure, Consequence, Likelihood	V9, R10, R11, R12, R13, R14, R15, R16, R17, R18, R19, R22, R23, R24, R25, R26, R27, R28, R29, E3, E4, E6, E8, E9
C2	Create a long term storage policy (recovery management Plan)	Exposure, Consequence,	V6, R9, R12, R13, R14, R15, R16, R17, R18, R22, R23, R24, R25, R26, R27, R28, R29, R30, E11,
C3	Create a protocol defining the workflow execution properties or create additional metadata, creating stronger bonds between the biological results	Exposure, Consequence	V12, R2, R3, R4, R7, R19
C4	Create data confidentiality criteria	Exposure, Consequence	V5, R0, R20
C5	Emergency budget for issues in the NCBI	Exposure, Consequence, Likelihood	V2, R9, E11
C6	Emergency budget in case a team member leaves	Exposure, Consequence, Likelihood	V2, R8, R21, E0, E5
C7	Get consent from data sources for data dissemination	Exposure, Consequence, Likelihood	V0, R0, R1, E3
C8	Improve security measures in servers, PCs, data bases and communications	Exposure, Consequence, Likelihood	V8, R13, R22, E6
C9	Insertion of alternative tools in the main workflow, if the main ones fail	Exposure, Consequence Likelihood,	V11, R28, E9, E10, E17
C10	Keep all the software and hardware components up to date	Exposure, Likelihood	V10, E7, E15, E17
C11	Modify formats used by the framework so each output references the associated input data (RDF style). interconnecting data elements	Consequence	R4, R5, R6
C12	Use several forms of documentation	Exposure, Consequence	V4, R4, R11, R19
C13	Usage of fall-back computational servers	Consequence	R22, R23, R24, R25, R26, R27, R28, R29
C14	Usage of fall-back data bases	Consequence, Likelihood	R22, R23, R24, R25, R26, R27, R28, R29, E4
C15	Use open-source tools and formats	Consequence	R4, R11

Table 4.6: MetaGen-FRAME controls with the respective type and entities it mitigates

#	Name	Control Numbers
P0	KDBIO Policy	C11 C3 C2 C10 C15 C9 C13 C6 C12 C8 C14 C1 C4 C7
P1	Repository/database organization Policy (e.g. NCBI Policy)	C2 C10 C15 C0 C5 C8 C1 C4 C7 C6

Figure 4.5: MetaGen-FRAME policies

Finally control measures are associated with DMP typical sections (see Table 4.7). This association helps justifying the treatment options and decisions made in a DMP.

Sections	Controls
Ethics and privacy	C4
Resourcing (Budget)	C2, C5
Legal Requirements	C7
Access and Sharing	C10, C9, C11, C13, C14, C15
Archiving and Preservation	C0, C1, C2, C3, C8, C9, C15
Stakeholders/Responsibilities	C6
Data Formats and Metadata	C2, C12
Data Quality Assurance	C3, C12

Table 4.7: MetaGen-FRAME control results association with DMP sections

Risk Reporting

As a final step for the creation of the RMP, more conclusions are drawn up from the previous results. These conclusions are useful for decision making, being structured for optimizing risk data communication to stakeholders.

In order to understand what type or category of assets is more relevant and valuable in MetaGen-FRAME project and by extension need more protection, using the information given in Figure 4.2, we present Table 4.8 where we group the identified assets into categories and measure the average value of them. This average is measured using all asset's value of a given category. Assets can be divided into categories or according to organization departments, if it becomes a more useful division for decision making. The categories presented correspond to the type of assets defined in collaboration with Miguel Coimbra.

Category	Asset	Value (Average)
Workflow/Business process	A7	Very-high
Data	A2, A3	Very-high
Staff	A4	Medium
Software	A5, A6	High
Hardware	A0, A1, A4	High

Table 4.8: MetaGen-FRAME asset categories and average values

As Table 4.8 shows, all assets that represent the internal structure, process or workflow of the project and the data itself are the most valuable assets and so, the ones that need more protection against vulnerabilities, threats and risks.

In order to understand what category of vulnerabilities has a higher exposure in MetaGen-FRAME project and by extension need more protection, using the information given in Figure 4.3, we present Table 4.9 where we group the identified vulnerabilities into categories and measure the average exposure of them. This average is measured using all vulnerability's exposure of a given category. Vulnerabilities can be divided into categories or according to organization departments, if it becomes a more useful division for decision making. These categories were established in collaboration with Miguel Coimbra.

Category	Vulnerability	Exposure (Average)
Financial (Strategic)	V2	low
Legal	V0, V7	medium
Data	V4, V5, V6, V9, V12	high
Operational (workflow/Hardware/software)	V8, V10, V11	medium
Staff/stakeholders (human)	V1, V3	medium

Table 4.9: MetaGen-FRAME vulnerability categories and average exposures

As Table 4.9 shows, data vulnerabilities have the highest exposure, needing more attention.

Table 4.10 shows the type or category of events that is more likely to happen, needing more attention, by presenting the average likelihood of the event categories. Table 4.10 also shows how many events there are for any category, which can also be helpful in decision making. Likelihood is presented in quantitative values (Ferreira, et al., 2013b). Events can be divided into categories or according to organization departments, if it becomes a more useful division for decision making.

Category	Event	Likelihood (Average)
Financial (Strategic)	E0, E5, E11	0.3
Legal	E13, E14, E16	0.3
Data	E2, E3, E4, E6, E9, E12	0.5
Operational (workflow/Hardware/software)	E1, E7, E9, E10, E15, E17	0.4
Staff/stakeholders (human)	E8	0.7

Table 4.10: MetaGen-FRAME events categories and average likelihoods

As it is showed in Table 4.10, human error or data misuses are the most common events, being the ones that should need more attention.

In what concerns risks, there are several conclusions that can be drawn up, using the previously created risk matrix presented in Figure 4.4, namely which are the risk categories that have a higher priority and so should be treated first. In Table 4.11 we present the risks (see Table 4.3) organized into categories and the average severity of each category. Risks can be divided into categories or according to organization departments, if it becomes a more useful division for decision making. Through Table 4.11 it is possible to see the severity of a certain category of risks or department, improving decision making. This view then complements the one given by the risk matrix (see Figure 4.4) where the same can be viewed concerning individual risks.

Category	Risk	Severity (Average)
Financial (strategic)	R9	3.5
Legal	R0, R1, R10, R20	2
Data	R2, R3, R4, R5, R6, R10, R11, R12, R13, R14, R15, R16, R17, R18, R19	3.4
Operational (workflow/Hardware/software)	R19, R22, R23, R24, R25, R26, R27, R28, R29, R30	2.7
Staff/stakeholders (human)	R7, R8, R21	3.1

Table 4.11: MetaGen-FRAME risk categories, average risk levels and priorities

The categories of risks with higher severity, are the financial and data risks. With this in mind, it is recommended that the data risks should be mitigated first.

In what concerns controls, there are also some conclusions that can be drawn, by organizing controls according to the categories presented in Table 4.9. This is shown in Table 4.12. Some controls may be related to more than one category, meaning they are versatile enough to mitigate several risks, belonging to different categories. Controls can also be divided into categories or according to organization departments, if it becomes a more useful division for decision making.

Category	Control
Financial, organizational or workflow (strategic)	C2, C3, C5, C6
Legal	C4
Data	C1, C2, C11, C12, C14
Operational (Hardware/software)	C0, C1, C7, C8, C9, C10, C13, C14, C15
Staff/stakeholders (human)	C6

Table 4.12: MetaGen-FRAME Control categories

From Table 4.12, we recommend that the controls related to the risk categories with higher severity should be the ones implemented first, especially those who are present in several categories, namely C1, C2, C6, C12 and C14, since with these controls more value is obtained for the money invested, by mitigating several risks from more than one category.

4.2. Case 2 – LNEC

The second case belongs to an engineering project, namely for the Laboratório Nacional de Engenharia Civil (LNEC). This case was also analysed in the scope of the European project TIMBUS (Redlich, et al., 2012).

4.2.1. Method Application and RMP Creation - LNEC case

This section presents the application of the risk management method, and corresponding RMP creation for the LNEC case. To facilitate the understandability of the application, the results are presented according to the steps of the proposed method. These results compose the RMP content, being the structure of this document similar to the succession of steps presented. To present these results, HoliRisk is used, although with certain exceptions since there are a large number of results for

certain entities, which decreases the quality of result display via HoliRisk. In Appendix C a series of proposed improvements to HoliRisk are presented.

