inesc id lisboa

Technical Report

A RISK MANAGEMENT PLAN IN METAGENOMICS

Version *1.0*

28/10/2013

This work was supported by national funds through FCT (Fundação para a Ciência e a Tecnologia), under projects PEst-OE/EEI/LA0021/2013, TAGS PTDC/EIA-EIA/112283/2009, and DataStorm EXCL/EEI-ESS/0257/2012, and by the project TIMBUS, co-funded by the European Commission under the 7th Framework Programme for research and technological development and demonstration activities (FP7/2007-2013) under grant agreement no. 269940.

CONTRIBUTORS

Name	Organization	Email
Filipe Ferreira	INESC-ID	filipe.ferreira@ist.utl.pt
Miguel Coimbra	INESC-ID	miguel.e.coimbra@ist.ult.pt
Raquel Bairrão	INESC-ID	raquel.bairrao@ist.utl.pt
Ricardo Vieira	INESC-ID	rjcv@ist.utl.pt
Ana T. Freitas	INESC-ID	atf@inesc-id.pt
Luís Russo	INESC-ID	lsr@kdbio.inesc-id.pt
José Borbinha	INESC-ID	jlb@ist.ult.pt

VERSION HISTORY

Version #	Implemented By	Revision Date	Approved By	Approval Date	Reason
1.0	Filipe Ferreira	28/10/2013	Miguel Coimbra	28/10/2013	Initial Risk Management Plan draft

TABLE OF CONTENTS

1 INTRODUCTION

Data Management (DM) is an activity, which involves actions such as 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 [9]. One of the main concerns of DM is Digital Preservation (DP), of the vast data sets used.

Within the DM and DP concerns, the concept of Data Management Plan (DMP) was developed. DMP represents the set of rules and good practices a project must follow in what concerns data, according with the objectives of stakeholders (usually, a funding organization).

Risks and challenges, namely in the data and workflows used, are increasingly emerging in e-Science projects. This report presents a solution for the previous risks and challenges, by presenting a Risk Management Plan (RMP) for the MetaGen-FRAME project [1]. This e-Science case study belongs to the field of Metagenomic, focused on sequence analysis and genome annotation. The method used for RMP creation is based on three distinct phases, namely the phase one, where the RMP's context is defined, phase two, where the planning is made and finally, phase three, where the proceeding are detailed. This process is based on ISO 31000 [2] good practices and a set of typical DMP sections [10].

2 PHASE ONE – CONTEXT

2.1 PROJECT DESCRIPTION

The MetaGenFRAME project [1] is focused in the study of relatively-controlled environments (possibly composed by several types of different bacteria, with each type being present in different quantities), whose chemical reactions may be influenced and enhanced. The project is therefore focused on the study of bacterias, also called prokaryotes. The origin of a metagenome typically consists of a closed environment (for example, a sample containing human gut bacteria) or an open environment (like the open ocean). The tools used for task execution are pre-selected via script.

The project's main tasks are shown in Fig. 1. in more detail.

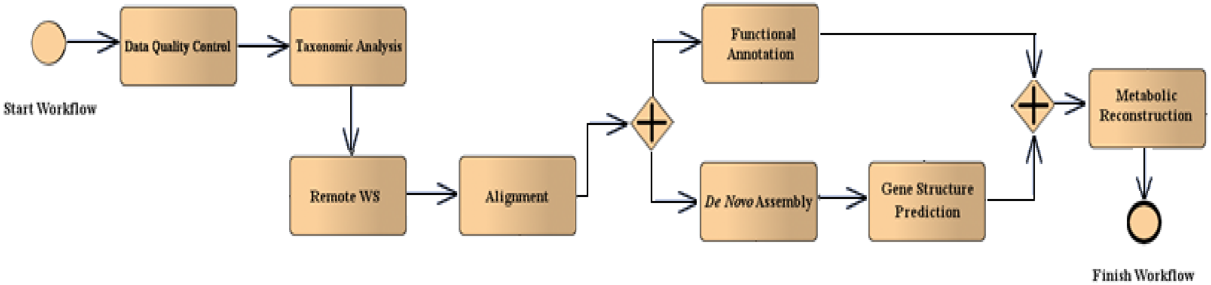

Fig 1 - The MetaGen-FRAME workflow.

Table 1 describes each task in more detail:

Table 1. List of MetaGen-FRAME tasks.

