Data Governance in Engineering and Science Projects

Filipe Ferreira
INESC-ID, Instituto Superior Técnico,
Universidade Técnica de Lisboa,
Portugal
filipe.ferreira@ist.utl.pt

ABSTRACT

Engineering and science projects are increasingly data driven, meaning data sets need to be created, stored, disseminated and re-used for future use. This raises Data Management concerns and challenges. To address these concerns, Data Management Plans (DMP) are created. However, we claim Data Management Plans don't cover all important concerns. 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 (RMP), which intends to support all the decisions described in the DMP. The motivation requirements and evaluation of our work are represented by the MetaGen-FRAME project, and a case of a Civil Engineering Laboratory. 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

1. INTRODUCTION

Science and Engineering projects are increasingly data intensive, high collaborative and highly computational at large scale [3]. 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, re-used and preserved. This leads to an increase of DM concerns along with demands for efficient and effective DM practices and procedures [5].

The DM challenges in research projects are currently being addressed through Data Management Plans (DMP), increasingly demanded by the research funding agencies. In general, those agencies require a DMP to record how data is managed and disseminated during the project, including how it is created, collected and stored [5]. In this work we acknowledge those concerns and related principles, but we propose and demonstrate they are not sufficient for a complete coverage of all the relevant Data Governance (DG) concerns, namely concerning Risk Management (RM). In that sense, this work therefore proposes to complement the existing DMP with Risk Management Plans (RMP), for a more solid DG.

1.1 Problem

Even though DMP try to impose DM good practices, the current guidelines that help create this document mainly describe what should be included in the DMP, without justifying the declared decisions or what techniques supported that decision. Also, funding agencies state that the DMP will be assessed and will serve as criteria for grant acceptance. However, project stake-holders 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 [3]. Other aspects often overlooked in DM procedures and decision, 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, RM also intends to protect data, meaning DM 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. Since DMP don't address RM concerns, we believe these documents are not enough to address all DG concerns in engineering and science projects.

In summary, science and engineering projects need a more concrete guidance on how to analyze, identify, assess and control data management problems [3]. 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.

1.2 Document Structure

This document is composed as follows: In chapter 2 the related work is detailed, namely DM in engineering and science projects, DMP, RM, RMP and a bond 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 [2] and the LNEC case [14]. In chapter 5, some conclusions are presented, some lessons learned are detailed and suggestions for future work are given.

2. Related Work

This section shows the state-of-the-art for DM in engineering and science projects and DMP. It also covers RM and RMP related work. Finally a bond between RMP and DMP is shown.

2.1 Data Management in Engineering and Science Projects

DM is an integral part of engineering and science projects. It allows researchers to produce quality data and protect it from being lost or misused [4] [5]. DM concerns are increasingly perceived, namely data's provenance, sharing, access and archival [5]. As a result of these concerns, research funders have been increasingly requesting the creation of DMP. A typical DMP describes how data will be created, stored and shared, with two purposes [5]: (i) guide researchers to reuse data; (ii) record the project's DM decisions. Each project has specific purposes creating different instances of DMP. However, there are always common issues allowing the definition of generic DMP sections.

Table 2.1 compares four DMP guidelines, associated with two funding agencies, the Australian Nation University (ANU) and the National Science Foundation (NSF, from the guidelines for Engineering and Biology) and also the guidelines of the Digital Curation Center (DCC). These sections are detailed at [7] [8].

Table 2.1. Comparison of DMP Guidelines (present(x), not present (-))

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	-

2.2 Risk Management

RM identifies, assesses and mitigates risks to an acceptable level. It manages risks i.e. the uncertainty associated with events which can affect assets [1] [12].

A risk can be positive (opportunity), or negative in which case it may be exploited by events/threats.

Standards, methods and tools for RM vary with the market sector, type of business or organizational activities [13]. There are standards that focus on defining guidelines for specific domains and other standards, like ISO 31000 that define generic RM terminologies, processes, principles, methods and techniques.

ISO 31000 states that RM creates value from a structured, iterative and tailored way. The standard proposes a reference process to proper execute RM [6] [10].

The referred process proposes that firstly the context, strategic objectives and risk criteria of RM must be defined. The intermediate step is Risk Assessment, which is composed of three distinct phases: risk identification, which generates the list of risks; risk analysis, to consider the impacts and probabilities determining the risk level; and risk evaluation, to rank risks according to their risk level.

The final step is risk treatment, where controls are assigned to risks. Communication and review takes place throughout all stages [6] [10]. ISO 31010 presents risk assessment techniques that support the ISO 31000 process [11].