Context Establishment

LNEC is a state owned research and development institution that plays a key role in advising the government in technical and scientific matters of the various domains of civil engineering. The context to be analyzed by the risk management method for data management is concerned on the tasks to long-term structural safety control of large dams. In order to produce updated structural safety information, large dams are continuously monitored by sensors that acquire measurements of important physical quantities to characterize the dam behavior (e.g., displacements, strains and stresses, discharges through the foundation). The sensor data is then stored and analyzed to determine if the structure is behaving as expected or, in case of divergences, the reasons for any exchange must be determined (this is performed by structural specialists, based on past data, statistical models, physical models and theoretical models).

LNEC is responsible to maintain an updated archive with safety information of large dams in Portugal, as well as interpreting the data stored in this archive in order to determine the structural safety. In fact, as depicted in Figure 4.6 this project can be seen as three major tasks:

- **Data upload**: the data captured by sensors installed in dams is continuously uploaded into a system that also stores all historical and relevant information to determine the structural safety. This task includes the gathering and ingesting activities;
- **Data store**: the uploaded data is processed (validated, normalized and transformed) and stored in a system. This task includes all activities related to the maintenance and preservation of data;
- **Data access**: the data must be always assessable to authorized users. Specialists request dam safety data to perform their analysis to ensure structural safety;

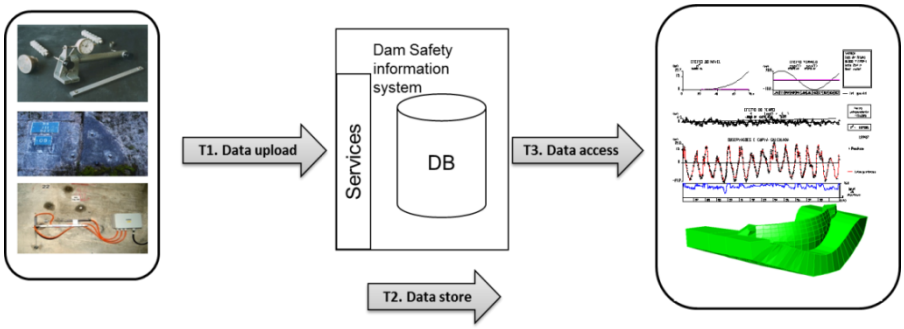

Figure 4.6: LNEC project tasks

Using the proposed RACI matrix (see Table 3.2), the roles and responsibilities were assigned to the corresponding entities, which was performed together with LNEC risk expert José Barateiro. As project sponsor, the LNEC board was considered. For project manager, the dam director was considered. For operational staff, the IT and dam technical staff was considered. For the risk expert role, the IT

research staff was considered. As risk owners, the dam research staff was pointed. Finally for DM experts, the IT research staff was considered.

Risk Assessment

For risk assessment, we opted to use SWIFT and HAZOP techniques. As proposed by the techniques, a workshop with LNEC stakeholders (José Barateiro) was conducted for brainstorming risk assessment activities. As a starting point for the discussion and risk assessment we used the list of typical assets, events, vulnerabilities and risks presented in section 3.7. Certain guidewords and what-if scenarios characteristic of HAZOP and SWIFT techniques also helped with risk assessment (see Appendix E). All criteria (for likelihood, exposure and consequence) used during risk assessment was defined in collaboration with José Barateiro. The likelihood and consequence criteria are presented in Table 4.13. For calculating the risk severity values, the following equation was used (also defined in collaboration with José Barateiro):

$$Severity_{risk} = 2 * Consequence_{risk} + Likelihood_{event} \quad (4.2)$$

This equation was defined in collaboration with the LNEC risk expert José Barteiro.

	Likelihood	Level	Consequence	
1	Never happened before	Irrelevant	No impact or not recognisable impact (risks with this impact can usually be ignored if not occurring at a high frequency)	1
2	Once every 10 years	Minor	Minor impact, affecting only a smaller group of individuals or already mitigated through backup systems	2
3	Once every 2 years	Noticeable	Risk occurrence would have a noticeable impact which should be avoided or mitigated	3
4	Once per half a year	Major	Occurrence of this risk frames the organisation or important parts of it inoperable for a shorter period	4
5	Once per two months	Crucial	Occurrence of this risk frames the organisation or crucial parts of it inoperable for a longer period	5
6	More than once per month	Catastrophic	A single occurrence of a catastrophic risk threatens the organisation to cease to exist (not recoverable)	6

Table 4.13: Likelihood and consequence criteria used for LNEC case

This case's assets are identified in Figure 4.7 as well as the vulnerabilities associated with them. Related to risk identification step, the vulnerabilities found for this case are presented in Table 4.14. In Table 4.15 the events of this case are presented with their respective likelihoods. Table 4.16, shows the case's risks. For each risk we identify it's consequence value, the event that could trigger it, the affected assets and the severity value. As visible on the Table 4.16, risks have relations between them in the sense that a risk could be a trigger event of another risk. As an example, lack of technical support for preservation solution can lead to loss of data, metadata or documentation. This shows a relation between risks and events, where risks can be events and events can be risks.

#	Name	Value	Vulnerabilities Number
A0	Business processes	very-high	V0 V20
A1	Data / metadata	very-high	V8 V10 V6 V17 V18 V4 V16 V5 V2 V15 V3 V14 V12 V7
A2	Databases	very-high	V4 V10 V17 V18 V16
A3	Infraestructures (hardware)	medium	V4 V17 V18 V6 V11
A4	Organization (LNEC)	very-high	V9 V19 V13 V1
A5	Personal	high	V21
A6	Tools (Software)	medium	V18 V11

Figure 4.7: LNEC assets with the respective value and vulnerabilities

ID	Vulnerability	Exposure
V0	Business processes inputs and outputs need to be preserved	High
V1	Change of laws	Medium
V2	Data fragmentation	Medium
V3	Data standards and formats need to be updated	Low
V4	Economic/organizational breakdowns	High
V5	High error rate in data	Medium
V6	Human dependency	High
V7	Lack of a standard for metadata/documentation representation	High
V8	Lack of data confidentiality criteria	Medium
V9	Need for governmental funding	Very-High
V10	Data preservation policy may not be fulfilled	Medium
V11	Possible updates in hardware/software	Medium
V12	Preservation system must comply to preservation plan	High
V13	The introduction or modification of data acquisition systems can cause interoperability problems	High
V14	The preservation community demands data preservation requirements	High
V15	Vital knowledge remains in certain personal	Medium
V16	The processing of personal data is subject to strict data protection rules which are harmonized within the EU by the Data Protection Directive	High
V17	Too large data sets in size or quantity	Very-High
V18	Unreliable hardware	Medium
V19	Unreliable software	Medium
V20	Business process can be unreliable or depend on non-reliable information	Medium
V21	Lack of authentication and authorization mechanisms	Medium

Table 4.14: LNEC vulnerabilities with the respective exposures

ID	Events	Likelihood
E0	Software tool can’t be used or accessed	4
E1	Abandonment of a stakeholder	2
E2	Budget cuts	4
E3	Creation and utilization of new technologies and techniques that increase substantially the quantity of data and metadata generated	4
E4	Data used inappropriately	4
E5	Data’s accuracy, consistency or completeness are not safeguarded	3

ID	Events	Likelihood
E6	Employee with vital knowledge leaves LNEC	3
E7	Corrupt preserved data/metadata	2
E8	Hardware failure	3
E9	Financial, legislative or organizational changes	4
E10	Hardware not enough for preservation system	2
E11	Hardware end of support	2
E12	Human errors	2
E13	Hardware maintenance	5
E14	Impossible to maintain and/or generate sufficient data to preserve semantic information of digital objects, and consumers can't understand the object with the same semantic level (metadata is insufficient or lost)	4
E15	Occurrence of natural disasters	1
E16	LNEC didn't notice a law change	2
E17	LNEC doesn't comply with laws and regulations	3
E18	LNEC´s information in the archive can't be accessed	4
E19	Non successful extrapolation of data/metadata meaning	3
E20	Personal data is processed during sensor data collection	3
E21	Hacker attack	3
E22	The introduction or modification of data acquisition systems lead to interoperability problems	2
E23	Preservation community demands new requirements that cannot be met by the preservation solution	2
E24	Business process depends on non-reliable information	3
E25	The preservation system preserves data without having adequate preservation rights and deposit agreements with the data owners	1
E26	No guarantee that the data is authentic, making it impossible to prove that the information is original and was produced by the responsible users	3
E27	There isn't enough information for render and execute digital objects as in their original form and technological context	3
E28	Tool or formats discontinuation or lack of support	2
E29	Individuals, entities or processes access and disclose classified information	2