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, although remote execution on a more powerful computing infraestructure is recommended and it's executed using NGS QC Toolkit[1]. The inputs are a text file with the sequences that are going to be analyzed, 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	Determine 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[2], being a local task. The input is the filtered data set 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; b) the number of taxons to be returned as valid, in tdecreasing order of percentage identiy and the output consists of several lists of organisms present in the sample, with respective resolutions and identity percentages (converted to query format for usage in the WS invocation sequence)
Remote WS	A sequence of WS that use the National Center for Biological Information (NCBI) data base. The WS sequence uses as input the lists obtained in the former task and produces a set of corresponding NCBI IDs. Later in the WS sequence, the NCBI is consulted using the IDs returning 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[3] mapper and is performed locally. It receives as 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. It may be a local or remote task. It is composed of two steps, starting with a separete execution of the NCBI BLAST program and then feeding its results in .xml format to the default tool Blast2GO[4]. It receives as input the .fasta file with aligment sequences produced in the Aligment 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[5] is the default program. This task may be run locally or remotely on a more powerful infrastructure. As input, it receives the set of non-aligned sequences and as output it returns contigs (junctions of several sequences)

[1] NGS QC Toolkit: http://59.163.192.90:8080/ngsqctoolkit/

[2] The Huttenhower Lab: http://huttenhower.sph.harvard.edu/metaphlan/

[3] TAPyR - Tool for Alignment of Pyrosequencing Reads :.: http://www.tapyr.net/

[4] Blast2GO: http://www.blast2go.com/b2ghome

[5] MetaVelvet: a short read assembler for metagenomics: http://metavelvet.dna.bio.keio.ac.jp/

Task	Description
Gene structure prevision	Used to obtain information about the sample's genes and to find if genetic structures are present. One tool that can execute this step is BG7[6]. It's a local task. As 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[7], .gbk[8], .tsv and .xml
Metabolic Reconstruction	One of the aims was to produce results associated with the sample's metabolism. Due to technical constraints, this task was implemented implicitly by the result display of Functional Annotation and Gene Structure Prediction

All tasks have a log file saved locally, and the view in which each task presents its results (using Taverna[9]) is customizable to the operator's needs.

2.2 PURPOSE OF THE RISK MANAGEMENT PLAN

A risk is an event or condition that, if it occurs, could have a positive or negative effect on a project's objectives. Risk Management is the process of identifying, assessing, responding to, monitoring, and reporting risks. This Risk Management Plan defines how risks associated with the MetaGen-FRAME project will be identified, analyzed, and managed. It outlines how risk management activities will be performed, recorded, and monitored throughout the lifecycle of the project and provides templates and practices for recording and prioritizing risks.

The intended audience of this document is the project team, project sponsor and management. This document also intends to complement a DMP related to the same use case, which is used to perform data management (DM) on the generated data, having as main objective the Digital Preservation (DP) of that same data.

2.3 AUTHORITY

The MetaGen-FRAME project was founded by FCT (Fundação para a Ciência e a Tecnologia). No official RM authority is involved in the RM analysis of this project.

2.4 SCOPE

The MetaGen-FRAME project risk management process aims to manage all foreseeable risks (both opportunities and threats) in a manner which is proactive, effective and appropriate, in order to maximise the likelihood of the project achieving its objectives, while maintaining risk exposure at an acceptable level. Due to the project mentioned

[6] BG7 – bacterial genome annotation system: http://bg7.ohnosequences.com/

[7] The Sequence Ontology Project: http://www.sequenceontology.org/gff3.shtml

[8] Sample GenBank Record: http://www.ncbi.nlm.nih.gov/Sitemap/samplerecord.html

[9] Taverna: http://www.taverna.org.uk/

above, the specific field of e-Science in where this RMP is developed is the area of Metagenomics.

3 PHASE TWO - PLANNING

3.1 STRATEGY AND APPROACH

The risk handling strategies that are going to be applied in this project are:

- **Likelihood control**: taking actions that reduce the likelihood of a given risk;
- **Risk sharing**: sharing or transferring the risks to other entities through contracts, finance or insurance;
- **Consequence control**: taking actions to reduce the consequence of a risk;
- **Exposure control**: taking actions to reduce the exposure of a vulnerability associated with a risk.

These strategies are applied in section 4.3.

3.2 ORGANIZATIONS AND RESPONSABILITIES

The several stakeholders involved and their respective responsibilities are expressed in the RACI chart expressed in table 2:

Table 2. RACI chart – (R = Responsibility, A = Accountable, C = Consulted, I = Informed)

Tasks/Positions	Project Sponsor	Project Manager	Risk Manager	Risk Owner
Taking decisions on project strategy	R	I		
Insurance of adequate resources for RM	R	I		
Definition of the acceptable levels of risks	C	R	I	
Risk Management Plan acceptance	I	R	C	
Control's efficiency and effectiveness monitorization	I	R	C	A
Risk control plans acceptance	I	R	C	
Overseeing and managing the risk management process		I	R	A
Preparation of the Risk Management Plan		I	R	A
Development of risk controls		I	C	R
Monitoring the progress of risk controls		I	C	R

The role of project sponsor is shared between Miguel Coimbra, Ana Teresa Freitas and Luís Russo. The role of Project Manager is performed by Miguel Coimbra. The role of Risk Manager and Risk Owner is performed by Filipe Ferreira.