2.3 Risk Management Plan

A RMP defines the scope and process for the identification, assessment and treatment of risks. The objective of this plan is to define the strategy to manage risks of an organization or project with the minimal impact on cost and schedule, as well as operational performance.

The RMP is considered a living document, being updated as needed [6]. A typical RMP comprises the following steps [10] (i) scope definition, (ii) Risk Assessment, (iii) Risk Treatment, (iv) Risk Control and Monitor. A RM and RMP overview is given in Figure 2.1. Every term presented in Figure 2.1 is detailed in [1] and [12] .

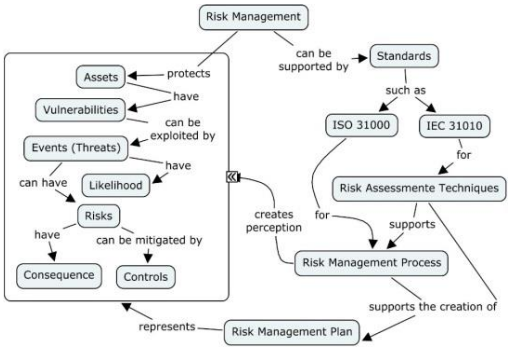

Figure 2.1. RM and RMP overview

2.4 A bond between DMP and RMP

Each research might have diverse purposes, resulting in different policies and instances of DMP. However, there are common issues allowing the definition of general DMP guidelines. From the analysis of different guidelines we defined a set of typical sections presented in a DMP (Table 2.1). We propose the complementation of DMP with corresponding RMP. To support this complementation, risks should be associated with the typical DMP sections presented in Table 2.1.

3. SOLUTION HYPOTHESIS AND PROPOSAL

This chapter reports to the definition of the objectives for a solution and the artifact development. 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, we also present some skills (required by the risk expert) and generic responsibilities/roles. A risk registry is also suggested, being useful for creating check-lists (presented in dissertation). All these elements combined represent our proposal for a DG approach in engineering and science projects.

3.1 Solution Hypothesis

To solve the presented problem, we try to explore the hypothesis of improving DM concerns in engineering and science projects by improving the DMP concept with RM guidelines, which means using a RMP to support the DMP. Our objectives are:

- **Propose 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 DG approach based on 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 developed, it's necessary to generalize DMP 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 the decisions, processes and controls implemented by the DMP**: The RMP, besides comprising all the process, results and techniques for RM analysis of a given case, 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 first objective.

3.2 Risk Management Method

The proposed method is illustrated in Figure 3.1 and intends to produce a RMP as final output. The structure of this method represents the structure of the RMP created. The process is based on ISO 31000 being suited for engineering and science projects.

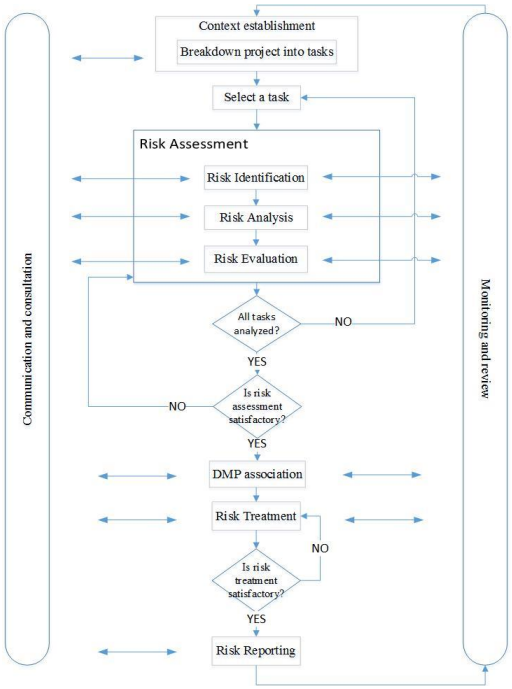

Figure 3.1. Method for RMP creation based on ISO 31000 guidelines

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 other elements that might affect DM. Internal context involves defining all the elements of the project, i.e. its objectives, resources, data, processes, systems, among others. 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 ranking system is suggested, being that decision made with the project experts. Risk assessment is performed using a set of techniques presented in section 3.3.

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. It is assumed the DMP in question already exists, and it must comply with the guidelines or requirements imposed by the funding agency (the creation of DMP and compliance issues are out of the scope of this dissertation).

The result of these steps eases the prioritization for risk treatment where controls are identified and associated to DMP sections.