Table 4.15: LNEC events with the respective likelihoods

ID	Risks	C	Event	Assets	Severity
R0	Copyright infringement due to preservation of data without the consent of data owners	2	E25	A1, A4	5
R1	Copyright infringement due to violation of data protection regulations	2	E17	A1, A4	7
R2	Lack of Support to New Data Acquisition Systems and Formats due to change of these systems and interoperability issues	3	E22	A1, A0	8
R3	Lack of Technical Support for Preservation Solution due to budget cuts	3	E2	A0, A3	10
R4	Lack of Technical Support for Preservation Solution due to lack of hardware requirements	3	E10	A0, A3	8
R5	Lack of Technical Support for Preservation Solution due to increase of data and metadata generated	3	E3	A0, A3	10
R6	Lack of Support to New Data Acquisition Systems and Formats due to new requirements demanded by the preservation community that cannot be met	3	E23	A0, A1, A4	8
R7	Loss of data authenticity	2	E26	A1	7
R8	Loss of expert knowledge due to stakeholder's abandonment of project	4	E1	A4, A5	10
R9	Loss of data integrity/traceability due to interoperability issues	5	E22	A0, A1	12
R10	Loss of data/metadata/documentation due to financial, legal or organizational changes	6	E9	A0, A1, A4	16
R11	Loss of data/metadata/documentation due to hacker	6	E21	A0, A1,	15

ID	Risks	C	Event	Assets	Severity
	attacks			A4	
R12	Loss of data/metadata/documentation due to hardware failure	6	E8	A0, A1, A3, A4	15
R13	Loss of data/metadata/documentation due to human errors	6	E12	A0, A1, A4, A5	14
R14	Loss of data/metadata/documentation due to insufficient/obsolete hardware for preservation system	6	E10	A0, A1, A4	10
R15	Loss of data/metadata/documentation due to natural disasters	6	E15	A0, A1, A4	13
R16	Loss of data/metadata/documentation due to software failure	6	E0	A0, A1, A4, A6	16
R17	Loss of data/metadata/documentation due to software obsolesce	6	E28	A0, A1, A4, A6	14
R18	Loss of data/metadata/documentation for inappropriate use of these assets	6	E4	A0, A1	16
R19	Loss of expert knowledge due to loss of employee with vital knowledge	4	E6	A4, A5	11
R20	Loss of integrity/traceability of data due to errors on data storage and dissemination	5	E7	A0, A1	12
R21	Loss of integrity/traceability of data due to non-successful extrapolation of data/metadata meaning	5	E19	A0, A1	13
R22	Loss of integrity/traceability of data due to unavailability to generate sufficient data for preserving the corresponding semantic characteristics	5	E14	A0, A1	14
R23	Loss of integrity/traceability of data due to unguarded data accuracy, consistency and completeness	5	E5	A0, A1	13
R24	Non-compliance with general legal obligations due to preservation of data without adequate rights	4	E25	A4	9
R25	Non-compliance with general legal obligations due to a law change overlook by LNEC	4	E16	A4	10
R26	Non-compliance with general legal obligations due to data regulations non-compliance by LNEC	4	E17	A4	11
R27	Non-compliance with general legal obligations due to hacker attacks and discloser of protected data	4	E29	A4	10
R28	Non-compliance with general legal obligations due to human data confidentiality violations	4	E20	A4	11
R29	Preservation Strategy/Plan cannot be Met due to inability of fulfilling new preservation community requirements	3	E23	A4	8
R30	Preservation Strategy/Plan cannot be Met due to insufficient hardware	3	E10	A4	8
R31	Shortcomings in semantic understandability	4	E27	A1	11
R32	Shortcomings in technical understandability	4	E27	A1	11
R33	Budget cuts due to Investor's abandonment of project	4	E1	A4	10
R34	Stakeholder lack of involvement due to loss of resources	4	E6	A4, A5	11
R35	System/service failure due to dependency of business processes on non-reliable information	3	E24	A0, A1, A2, A3, A4, A6	9
R36	System/service failure due to hacker attack	3	E21	A0, A1, A2, A3, A4, A6	9
R37	System/service failure due to hardware failure	3	E8	A0, A1, A2, A3, A4, A6	9
R38	System/service failure due to hardware maintenance	3	E13	A0, A1, A2, A3, A4, A6	11

ID	Risks	C	Event	Assets	Severity
R39	System/service failure due to hardware obsolesce	3	E11	A0, A1, A2, A3, A4, A6	8
R40	System/service failure due to human errors	3	E12	A0, A1, A2, A3, A4, A6	8
R41	System/service failure due to natural disasters	3	E15	A0, A1, A2, A3, A4, A6	7
R42	System/service failure due to non-access of stored data	3	E18	A0, A1, A2, A3, A4, A6	10
R43	System/service failure due to non-fulfilment of preservation requirements	3	E23	A0, A1, A2, A3, A4, A6	8
R44	System/service failure due to software failure	3	E21	A0, A1, A2, A3, A4, A6	9
R45	System/service failure due to software obsolesce	3	E28	A0, A1, A2, A3, A4, A6	8
R46	Unavailable storage capacity concerning data growth	6	E3	A1, A2, A3, A4	16
R47	Violation of Intellectual property	3	E25	A1, A4	7

Table 4.16: LNEC risks with the respective trigger events, assets, consequence values and severities. Consequence values are presented by the designation (C)

The tasks presented in Figure 4.6 also represent the decomposition of the project. Therefore, risk assessment was individually performed for each task. The result of each individual assessment is presented on Table 4.17.

Task	Risks
Data Upload	R1, R2, R6, R8, R19, R29, R20, R21, R22, R23, R30, R31, R32, R33, R35, R36, R37, R38, R39, R40, R41, R42, R43, R44, R45, R47
Data Store	R0, R3, R4, R5, R8, R10, R11, R12, R13, R14, R15, R16, R17, R18, R19, R20, R21, R22, R23, R24, R25, R26, R27, R28, R29, R30, R34, R35, R36, R37, R38, R39, R40, R41, R42, R43, R44, R45
Data Access	R1, R2, R6, R7, R8, R9, R10, R11, R12, R13, R14, R15, R16, R17, R18, R20, R21, R22, R23, R24, R25, R26, R27, R28, R30, R31, R32, R35, R36, R37, R38, R39, R40, R41, R42, R43, R44, R45, R46, R47

Table 4.17: Results of each individual task assessment

For risk evaluation we used a consequence/probability matrix generated by the HoliRisk and reproduced on Figure 4.8. The technique allows a consistent and holistic view of the risks in terms of their severity. With Figure 4.8 it is possible to conclude that, R10 - Loss of data/metadata/documentation due to financial, legal or organizational changes; R11 - Loss of data/metadata/documentation due to hacker attack; R12 - Loss of data/metadata/documentation due to hardware failure; R16 - Loss of data/metadata/documentation due to software failure; R18 - Loss of data/metadata/documentation for inappropriate use of these assets and R46 – Unavailability of storage capacity concerning data growth are the more severe risks that should become a priority for risk treatment.

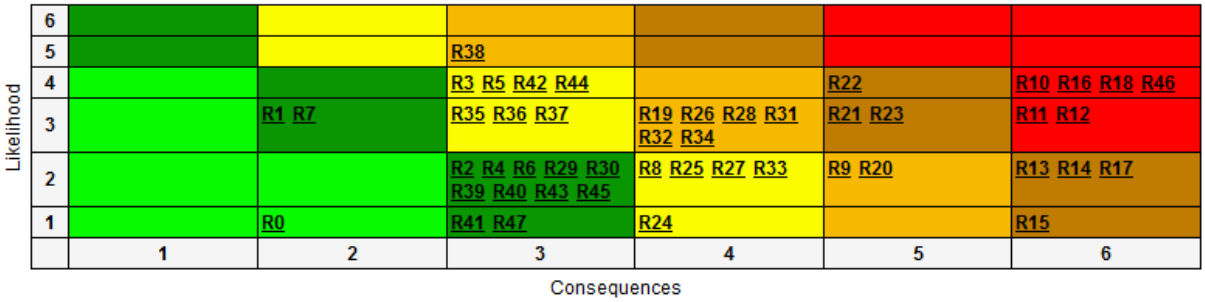

Figure 4.8: LNEC risk matrix

DMP Association

After validation of the risk assessment results, risks were associated with typical DMP sections which is presented in Table 4.18. As we can see, resourcing, archiving and preservation and data formats and metadata are the DMP sections with risks with higher severity and in need of more attention.