3.3 TECHNIQUES

To perform the risk assessment phase, following set of techniques [3], the subset of used techniques and their were used (also in table 3):

- For Risk Identification:
  - **Check-lists**: E-Science has a list of well-known risks like any other area. These standard risks and challenges provide a good starting point to the identification of risks; - allows the identification of the first risk set for the scenario at hand;
  - **Brainstorming**: In e-Science, like in any other area, there is space for imagination in what concerns finding risks. Starting with the risks given by the check-lists technique, it's possible to find new risks regarding the scenario at hand in a systematic manner, through the gathering of several project stakeholders, like the research team, where ideas and thoughts are shared and discussed; - As check-lists don't find new risks, brainstorm is useful to find new risks;
  - **Structured "What-if" Technique (SWIFT)**: Identify potential risks arising when change is eminent; - The use case is based on procedures and systems (local and remote) which are the main applications for this technique, and so this technique becomes useful in finding new and specific risks associated with the procedures and systems used. E-Science is based in scientific workflows that are composed by a set of systems and tools, used in a number of procedures. These systems, tools and the respective procedures are susceptible to change. These changes can have diverse effects on the performance of a system and procedure, causing potentially positive or negative changes in the main workflow. These facts make SWIFT a suitable technique to use in an e-Science scenario;
  - **Failure modes and effects analysis (FMEA) and failure modes and effects and criticality analysis (FMECA)**: Identification of the ways in which components, systems or processes may not fulfill their design objectives; - The e-Science use cases are based on procedures and systems (local and remote), being their sequence represented through the main workflow. These systems and procedures represent the focus of this technique, and so this technique becomes useful in finding failure modes specific to the procedures and systems used, allowing the detection of risks in e-Science projects;
  - **Human reliability assessment (HRA)**: To assess possible human errors; - If an e-Science project is not fully automated, there is a dependency of an human operator, which can happen in any task of a scientific workflow, for example in the beginning of the workflow (to prepare the input files). This raises risks of the human nature, requiring a technique that can identify possible human errors that can compromise the workflow and the results;

- Risk Analysis:
    - Decision tree analysis: When decisions are needed. Estimates, for each path coming from a certain decision/event, the value/cost of its outcome, providing means to choose the best from the available set of options; - As it has already been stated earlier, in e-Science scenarios are based on scientific workflows, that depend on a large number of systems and tools that support a set of procedures, generating large quantities of data. Procedures may be executed using different sets of tools and systems. It becomes vital to decide, which tools or systems are going to be used. In order to take that decision, the costs and value pertaining each of the several possibilities needs to be calculated. Since this is the purpose of this technique, for every decision that needs to be made in an e-Science case, the usage of this technique becomes very helpful. In this method, this technique is also intended to be used to analyze the risks that were identified by techniques that can't analyze their identified risks;
    - Structured "What-if" Technique (SWIFT): This technique is bases upon meeting involving the several stakeholders, much like brainstorm, and during these meetings the values of probability and consequence are discussed and agreed by all the involved parties, leading to the determination of the respective risk levels. From this procedure results a list of ranked risks;
    - Failure modes and effects analysis (FMEA) and failure modes and effects and criticality analysis (FMECA): For each risk it identified, calculates the risk's criticality (level of risk), in order to prioritize the same risks;
    - Human reliability assessment (HRA): Calculates the probabilities and possible consequences of the risks associated with human errors identified in the previous step.

- Risk Evaluation:
    - Consequence/probability matrix: Support method to help decide which risks need to be treated and the ones that do not. It also gives a visualization of the risk evaluation;

Table 3. Proposal Techniques used in the MetaGen-FRAME RMP (RI – risk identification, RA – risk analysis, RE – risk evaluation).