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 to stakeholders 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.

3.3 Skills and Responsibilities

Despite the diversity in engineering and science projects we can identify generic roles that are present in any project. 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 [9] and other resulted from interactions with the interviewers:

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

Role	Description	Responsibility
Project Sponsor	Assumed by the funding agency	Informed of all risks and controls
Group Leader	Represented by the project PI (principal investigator). This role is more relevant for scientific projects	Responsible for risk and decision making, together with the project sponsor
Project Manager	Researcher that is responsible for coordinating the project	Defining the context and communicate risks. Accountable for all decision making
Risk Expert	Responsible for identifying, analyse, evaluate and treat risks	Should be consulted in all steps of the risk management process
Risk Owner	Person that controls and monitors specific risks	Informed of all decisions regarding that risk, and communicate issues to project manager
Operational/Scientific Team	Persons in charge of executing the project	Responsible for control implementation
DM expert	Person in charge of executing the project's DMP	Create the DMP and assist in the association of its sections with RM results

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. The following skills are 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, analyze, evaluate and treat any risk surrounding data;
- **Engineering or science field focus**: Knowledge of the field in question is mandatory;

3.4 HoliRisk: a Risk Management Tool

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

4. Demonstration and Evaluation

In this chapter we demonstrate and evaluate the proposed method with two different cases.

4.1 Case 1 – Metagenomics (MetaGen-FRAME Project)

This section shows the application of the proposed method and corresponding RMP creation on the MetaGen-FRAME project [2].

4.1.1 Context Establishment

The MetaGen-FRAME 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 [7].

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

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.

4.1.2 Risk Assessment

For risk assessment, we opted to use the HAZOP and SWIFT techniques. 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 typical assets, events, vulnerabilities and risks. Certain guidewords and what if scenarios characteristic from the HAZOP and SWIFT techniques also helped with risk assessment. All the results are further detailed at [7].

The project assets are presented in Figure 4.1. Vulnerabilities are presented in Figure 4.2. Events are presented in Table 4.1 Risks are presented in Table 4.2, with their respective consequence and severity values, as well as their trigger events and related assets. For result presentation, not all results are displayed, being the remaining in the dissertation. For calculation of risk severity, the following formula was used:

$$Severity_{risk} = Consequence_{risk} * Likelihood_{event}$$

#	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 4.1. 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 4.2. MetaGen-FRAME vulnerabilities

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

[2] http://timbusproject.net/

Table 4.1. MetaGen-FRAME events

ID	Events	Likelihood
E1	Creation and utilization of new technologies that increase substantially the quantity of data and metadata generated	0,5
E4	Errors on search, access and delivery of preserved data	0,5
E5	Financial, legislative or organizational changes	0,3
E7	Hardware obsolesce	0,3
E8	Human errors	0,7
E12	Non successful extrapolation of the data, metadata and documentation's meaning	0,5
E15	Software failure	0,5

Table 4.2. MetaGen-FRAME risks. Consequence values are presented by (C). The affected assets (A), trigger events (E) and severity values (S) are also presented

ID	Risks	C	E	A	S
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
R9	Lack of financial requirements	7	E5	A3, A7	3.5
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
R14	Loss of data/metadata/documentation due to human errors	9	E8	A2	6.3
R17	Loss of data/metadata/documentation due to software failure	9	E15	A2	4.5
R19	Loss of metadata denying the representation of the output to the user via Taverna	5	E12	A2	2.5
R30	Unavailability of storage capacity to respond to data growth	9	E1	A2	4.5

For each risk we identify the event that could trigger it. As visible on the table, 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.

For risk evaluation we used a consequence/likelihood matrix generated by the HoliRisk and reproduced on Figure 4.3. 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 severity where 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 input values from the 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 be the first treated.

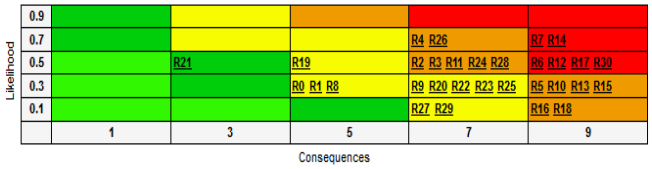

Figure 4.3. MetaGen-FRAME risk matrix

4.1.3 DMP Association

After validating risk assessment results, risks were associated with typical DMP sections. The association of the risks with the DMP sections is presented in Table 4.3, where we can see data quality assurance section has the risks with higher severity. The average severity was calculated using the values in Table 4.2.