Sections	Risks	Severity (Average)
Ethics and privacy	R0, R1, R27, R28, R47	8
Resourcing (Budget)	R10, R33	13
Legal Requirements	R0, R1, R24, R25, R26, R27, R28	9
Access and Sharing	R35, R36, R37, R38, R39, R40, R41, R42, R43, R44, R45	9
Archiving and Preservation	R3, R4, R5, R10, R11, R12, R13, R14, R15, R16, R17, R18, R29, R30, R46	13
Stakeholders/Responsibilities	R8, R19, R33, R34	11
Data Formats and Metadata	R2, R6, R10, R11, R12, R13, R14, R15, R16, R17, R18,	13
Data Quality Assurance	R7, R9, R20, R21, R22, R23, R31, R32	12

Table 4.18: LNEC risk results association with DMP sections

This association supports the definition of a DMP by easing the identification of the risks that need to be covered in each section of the document. As an example, the section of resourcing of LNEC's DMP should describe which actions are going to be taken to mitigate the risks of loss of data/metadata or documentation due to financial loss (R10) and the risk of budget cuts due to investor's abandonment of project (R33).

Risk Treatment

For risk treatment, controls were identified using the SWIFT technique. The controls from section 3.7 were also useful. A workshop was conducted with José Barateiro where risks were analyzed individually in order to identify possible controls. Controls were then evaluated in terms of feasibility to verify if the LNEC team could apply the controls. The final proposed set of controls is presented in Table 4.19, with the respective type of control indicated and the entities (risks, events or vulnerabilities) they mitigate. From this set, three policies were defined (see Figure 4.9). Some controls are shared in more than one policy, meaning they should be implemented by more than one LNEC entity/team.

ID	Controls	Type	Entities
C0	Be on top of the most recent formats and data standards	Exposure, Consequence	V3, R2, R6, R9, R20, R21, R22, R23, R29
C1	Create/improve recovery management Plan	Exposure, Consequence, Likelihood	V10, R3, R4, R5, R10, R11, R12, R13, R14, R15, R16, R17, R18, R30, R33, R35, R36, R37, R38, R39, R40, R41, R42, R43, R44, R45, R46, E7, E8, E15, E18
C2	Creation of data, metadata and documentation standards for preservation community information sharing	Exposure, Consequence, Likelihood	V7, R2, R9, R20, R21, R22, R23, R29, R31, R32, E3, E7, E23
C3	Development of appropriate software to accommodate new accepted techniques in the scientific community	Likelihood	E22, E23
C4	Encourage formation and training	Exposure, Consequence, Likelihood	V15, R8, R19, E1, E6, V6
C5	Enforce documentation of systems/processes	Likelihood, Consequence	V20, R8, R11, R13, R18, R19, R21, E24, E27
C6	Ensure and enforce copyrights of data	Likelihood, Consequence	R0, R1, R27, R28, R47, E29
C7	Improve security measures	Likelihood, Consequence	R11, R36, E21
C8	Internal and external legal audits	Consequence	V1, R0, R1, R24, R25, R26, R27, R28, R47, E17, E18, E20
C9	Internal data monitoring	Consequence	R7, R10, R11, R12, R13, R14, R15, R16, R17, R18, R20, R21, R22, R46
C10	Keep all software and hardware components up to date	Likelihood, Consequence	V11, R14, R17, R39, E10, E28
C11	Redundancy: backup system	Exposure, Consequence, Likelihood	V0, V17, R10, R18, E4, E7, E8, E12, E21, E27
C12	Separate personal data from the remaining data	Consequence	R27, R28
C13	Store backup hardware components	Likelihood, Consequence	R12, R14, R37, R38, R39, E11
C14	Sustainability plan	Exposure, Consequence, Likelihood	V4, V9, R3, R10, R33, R34, E2, E9
C15	Usage of fall back data bases	Likelihood, Consequence	R10, R11, R12, R13, R14, R15, R16, R17, R18, E7, E8, E12, E21, E27
C16	Use fall back tools	Consequence	R16, R17, R44, R45, E28
C17	Use open-source tools and formats	Likelihood, Consequence	R9, R29, E28
C18	Use several forms of documentation	Consequence	R10, R11, R12, R13, R14, R15, R16, R17, R18
C19	Use third party to enforce data protection regulations	Likelihood, Consequence	R24, R30, R43, E23

Table 4.19: LNEC controls, with the respective entities they mitigate. They can mitigate risks (type Consequence), mitigate events (type Likelihood) or mitigate vulnerabilities (type Exposure)

#	Name	Control Numbers
P0	Law protection	C8 C12 C6
P1	Protection and reutilization of data	C11 C7 C1 C18 C6 C10 C17 C2 C19 C0 C8 C4 C12 C5 C9 C15
P2	Protection of software / Hardware	C10 C14 C17 C3 C4 C7 C16 C13

Figure 4.9: LNEC policies

Finally control measures are associated with DMP typical sections (see Table 4.20). This association of risk controls to DMP sections helps justify the application of control measures in each DMP section.

Sections	Controls
Ethics and privacy	C6
Resourcing (Budget)	C14
Legal Requirements	C6, C8, C12, C19
Access and Sharing	C7, C9, C10, C12, C16
Archiving and Preservation	C1, C7, C9, C10, C11, C12, C13, C15, C16
Stakeholders/Responsibilities	C4, C12
Data Formats and Metadata	C0, C2, C5, C9, C17, C18
Data Quality Assurance	C0, C3, C9

Table 4.20: LNEC control results association with DMP sections

Risk Reporting

As a final step for the creation of the RMP, some conclusions are drawn up from the previous results. These conclusions are then useful for decision making, summarizing all the important risk data and conclusions easing the communication of this information to stakeholders.

In order to understand what type or category of assets is more valuable for LNEC and by extension need more protection, we present Table 4.21 where we group the identified assets into categories and measure the average value of them. This average is measured using all asset's value of a given category. Assets can be divided into categories or according to organization departments, if it becomes a more useful division for decision making. These types or categories of assets were defined in collaboration with José Barateiro.

Category	Asset	Value (Average)
Workflow/Business process	A0, A4	Very-high
Data	A1, A2	Very-high
Staff	A5	High
Software	A6	Medium
Hardware	A3	Medium

Table 4.21: LNEC asset categories and average values

As Table 4.21 shows, assets that represent the internal processes or workflows and the data itself are the most valuable and in need for more protection against vulnerabilities, threats and risks.

In order to understand what category of vulnerabilities has a higher exposure in the LNEC case and by extension need more attention, using the information given in Table 4.14, we present Table 4.22 where we group the identified vulnerabilities into categories and measure the average exposure of them.

This average is measured using all vulnerability's exposure of a given category. Vulnerabilities can be divided into categories or according to organization departments, if it becomes a more useful division for decision making. These categories were established in collaboration with José Barateiro.

Category	Vulnerability	Exposure (Average)
Financial (Strategic)	V4, V9	High
Legal	V1	Medium
Data	V0, V2, V3, V5, V8, V10, V12, V13, V14, V16, V17	High
Operational (workflow/Hardware/software)	V11, V18, V19, V20, V21	Medium
Staff/stakeholders (human)	V6, V15	Medium

Table 4.22: LNEC vulnerability categories and average exposures

As Table 4.22 shows, financial and data vulnerabilities have the highest exposure, needing more attention. Table 4.23 shows the type or category of events/threats that is more likely to happen, needing more attention, by presenting the average likelihood of the event categories. Table 4.23 also shows how many events there are for any category, which can also be helpful in decision making. Events can be divided into categories or according to organization departments, if it becomes a more useful division for decision making.

Category	Event	Likelihood (Average)
Financial (strategic)	E2, E9, E15, E25	2.6
Legal	E16, E17,	2.5
Data	E3, E4, E5, E7, E14, E19, E20, E21, E26, E27, E29	3
Operational (workflow/Hardware/software)	E0, E8, E10, E11, E13, E18, E22, E23, E24, E25, E28,	2.6
Staff/stakeholders (human)	E1, E6, E12	2.3

Table 4.23: LNEC events categories and average likelihoods

As it is showed in Table 4.23, data related events are the most common, being the ones that should need more attention. In what concerns risks, there are several conclusions that can be drawn up, using the previously created risk matrix presented in Figure 4.8, namely which are the risk categories that have a higher severity and so should be treated first. In Table 4.24 we present the risks (see Table 4.16) organized into categories and the average severity (using the values from Table 4.16). Risks can be divided into categories or according to organization departments, if it becomes a more useful division for decision making. This view complements the one given by the risk matrix where the same can be viewed concerning individual risks.

Category	Risk	Severity (Average)
Financial (strategic)	R10, R33	13
Legal	R0, R1, R24, R25, R26, R27, R28, R47	9
Data	R7, R9, R10, R11, R12, R13, R14, R15, R16, R17, R18, R20, R21, R22, R23, R31	13
Operational (workflow/Hardware/software)	R2, R3, R4, R5, R6, R29, R30, R32, R35, R36, R37, R38, R39, R40, R41, R42, R43, R44, R45, R46	9
Staff/stakeholders (human)	R8, R19, R34	11

Table 4.24: LNEC risk categories, average risk levels and priorities

The categories of risks with higher severity are the data and financial risks. With this in mind, it is recommended that these risks should be threated first.

In what concerns controls, there are also some conclusions that can be drawn, by organizing them according to the categories presented in Table 4.24. This is shown in Table 4.25. Some controls may be related to more than one category, meaning they are versatile enough to mitigate several risks, belonging to different categories. Just like risks, controls can also be divided into categories or according to organization departments, if it becomes a more useful division for decision making.

Category	Control
Financial, organizational or workflow (strategic)	C1, C5, C14, C19
Legal	C6, C8
Data	C0, C2, C3, C7, C9, C11, C12, C15, C18
Operational (Hardware/software)	C3, C7, C10, C11, C13, C16, C17
Staff/stakeholders (human)	C4

Table 4.25: LNEC control categories

From Table 4.25, we recommend that the controls related to the more prioritized risk categories should be the ones implemented first, especially those who are present in several categories, namely C3, C7, and C11 since with these controls more value is obtain for the money invested, by mitigating several risks from more than one category.

4.2.2. Analysis of Two Approaches

Since a RM analysis was made for the LNEC case in TIMBUS project (Redlich, et al., 2012), it is possible to compare both approaches, finding some differences and common aspects.

- **Both approaches follow ISO 31000 guidelines**.
- **Both approaches use a similar baseline of 31010 risk assessment techniques**: Brainstorm, SWIFT, check-lists and likelihood/consequence risk matrix are used in both cases. Although, in the TIMBUS analysis, the HAZOP technique wasn't used for risk assessment. The proposed method, by using both SWIFT and HAZOP together, manages to find more RM results related to the systems and business processes of LNEC, namely risks, events or vulnerabilities that SWIFT alone could not find. This happens due to the usage of both guidewords and what-if scenarios, which stimulates the brainstorming process on interviews.
- **TIMBUS approach doesn't consider the existence of a DMP**: This can become troublesome in scenarios where DMP are required. Although, the method proposed, by considering a hypothetical DMP made for the scenario in question and by associating risk results with the requirements of this DMP, creates an overhead of time and effort to perform the RM analysis (this overhead isn't present in the TIMBUS analysis, since no DMP is considered).
- **TIMBUS analysis only used the risk matrix as reporting tool for the involved stakeholders:** It doesn't supply reporting aids more friendly for high executives, who would be more interested in finding details concerning a certain set or category of risks, or in finding details concerning risks related to a certain department of LNEC.

- **On the TIMBUS analysis, only three risk types were identified, namely strategic, operational and legal**: In the proposed method, for risk reporting needs, more categories were considered. This allows a more clear knowledge of the several types of risks related to LNEC, as well as the several types of controls that are needed to mitigate the former risks. The category definition in the proposed method, supported by risk expert José Barateiro, was inspired from the study of DMP typical sections. This study was ignored in TIMBUS analysis.
- **In the proposed method, there was a detection of irrelevant risks**: This was possible through the association of risks with DMP sections, thus identifying if a risk is relevant for DM. A risk that does not suit any of the DMP sections is probably out of the context of DM and can be discarded;
- **Better starting point for RM analysis using the proposed method**: Having applied the method to the MetaGen-FRAME case first, more complete check-lists were developed.
- **A RACI matrix is detailed in the proposed method**: This matrix is used to support scope definition by defining generic roles and responsibilities detailed for each involved task that needs to be performed. TIMBUS analysis doesn't use such technique.
- **LNEC has a specific DM concern, namely DP**: The analysis made in TIMBUS is more oriented for the DP concern. The proposed method addresses DM concerns in a generic sense, so it doesn't specialize only in a certain DM concern like DP, but rather in all DM concerns, although DP concern is also included in the RM analysis performed.

The former differences and common aspects between both approaches show that they are not so different, since they share the same principle baseline, which means the proposed approach doesn't require impractical requirements. Although, our approach presents some enhancements, bringing extra value to RM analysis.

5. Conclusions and Future Work

In engineering and science projects DM is a well-known concern. Currently, to answer DM concerns, DMP are developed as part of the project's proposal. However, DMP don't take into consideration RM concerns, although DM and RM share objectives and represent two different ways of addressing DG concerns. Also, stakeholders responsible for creating the document have difficulty identifying DM problems and justifying the respective solutions which consequently hardens the assurance that their DMP is exemplary of good practices. As DMP do not take into consideration all DG concerns, namely the concerns related to RM, this document isn't sufficient to assure a solid DG. This means that there is room for DMP improvement.

In this dissertation we propose a risk management method to support the definition of a DMP. The method means to create a RMP that justifies all the decisions and controls defined in each section of the DMP. This method is based on well-known risk management references, namely ISO 31000 guidelines and ISO 31010 techniques, being specifically designed for engineering and science projects although, potentially, it could be used in other domains. The method is applied in two different scenarios, one regarding a science project in the field of Metagenomics (MetaGen-FRAME) and another, representing the field of engineering, which regards a civil engineering project for dam safety belonging to LNEC.

After the application of the method to both cases, two sets of results were gathered, from where several conclusions were drawn. To support the application storage, management and presentation of these results and associated conclusions, a web tool called HoliRisk was used.

Besides the method, also a risk registry is proposed concerning engineering and science projects. This registry became useful for the beginning of the RM analysis, helping the definition of check-lists, which then can be used to support risk assessment. Finally, since both DMP and RMP try to govern data and other valuable project assets, our proposal unifies DM and RM principles and techniques, and in practice DMP and RMP efforts, to promote a more solid DG for engineering and science projects. In other words our proposed method promotes DG for projects belonging to the domain areas of this dissertation. To support this, our proposal includes a set of skills we deemed necessary for any risk expert that applies our method and a set of typical stakeholder roles and responsibilities to help with decision making and accountability assignment.

The results obtained were deemed relevant by both case surveyed representatives that were interviewed during the application of our method, namely Miguel Coimbra and José Barateiro. Concerning the LNEC case, since it was also analyzed in TIMBUS project, it became possible a comparison between both approaches (see section 4.2.2), which lead to the detection of several similar aspects and differences. With all our results and achievements we can conclude that our objectives were fulfilled and our hypotheses were confirmed.

Our work presents the following main contributions:

- DG approach developed for engineering and science projects, through our proposed method based on the combine utilization of DMP and RMP, with the support of the suggested skill set and the presented stakeholder roles and responsibilities.
- Detection of irrelevant risks and having better understanding of the DM problem through the association of risks and controls with DMP sections, determining the relevant ones for DM.
- Improvement of DMP, by justifying the decisions and controls through the application of suited RM techniques and guidelines in a form of a RMP.
- Method suited for engineering and science projects, being possible to use it for other domains due to its generic characteristics.
- The new set of skills suited for a risk expert represents an opportunity for librarians and archivists, since they can assume, in the future, the role of risk experts.

5.1. Lessons Learned

During the development of this dissertation, several lessons were learned and some difficulties were met.

First, it was impossible to determine beforehand the required number of iterations needed to achieve a good set of results (being this number unpredictable). Determine if a new iteration was required was also difficult, since opinions diverged in the analysis team to determine if risk assessment or risk treatment results were sufficient. This difficulty was enhanced due to the several changes on the method itself, since any change meant the need for a new iteration.