Table 4.3. MetaGen-FRAME risk results association with DMP sections

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

This association supports the definition of a DMP by easing the identification of the risks that need to be covered in each section. 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).

4.1.4 Risk Treatment

For risk treatment, controls were identified using the SWIFT technique. A workshop was conducted were risks were analyzed in order to identify controls. Controls were 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 4.4. From this set, two policies were defined (see Figure 4.4), 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. Each control can reduce the consequence of a risk, decrease the likelihood of events or reduce vulnerability's exposure.

Table 4.4. MetaGen-FRAME controls with the respective type and entities it mitigates. For type (T), L = Likelihood, C = Consequence, E = Exposure

ID	Controls	T	Entities
C1	Backup system	C	R10, R11, R12, R13, R14, R15, R16, R17, R18, R19, R22, R23, R24, R25, R26, R27, R28, R29
C2	Create a long term storage policy (recovery management Plan)	C, E	V6, R9, R12, R13, R14, R15, R16, R17, R18, R22, R23, R24, R25, R26, R27, R28, R29, R30
C3	Create a protocol defining the workflow execution properties or create additional metadata, creating stronger bonds between the biological results	C	R2, R3, R4, R7, R19
C5	Emergency budget for issues in the NCBI	L, E,C	E11, V2, R9
C11	Modify formats used by the framework so each output references the associated input data (RDF style). interconnecting data elements	C	R4, R5, R6

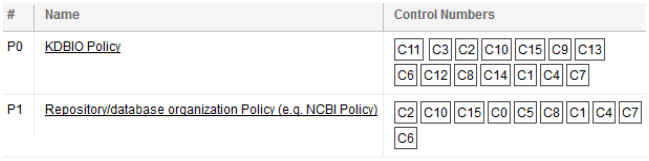

Figure 4.4. MetaGen-FRAME policies

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.

Table 4.5. MetaGen-FRAME control results association with DMP sections

Sections	Controls
Ethics and privacy	C4
Resourcing (Budget)	C2, C5
Legal Requirements	C7
Access and Sharing	C10, C9, C11, 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

4.1.5 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. As an example of these conclusions, in Table 4.6 we grouped risks into categories that were defined on collaboration with the consulted experts (in this case Miguel Coimbra). Each risk category has the corresponding average severity. Risks can be divided into categories or according to organization departments, if it becomes a more useful division for decision making. Through Table 4.6 it is possible to see the severity of a certain category of risks or department, improving decision making. This view complements the one given by the risk matrix (see Figure 4.3) where the same can be viewed concerning individual risks.

Table 4.6. Risk categories, average risk levels and priorities

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

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.

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 [14].

4.2.1 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. 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.

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. This project can be seen as three major tasks:

- **Data upload**: the data captured by sensors 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;

Typical roles and responsibilities were assigned to certain entities, which was performed together with LNEC 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.

4.2.2 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 a list of typical assets, events, vulnerabilities and risks.

Certain guidewords and what-if scenarios characteristic of HAZOP and SWIFT techniques also helped with risk assessment. All criteria (for likelihood, exposure and consequence) used during risk assessment was defined in collaboration with LNEC expert José Barateiro. For calculating the risk severity values, the following equation was used:

$$Severity_{risk} = 2 * Consequence_{risk} + Likelihood_{event}$$

This case's assets are identified in Figure 4.5. The vulnerabilities found are presented in Table 4.7. In Table 4.8 the events of this case are displayed. Table 4.9 shows the case's risks. For each risk we identify its consequence, the event that could trigger it, the affected assets and the severity value.

As visible on the Table 4.9, 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. For result presentation, not all results are displayed, being the remaining in the dissertation.

#	Name	Value	Vulnerabilities Number
A0	Business processes	very-high	V0
A1	Data / metadata	very-high	V8 V10 V6 V17 V18 V4 V16 V5 V2 V15 V3 V14 V12
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 4.5. LNEC assets with values and vulnerabilities

Table 4.7. LNEC Vulnerabilities

ID	Vulnerability	Exposure
V0	Business processes 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

Table 4.8. LNEC events

ID	Events	Likelihood
E0	Software tool can't be used or accessed	4
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
E8	Hardware failure	3
E9	Financial, legislative or organizational changes	4
E21	Hacker attack	3

Table 4.9. LNEC risks with the respective trigger events (E), assets (A), consequence values (C) and severities (S)

ID	Risks	C	E	A	S
R3	Lack of Technical Support for Preservation Solution due to budget cuts	3	E2	A0, A3	10
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 attacks	6	E21	A0, A1, A4	15
R12	Loss of data/metadata/documentation due to hardware failure	6	E8	A0, A1, A3, A4	15
R16	Loss of data/metadata/documentation due to software failure	6	E0	A0, A1, A4, A6	16
R18	Loss of data/metadata/documentation for inappropriate use of these assets	6	E4	A0, A1	16
R46	Unavailable storage capacity concerning data growth	6	E3	A1, A2, A3, A4	16

For risk evaluation we used a consequence/likelihood matrix generated by the HoliRisk and reproduced on Figure 4.6. With this matrix we 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.