Meetings with project or organization personal were required due to the techniques that were chosen. This raised the difficulty of combining all requirements and concerns of different stakeholders and translate them into assets, vulnerabilities, threats, risks and controls. This fact sometimes delayed or stalled the evolution of the RM analysis.

For measuring event likelihood and risk consequence values, it became clear that it is impractical to determine beforehand the range values or risk criteria used (for risk analysis and evaluation) and if they should be qualitative, semi-quantitative or quantitative. In the majority of times, quantitative evaluations, despite being desirable, can´t be performed and so qualitative measures need to be applied, which means dealing with possible disadvantages, like less value accuracy. From this it becomes clear that the definition of these values must be guided and elaborated together with the project or organization experts, since they are the most qualified to determine these values. This happens due to the subjectivity of the stakeholders involved and the unique business processes of each case study. Also, in most cases it becomes impossible to quantify (quantitatively) the risk indicators.

Another difficulty was to determine, not only which risk data was relevant to present to stakeholders, but also to understand the best way this risk data would be presented.

Choosing the most suited risk assessment techniques for every stage of risk assessment from ISO 31010 was also challenging. Since every engineering and science project is different, where there's some difficulty in certain cases for quantifying risk indicators in a quantitative manner, we noticed that more subjective techniques, based on brainstorm, were more suited than mathematical or analytical techniques, producing more relevant results. This kind of techniques were also more suited, since they allow the vital cooperation of the interview personal, as well as the already existing and established expert knowledge in the field in question (knowledge expressed through check-lists).

Finally, yet another difficulty rose. During the analysis of both methods, it became clear that some events could be seen as risks and some risks could also be seen as events. Until this point, both elements were considered distinct entities. This became particularly troublesome, since HoliRisk didn't took this fact into consideration, hardening the work developed.

5.2. Future Work

HoliRisk could be used to fully support the proposed method. At this point only supports the risk registry functionality (responsible for storing risk related data) and some basic requirements for the risk reporting service. A specific risk reporter for the domain of engineering and science still needs to be developed. In Appendix C a more concrete list of requirements for the development of this new risk reporter is presented.

During this dissertation, deliverable 8.2 from TIMBUS analysis on LNEC was analysed. Deliverable 8.4 could also be analysed as a future scenario, since it also presents an issue of DM, namely the preservation of CAD models, being this issue also suited for application of the presented method.

Linked data and open linked data also fit the criteria for a RM analysis using the presented method, since both topics involve large issues of DM, namely the management of large data sets, including their reutilization, in a private or open environment, which can also raise ethical and licensing problems and risks.

The scope of this dissertation was mainly focused on RM and understanding how these principles can be used to help improve DM. The making of DMP was always implicit. Currently, some new tools are being developed, allowing the creation of DMP online according to the guidelines of a certain funding agency. These tools, like DMPTool[4] and DMPOnline[5], are being currently used but they still need to evolve and mature in order to become truly useful. Once these tools are mature enough, they can be used to create DMP, which in turn, can be used to test the method suggested in this dissertation, in order to access the efficacy of the results generated, giving a more clear understanding of how the generated results can be used to complement the original DMP.

Projects with small budgets (up to one million) probably won't be able to produce both DMP and RMP. The presented skills can give a response, where in a near future, librarians and archivists can learn to

[4] https://dmp.cdlib.org/
[5] https://dmponline.dcc.ac.uk/

create RMP besides the DMP, minimizing duplication of costs and efforts. Although this issue stills needs a more detailed analysis.

The presented proposal can also be used for compliance insurance, where it can be adapted to an approach to perform data repositories auditing.

Finally, the method presented was developed concerning engineering and science requirements and projects. For future work, this method can be adapted to other fields of interest, namely any field or type of project, where DM concerns are present and relevant. In other words, this method has the potential to be extended to other fields of action, given the need for a DMP is present to address the corresponding DM concerns involving the particular field or project. This would involve improving the risk registry associated with the proposed method, expanding it to other fields. New techniques could also be considered to perform risk assessment and risk treatment.

References

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

Barateiro, J. (2012). A Risk Management Framework Applied to Digital Preservation. Universidade Técnica de Lisboa, Instituto Superior Técnico.

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

Beagrie, N. (2006). Digital Curation for Science, Digital Libraries, and Individuals. *The International Journal of Digital Curation*.

Bimholtz, J., & Bietz, M. (2003). ata at Work: Supporting Sharing in Science and Engi-neering. *GROUP '03 Proceedings of the 2003 international ACM SIGGROUP conference.*

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

Braga, R. (2007). Automatic capture and efficient storage of e-Science experiment provenance, . *Wiley InterScience*.

Brennan, R. (2011). *Engineering Data Lifecycle Management Overview.* IBM NRSC Perth.

Canteiro, S. (2011). *Risk Assessment in Digital Preservation.* Universidade Técnica de Lisboa, Instituto Superior Técnico.

Caralli, R., Stevens, J., Young, L., & Wilson, W. (2007). OCTAVE Allegro: Improving the Information Security Risk Assessment Process. *oftware Engineering Institute at Carnegie Mellon University*.

Coimbra, M. (2013). *MetaGen-FRAME: Metagenomics Data Analysis Framework Focused on Stressed Microbial Communities.* Universidade Técnica de Lisboa, Instituto Superior Técnico.

Commerce, O. o. (2007). *Management of Risk: Guidance for Prac-titioners (M_o_R).* United Kingdom,: Office of Government Commerce (OGC).

Council, A. C. (2010). *Corporate Governance Principles and Recommendations.*

Crowston, K., & Qin, J. (2011). A Capability Maturity Model for Scientific Data. *Proceedings of the American Society for Information Science and Technology*, 10-19.

Darlington, M., Ball, A., Howard, T., Culley, S., & McMahon, C. (2010). *Principles for Engineering Research Data Management.* University of Bath.

Day, M. (2004). Preservation metadata initiatives: practically, sustainability, and interop-erability. *University of Bath*.

Deelman, E., & Chervenak, A. (2008). Data Management Challenges of Data-Intensive Sci-entific Workflows. *USC Information Sciences Institute*.

Development, E. C. (1947). *Canons of ethics for engineers.* New York: Engineers' Council for Professional Development.

Fernandes, D., Bakhshandeh, M., & Borbinha, J. (2012). *Survey of data management plans in the scope of scientific research.* INESC-ID. TIMBUS Timeless Business.

Ferreira, F., Coimbra, M., Bairrão, R., Vieira, R., Freitas, A. T., Russo, L. M., et al. (2014). Data Management in Metagenomics: A Risk Management Approach. *IDCC14*.

Ferreira, F., Coimbra, M., Vieira, R., Bairrão, R., Freitas, A. T., Russo, L. N., et al. (2013b). *A Risk Management Plan in Metagenomics.* INESC-ID.

Ferreira, F., Coimbra, M., Vieira, R., Proença, D., Freitas, A. T., Russo, L. N., et al. (2013a). Risk aware Data Management in Metagenomics. *InForum.* Évora.

Griffin, J. (2010). *Principles and benefits of effective data governance.* Deloitte.

Hey, T., & Trefethen, A. (2003). E-Science and its Implications. *The Royal Society*.

Hillson, D., & Simon, P. (2007). *Risk Management Plan template.* ATOMrisk.

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

Institute of Management Accountants (IMA). (2007). *Enterprise Risk Management: Tools and Techniques for Effective Implementation.*

ISO FDIS 31000. (2009). *ISO (2009) Risk Management - Principles and guidelines.* Geneva, Switzerland: ISO FDIS.

ISO Guide 73. (2009). *ISO (2009) Risk management – Vocabulary.* Geneva, Switzerland: ISO.

ISO IEC 27001. (2013). *Information technology — Security techniques — Information security management systems — Requirements.*

ISO IEC 27005. (2011). *Security technologies – Information security risk management.* Geneva, Switzerland: ISO IEC.

ISO IEC 31010. (2009). *ISO (2009) Risk management - Risk assessment techniques.* Geneva, Switzerland: ISO IEC.

IT Governance Institute. (2009). *The Risk IT Framework (Exposure Draft).* IT Governance Institute.

Jankowski, N. (2007). Exploring e-Science: An Introduction. *ournal of Computer-Mediated Communication*.

Kaye, J., Boddington, P., Vries, J., Hawkins, N., & Melham, K. (2010). Ethical implications of the use of whole genome methods in medical research. *Europe Journal of Human Genetics*.

Khatri, V., & Brown, C. (2010). Designing data governance. *Communications of the ACM*, 53.1: 148-152.

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

Nguyen, T., Guellec, P., Féru, F., Maillé, B., & Yannou, B. (2007). Definition of an Engineering Data Management for Collaborative Product Development. *INTERNATIONAL CONFERENCE ON ENGINEERING DESIGN*, 28 - 31.

NIST. (2012). *NIST Special Publication 800-30 (Revision 1).* Risk Management Guide for IT Systems.

Oracle. (2011). *Entermprise Information Management: Best Practices in Data Governance.*

Peffers, K., Tuunanen, T., Rothenberger, M., & Chatterjee, S. (2007). A Design Science Research Methodology for Information Systems Research. *Journal of Management Information Systems*, 45–77.

PEFFERS, K., TUUNANEN, T., ROTHENBERGER, M., & CHATTERJEE, S. (2007). A Design Science Research Methodology for Information Systems Research. *Journal of Management Information Systems*, 45–77.

Ramirez, D. (2008). Risk Management Standards: The Bigger Picture. *Information Systems Control Journal*, Vol. 4.

Redlich, D., Molka, T., Gilani, W., Barateiro, J., Miranda, P., Lucas, A., et al. (2012). *Deliverable 8.2: Use Case Specific Risks.* INESC-ID. TIMBUS Timeless Buiseness.

Schmitt, C., & Burchinal, M. (2011). Data management practices for collaborative re-search. *Frontiers in Psychiatry*.

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

Tread, C. o. (2004). Enterprise Risk Management — Integrated Framework. *NJ: AICPA.* Jersey City: Committee of Sponsoring Organizations of the Tread.

Vermaaten, S., Lavoie, B., & Caplan, P. (2012). Identifying Threats to Successful Digital Preservation: the SPOT Model for Risk Assessment. *D-Lib Magazine*.

Wooley, J., Godzik, A., & Friedberg, I. (2010). A Primer on Metagenomics. *Plos Computational Biology*.

Appendix A - Glossary

The most relevant concepts (including abbreviations) related to RM, DP, engineering, science, DM and DMP are presented. These concepts are in accordance with (ISO Guide 73, 2009), (ISO FDIS 31000, 2009), (Hey & Trefethen, 2003), (Braga, 2007), (Deelman & Chervenak, 2008), (Vermaaten, Lavoie, & Caplan, 2012), (Barateiro, Antunes, Freitas, & Borbinha, 2010), (Fernandes, Bakhshandeh, & Borbinha, 2012), (Coimbra, 2013) (Darlington, Ball, Howard, Culley, & McMahon, 2010) (Development, 1947).

Risk – Effect of uncertainty on objectives.

Risk Management (RM) - Coordinated activities to direct and control an organization with regard to risk.

Asset – Anything of value to the organization.

Event - Occurrence or change of a particular set of circumstances.

Risk Management Policy - Statement of the overall intentions and direction of an organization related to risk management.

Risk Management Framework - Set of components that provide the foundations and organizational arrangements for designing, implementing, monitoring, reviewing and continually improving risk management throughout the organization.

Risk Management Process - Systematic application of management policies, procedures and practices to the activities of communicating, consulting, establishing the context, and identifying, analysing, evaluating, treating, monitoring and reviewing risk.

Risk Management Plan (RMP) **-** Scheme within the risk management framework specifying the approach, the management components and resources to be applied to the management of risk.

Stakeholder – Person or organization that can affect, be affected by, or perceive themselves to be affected by a decision or activity.

Risk Perception - Stakeholder's view on a risk.

Vulnerability - Intrinsic properties of something resulting in susceptibility to a risk source that can lead to an event with a consequence.

Threat - Circumstance or event with the potential to adversely impact an asset through unauthorized access, destruction, disclosure, modification of data, and/or denial of service.

Likelihood - Chance of something happening**.**

Consequence - Outcome of an event affecting objectives.

Severity – Results in the combination of risk consequence and event likelihood.

Risk Assessment – Overall process of risk identification, risk analysis and risk evaluation.

Risk Identification - Process of finding, recognizing and describing risks.

Risk Analysis - Process to comprehend the nature of risk and to determine the level of risk.

Risk Treatment – Process to modify risk.

Risk Owner - Person or entity with the accountability and authority to manage a risk.

Risk Matrix - Tool for ranking and displaying risks by defining ranges for consequence and likelihood.

Risk Level - Magnitude of a risk or combination of risks, expressed in terms of the combination of consequences and their likelihood.

Risk Criteria - Terms of reference against which the significance of a risk is evaluated.

Monitoring - Continual checking, supervising, critically observing or determining the status in order to identify change from the performance level required or expected.

Review - Activity undertaken to determine the suitability, adequacy and effectiveness of the subject matter to achieve established objectives.

Control - Measure that is modifying risk.

Communication and consultation - Continual and iterative processes that an organization conducts to provide, share or obtain information, and to engage in dialogue with stakeholders regarding the management of risk.

Residual Risk - Risk remaining after risk treatment.

Internal Context – Internal environment in which the organization seeks to achieve its objectives.

External Context - External environment in which the organization seeks to achieve its objectives.

Digital preservation – Consists on the long term maintenance of the accessibility of a digital object.

Authenticity - Ability to prove that the preserved digital object corresponds to the one produced by the original owner.

Heterogeneity - Being associated with several hardware or software providers, having different systems and components, being able to replace and add new ones whenever necessary.

Integrity – Guarantee that the informational content of a digital object has not been modified.

Provenance – Origin and history of the digital object, like its creator or the entity responsible for it.

Scalability - Ability to face technologic evolution and adjust to it.

Reliability – Ability to keep the digital object intact, accessible and authentic through time.

Trustworthiness – Ability to prove the information contained in a digital repository is worthy of trust, maintaining its authenticity, integrity, correctness, provenance, and overall quality.

Science – Global collaboration in certain fields or areas of research, the next generation infrastructure that enables it and a series of workflows for supporting the scientific pipelines and tasks.

Workflow – Means by which team members or scientists can model, design, execute, debug, re-configure and re-run their analysis and visualization pipelines, through a structured, repeatable and verifiable way, involving a series of steps, accessing large quantities of data and generate similar amounts of intermediate and final products.

Data Management (DM) – Activity which involves organizing, protecting, and sharing through actions such as data backups, cooperative work, version control, metadata management, data security, and archiving.

Data Management Plan (DMP) - Document that describes what data will be created, collected, stored, managed and disseminated during a project.

Engineering - Application of scientific principles to design or develop structures, machines, apparatus, or manufacturing (business) processes, or works utilizing them singly or in combination; or to construct or operate the same with full cognizance of their design; or to forecast their behaviour under specific operating conditions.

Appendix B - Risk Assessment Techniques Descriptions

These are the descriptions of the techniques from ISO 31010 (ISO IEC 31010, 2009) presented in Table 2.4.

Method Category	Technique	Description
Lock-up methods	Check-lists	Risk identification using previously developed lists, codes or standards
	PHA	Identification of hazards and events that harm activities, facilities or systems
Supporting methods	Structured Interview and brainstorming	Exchange of ideas through conversation with or without formal questions
	Delphi Technique	Combination of expert opinions to estimate risk likelihood and consequence
	SWIFT	Identification of risks by a team
	HRA	Used to evaluate human error influences on a system
Scenario analysis	Root cause analysis	Analyse of a losses, their causes and possible improvements identification
	Scenario analysis	Risk identification using future or imaginative scenarios
	Toxicological risk assessment	Identification and analysis of hazards
	Fault tree analysis	Identification of means of occurrence of an event through a logical tree diagram
	Event tree analysis	Inductive reasoning used to translate probabilities into possible outcomes
	Cause/consequence analysis	Combination of fault and event tree analysis allowing inclusion of time delays
	Cause-and-effect analysis	Identification of possible causes to an effect or risk
	Decision tree analysis	Representation of sequential decisions and results originated from an initial decision
	CBA	Analyse several options and chose the most profitable
Function analysis	FMEA/FMECA	Determine how components, systems or processes may not fulfil their objectives
	Reliability-centered maintenance	Identification of policies to manage failures
	HAZOP	Identification of deviations from the expected or intended performance
	HACCP	System for assuring product quality, reliability and safety of processes
Controls Assessment	LOPA	Evaluation of controls and their effectiveness
	Bow tie analysis	Describing and analysing risk's paths
Statistical methods	Markov analysis	Analysis of the current state of a system, from which its future state depends
	Bayesian analysis	Based on the prior distribution data to assess the likelihood of the results
	FN curves	Measure of the likelihood of events that cause damage to a population
	Risk indices	Risk indices are used to rate risks using similar criteria so that they can be compared
	Consequence/likelihood Matrix	Rank risks, sources of risk or risk treatments on the basis of the risk level