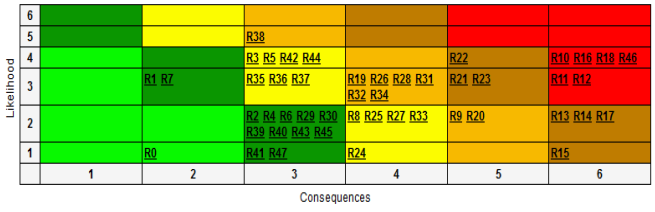

Figure 4.6. LNEC risk matrix

4.2.3 DMP Association

After validation of the risk assessment results, risks were associated with typical DMP sections which is presented in Table 4.10. Resourcing, archiving and preservation and data formats and metadata are the DMP sections with risks with higher severity.

Table 4.10. LNEC risk results association with DMP sections

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

4.2.4 Risk Treatment

For risk treatment, controls were identified using the SWIFT technique. 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 could apply them. The final proposed set of controls is presented in Table 4.11, 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 4.7), Some controls are shared in more than one policy, meaning they should be implemented by more than one entity.

Table 4.11. LNEC controls, with the entities they mitigate. For type (T), L = Likelihood, C = Consequence, E = Exposure

ID	Controls	T	Entities
C1	Create/improve recovery management Plan	E, C	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
C5	Enforce documentation of systems/processes	C	R8, R11, R13, R18, R19, R21
C15	Usage of fall back data bases	C	R10, R11, R12, R13, R14, R15, R16, R17, R18

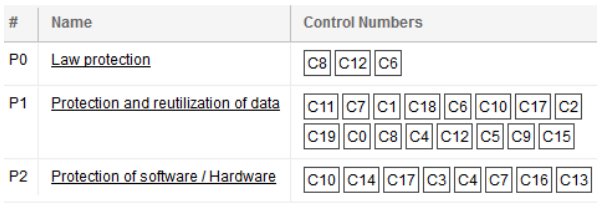

Figure 4.7. LNEC policies

Finally control measures are associated with DMP typical sections (see Table 4.12). This association helps justify the application of control measures in each DMP section.

Table 4.12. LNEC controls association with DMP sections

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, C9, C17, C18
Data Quality Assurance	C0, C3, C9

4.2.5 Risk Reporting

As a final step for the creation of the RMP, more conclusions are drawn up from the previous results, being useful for decision making and structured for optimizing risk data communication to stakeholders. As an example, in Table 4.13 we grouped risks into categories that were defined on collaboration with the consulted experts (in this case José Barateiro). Risks can be divided into categories or according to organization departments, if it becomes more useful for decision making. This view complements the one given by the risk matrix (see Figure 4.6).

Table 4.13. Risk categories, average risk levels and priorities

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

4.3 Analysis of Two Approaches

Since a RM analysis was made for the LNEC case in TIMBUS project [14], 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 risk assessment techniques, namely brainstorm, SWIFT, check-lists and likelihood/consequence risk matrix. Although, in the TIMBUS analysis, the HAZOP technique wasn't used for risk assessment. The proposed method, by using both SWIFT and HAZOP, finds more RM results related to the systems and business processes of LNEC, namely risks, events or vulnerabilities that SWIFT alone couldn't 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.