Table B.1: Risk Assessment technique's descriptions

Appendix C - Contributions for HoliRisk Development

In this annex, the most relevant contributions made for the improvement of HoliRisk, assuming a role of beta tester, are presented:

- Creation of the engineering and science model.
- Creation of the Metagenomics and LNEC model instances (contexts).
- Modification of the HoliRisk domain model. This was motivated by certain limitations identified in the tool during the application of the method to the case studies, These limitations happened due to certain aspects not considered in the domain model used (Barateiro, 2012). These aspects were, for example, the associations of several risks or other elements to a unique control, the fact that risks can also be considered events and vice versa and the association of a single risk to a certain asset, where several are in most cases required.
- Suggest a connection between the model, context and risk report sections of the tool, thus eliminating the issue of several separate access links being only a single login required for access to all three different components of HoliRisk.
- Giving of several small suggestions to the improvement of the HoliRisk (like the color system for the risk matrix, improvement of control and policy tables preview and format, especially for the policies and controls tables.
- Suggest the ordination of risk repository results by name.
- Identification of a large number of errors/bugs, in the context, model and reporting sections of the tool (for example, errors in the element creation formularies (problems in save/update content); errors in importing and exporting of results; error in the creation of a risk matrix, namely the colors definition.

There are other features that would bring value to the tool, whose absence limited the utilization of HoliRisk in this dissertation:

- Automatic updates of content between module and context sections. This would be useful so that each insertion or removal of a given element in the model, could be automatically expressed in the related contexts.
- Transfer of elements between different models or contexts.
- Better visualization of the data in risk reporter, with better resizing of tables and columns.
- Export of tables of elements like risks, assets, vulnerabilities and so on, to JPEG or PNG formats so data can be presented in a good viewing quality.
- Finalizing the base domain model change of the tool, so relations between elements can be more agile, for example, so we can treat an event also like a risk or vice versa, or to relate several assets to a given risk.

- Create a risk reporter specific for the engineering and science domain according to our proposal. This involves the following tasks:
    - Support the definition of RACI matrix.
    - Allow the application of SWIFT and HAZOP techniques, where the risk registry data is given as input and the corresponding results generated by the application of each technique are presented to the user.
    - Produce the tables from risk reporting.

Appendix D – MetaGen-FRAME project modules

This appendix describes the MetaGen-FRAME project modules that are presented in Figure 4.1 (Ferreira, et al., 2014).

Task	Description
Data quality control	Before a data set is processed, the information needs to respect certain quality thresholds. This step may be local or remote. The tool used is NGS QC Toolkit. The inputs are a text file with the sequences that are going to be analysed, a string with the format used by the previous file, a string detailing which sequence technology was used, and a variable to filter sequences by size. The output is a filtered version of the original data set, as well as statistics regarding the removed sequences
Analysis of taxonomy	This analysis determines the sample's microbial diversity, to determine the different organisms that are present and, if possible, their resolution levels (species, kingdom, etc). The tool used is MetaPhlAn, being a local task. The input is the filtered dataset produced previously, as well as a value which may represent a) the minimum percentage identity that a taxon (a group of one or more populations of organism(s)) needs to have to be considered valid; or b) the number of taxons to be returned as valid, in decreasing order of percentage identity. The output consists of several lists of organisms present in the sample, with respective resolutions and identity percentages
Remote web service	A sequence of web services that use the NCBI database. The web service sequence uses as an input the lists obtained in the former task and produces a set of corresponding NCBI IDs. Later in the web service sequence, the NCBI is consulted using the IDs and returns a list of sequences associated to the existing taxonomic results, in .fasta format
Alignment	Establishment of an order between the sequences by comparison with the sequences obtained previously. This step uses a parallel version of TAPyR mapper and is performed locally. It receives as an input the former list of sequences and generates as outputs a set of aligned sequences in .SAM format, a set of non-aligned sequences in a .fasta file, a set of aligned sequences also in a .fasta file
Functional annotation	The set of consensus sequences are submitted to a functional annotation procedure. This may be a local or remote task. It is composed of two steps, starting with a separate execution of the NCBI BLAST program and then feeding its results in .xml format to the default tool Blast2GO. It receives as an input the

Task	Description
	.fasta file with alignment sequences produced in the alignment task and produces image and texts identifying the main genes and components that were found to be associated to the aligned reads
De novo assembly	Sample identification by reconstruction. MetaVelvet is the default program. This task may run locally or remotely on a more powerful infrastructure. As an input, it receives the set of non-aligned sequences and as an output it returns contigs (junctions of several sequences)
Gene structure prediction	This is used to obtain information about the sample's genes and to find out if genetic structures are present. One tool that can execute this step is BG7. This is a local task. As an input, it receives the set of contigs generated in the de novo task and the output contains information regarding predicted genes in the following formats: .gff, .gbk, .tsv and .xml
Metabolic reconstruction	The aim was to produce results associated with the sample's metabolism. Due to technical constraints, this task was implemented implicitly by the result display from the Functional Annotation and Gene Structure Prediction steps

Table D.1: Description of MetaGen-FRAME project Modules

Appendix E – HAZOP and SWIFT Guidewords and Scenarios

In this appendix we present the guidewords used for HAZOP and the what-if scenarios used for SWIFT technique, being these elements presented for each case study. Examples of each utilization are also presented.

For MetaGen-FRAME project:

HAZOP Guidewords	Example
No or Not	Not all inputs are inserted; No data is returned by the NCBI;
More (Higher)	A higher percentage is given as initial input;
Less (Lower)	A lower percentage is given as initial input;
Other	Other tool/database is used;
All/Part of	All/Part of data is lost
Compatibility	Lank of compatibility between formats;
Down	NCBI is down;
Disabled	NCBI disabled;
Incomplete	Incomplete data or metadata;

Table E.1: MetaGen-FRAME HAZOP guidewords used

The following scenarios or questions were the ones used for the SWIFT technique for the MetaGen-FRAME project:

- What if inputs are wrongly inserted into the method?
- What if NCBI or other database used fails?
- What if NCBI, other database or local pc is hacked?
- What if local data or metadata is deleted?
- What if data or metadata grows exponentially?
- What if NCBI or other database facilities suffers a natural disaster?
- What if local PC is destroyed or robbed?
- What if data used is proprietary or human related (confidential/copyright)?
- What if financial support disappears?
- What if Metadata or data is corrupted?
- What if preservation laws and regulations change?
- What if certain stakeholders become less or non-involved?
- What if Software tools fail or become obsolete?
- What if hardware or network components fail or become obsolete?
- What if error occur in the access, search or delivery of data?

For LNEC case:

HAZOP Guidewords	Example
No or Not	No preservation services available;
More (Higher)	More volume of data generated;
Less (Lower)	Less governmental funding;
All/Part of	All/Part of data is lost
Compatibility	Lank of compatibility between formats;
Down	Hardware/software components are down;
Late	Data is synchronized later than required;
Ignored	Law changes ignored
Change	Preservation requirements change; Organizational structure change;
Disabled	Preservation system disabled;

Table E.2: LNEC HAZOP guidewords used

The following scenarios or questions were the ones used for the SWIFT technique for the LNEC case:

- What if databases fail?
- What if databases or other infrastructures are hacked?
- What if local data or metadata is deleted?
- What if data or metadata grows exponentially and the preservation solution infrastructure becomes insufficient?
- What if LNEC facilities suffer a natural disaster?
- What if data used is proprietary or human related (confidential/copyright)?
- What of human data is mixed with other kind of data?
- What if governmental support disappears?
- What if other financial supports disappear?
- What if Metadata or data is corrupted?
- What if certain stakeholders/partners become less or non-involved?
- What if workers with vital knowledge leave LNEC?
- What if software tools fail or become obsolete?
- What if hardware or network components fail or become obsolete?
- What if an error occurs in the access, search or delivery of data?
- What if internal LNEC services fail?
- What if the preservation community requirements change?
- What if preservation laws and regulations change?
- What if LNEC strategy plan is adjusted or changed?
- What if utilities fail?
- What if LNEC can’t fulfil its duties to its clients (dams)?