- TIMBUS analysis only used the risk matrix as reporting tool for stakeholders. It doesn't supply friendly reporting aids for high executives, who would be more interested in finding details concerning a certain category of risks, or in finding details concerning risks related to a department of LNEC.
- On the TIMBUS analysis, only three risk types were identified, namely strategic, operational and legal. In the proposed method, more categories were considered. This allows a clear knowledge of the risk types related to LNEC, as well as the control types that are needed. The category definition in the proposed method, supported by risk expert José Barateiro, was inspired in the study of DMP sections.
- In the proposed method, there was a detection of irrelevant risks due to the association of risks with DMP sections, thus identifying the risk relevancy for DM. A risk that doesn't suit any DMP section is out of DM context DM and is discarded;
- Having applied the method to the MetaGen-FRAME case first, more complete check-lists were developed, meaning a better starting point for the RM analysis.
- A RACI matrix is detailed in the proposed method and is used to support scope definition by defining generic roles and responsibilities detailed for each task that needs to be performed. TIMBUS analysis doesn't use such technique.
- LNEC has a specific DM concern, namely digital preservation (DP) and so, TIMBUS analysis is more oriented for that concern. The proposed method addresses generic DM concerns, 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 that DM and RM share objectives and have similar concerns. Also, stakeholders have difficulty identifying DM problems and justifying the respective solutions. 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 the decisions and controls defined in each DMP section. This method is based on ISO 31000 guidelines and ISO 31010 techniques, being specifically designed for engineering and science projects. The method is applied in two scenarios, the MetaGen-FRAME project and the LNEC case.

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 improves DG for projects belonging to the domain areas of this dissertation.

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 in 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 the creation of 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 Main Contributions

Our main contributions were the following:

- **DG approach proposed for engineering and science projects**: 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 promotes a more solid DG.
- **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 are justified through the detection of risks, threats, vulnerabilities controls and polices.

- **Development of a generic risk management method for the creation of risk management plans for 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. For this we suggest a set of skills that we recommend any risk expert should master.
- **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.
- **Testing and improving a risk management tool – HoliRisk**: During this dissertation, we assumed the role of beta tester of HoliRisk. This tool is currently being developed and our work contributed to its development through the detection of several bugs and other conceptual errors, as well as the indication of several suggestions.

5.3 Future Work

HoliRisk could be used to fully support the proposed method. At this point only supports the risk registry functionality and some basic requirements for risk reporting. A specific risk reporter for the domain of engineering and science still needs to be developed.

During this dissertation, deliverable 8.2 from TIMBUS analysis on LNEC was analyzed. Deliverable 8.4 could also be analyzed as a future scenario.

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 and DMPOnline, 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 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 proposed method 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.

6. REFERENCES

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

[2] Coimbra, M. 2013. "MeatGen-FRAME: Metagenomics Data Analysis Framework Focused on Stressed Microbial Communities". Universidade Técnica de Lisboa, Instituto Superior Técnico.

[3] Crowston, K, Qin, J. 2011. "A Capability Maturity Model for Scientific Data". Proceedings of the American Society for Information Science and Technology, 10-19

[4] Darlington, M., Ball, A., Howard, T., Culley, S., & McMahon, C. 2010. "Principles for Engineering Research Data Management". University of Bath.

[5] Fernandes, D., Bakhshandeh, M., Borbinha, J. 2012. "Survey of data management plans in the scope of scientific research". TIMBUS Timeless Business. INESC-ID.

[6] Ferreira, F., Coimbra, R., Bairrão, M., Vieira, R., Freitas, A. T., Russo, L., N., S., Borbinha, J. 2014. "Data Management in Metagenomics: A Risk Management Approach". In IDCC Conference.

[7] Ferreira, F., Coimbra, M., Vieira, R., Bairrão, R., Freitas, A. T., Russo, L., N., S., Borbinha, J. 2013. "A Risk Management Plan in Metagenomics". Technical Report. INESC-ID.

[8] Ferreira, F., Coimbra, M., Vieira, R., Proença, D., Freitas, A. T., Russo, L., N., S., Borbinha, J. 2013. "Risk aware Data Management in Metagenomics". In Inforum Conference.

[9] Hillson, D., & Simon, P. 2007. "Risk Management Plan template". ATOMrisk.

[10] ISO 2009. Risk Management - Principles and guidelines. ISO FDIS 31000 Geneva, Switzerland.

[11] ISO 2009. Risk Management – Risk assessment techniques. ISO IEC 31010 Geneva, Switzerland.

[12] ISO 2009. Risk Management – Vocabulary. ISO Guide: 73 Geneva, Switzerland.

[13] Ramirez, D. 2008. "Risk Management Standards: The Bigger Picture". Information Systems Control Journal, Vol. 4.

[14] Redlich, D., Molka, T., Gilani, W., Barateiro, J., Miranda, P., Lucas, A., Kolany, B., Yankova, S., Hecheltjen, M., Vieria, R., Borbinha, J., Nolan, M. Trezentos, P., Simon, F. 2012. "Deliverable 8.2: Use Case Specific Risks". TIMBUS Timeless Business. INESC-ID