Technique	Risk Identification	Risk Analysis	Risk Evaluation
Check-lists	X		
Brainstorming	X		
SWIFT	X	X	
FMEA/FMECA	X	X	
HRA	X	X	
Decision tree analysis		X	
Risk matrix			X

4 PHASE THREE – PROCEEDINGS

4.1 ASSET, EVENT AND VULNERABILITY'S IDENTIFICATION

The assets that need protection, in what concerns the MetaGen-Frame project are:

- **A1** - Data (including the metadata, and documentation);
- **A2** - Tools (Taverna, Blast2GO, NGS QC Toolkit, BG7, MetaPhlAn, TAPyR);
- **A3** - Computational servers;
- **A4** - Data bases (NCBI);
- **A5** - Local PC;
- **A6** - WS;

The vulnerabilities that are associated with the former assets are:

- **V1** - Unreliable storage hard drive in the local PC. PCs and external HD are useful for short term storage, but inadequate for long term storage due to high failure rate;
- **V2** - Security breaches in the Local PC, as well as, in the NCBI and computational servers, since these servers and data bases can be configured by agents with formation on bioinformatic, lacking the necessery formation in security;
- **V3** - Poor debug capabilities of Taverna. In the case of failure, it's problematic if the SWMS doesn't provide debugging capabilities that show the failure cause. For example, if the WS fails and no information is provided the user is left wondering whether the service failed locally or remotely;
- **V4** - Lack of syntactic and semantic verification mecanisms to check the initial inputs given by the human operator;
- **V5** - Lack of a long storage policy;
- **V6** - Communication channel overload, leading to a slow or non existing connection;
- **V7** - Economic or organizational breakdowns can also influence the organization running the NCBI, causing its termination;
- **V8** – Lack of a criteria set, defining if a certain data set is confidential or not;

The events that can exploit the given vulnerabilities are:

- **E1** - Local media units, data bases, WS, computational servers or communication failures;
- **E2** - Media units, data bases or computational servers maintenance;
- **E3** - Hacker attacks to the infrastructures or communication channels;
- **E4** - Natural disasters (fires, floods, earthquakes);
- **E5** - Insertion of wrong input values by the human operator;
- **E6** - Tool discontinuation and lack of support;

- **E7** - Financial, legislative or organizational changes in the organization running the data bases used, leading to changes on the policies surrounding the data preservation;
- **E8** – Sharing of information without consent;
- **E9** – Project's abandonment from a stakeholder;

4.2 RISK ASSESSMENT

4.2.1 Risk Identification

The MetaGenFRAME project extrapolates several types of information from the data set it receives as input, such as the composition of the organism community present in the sample. It also aims to produce information pertaining the metabolism and main chemical reactions. With a particular focus on prokaryotic organisms, this raises important issues associated with the secrecy and storage of data, as it will potentially convey information that is important to the client or entity's activity. An example of such an activity is the process of analyzing and enhancing biomass decomposition, fuel refinement, crude extraction, among others. Such processes are trade secrets, and their study must undertake the precautions mentioned earlier. The project also uses remote web services, so ensuring that the information and services available remotely will remain active is a key-necessity for biologists and other professionals. The identified risks are organized by categories of risks. For each risk, the assets, vulnerabilities and events directily associated are also proesented. The identified risks are presented in table 4:

Table 4. Identified risks, with the respective assets, vulnerabilities and events

Category of Risk:	Risk	Assets	Vulnerabilities	Events
Human errors	R1 - Accidental alteration or deletion of digital objects;	A1	V4	E5
	R2 - Insertion of wrong input values: One example is the introduction of the wrong value in variables that indicate the percentage of a sequence's nucleotides that must be of quality regarding the total length of the sequences which are filtered in the data quality control task, therefore influencing all the following results;	A1	V4	E5
Intentional (internal or external) attacks	R3 - Alteration of the external WS, NCBI, computational servers or local PC used causing their unavailability or failure;	A3, A4, A5, A6	V1, V2	E1, E2, E3, E4
	R4 - Loss of information due to communication failures;	A1	V6	E3

Category of Risk:	Risk	Assets	Vulnerabilities	Events
Loss of data	R5 - Loss of information and data traceability due to a media fault, compromising the workflow's recreation, with the same inputs;	A1, A2	V1, V7	E1, E2, E7
	R6 - Loss of metadata denying the representation of the output information to the user via Taverna;	A1	V1, V7	E1, E2, E3, E4, E7
	R7 – Lack of financial or legal requirements to preserve data;	A1	V7	E7
Workflow execution failures	R8 - Obsolesce of the tools used in the workflow or in the NCBI or local PC;	A2		E6
	R9 – Ocurrence of an unexplicable error that cant be explaned;	A2	V3	E6
Data sharing and missuse of information	R10 - Sharing of confidential data;	A1	V8	E8
	R11 - Difficulties sharing the information and the workflow's execution in other future scenarios;	A1, A2	V8	E8
Stakeholders and data owners	R12 – Stakeholder's lack of involvment;	A1		E9

4.2.1.1 DMP Risk Allocation

As this document intents to complement a DMP referent to the same project, the risks presented above must be allocated according to the generic sections of the DMP, leading to the following distribution expressed in table 5:

Table 5. Relation between the typical sections of a DMP and the identified risks

	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	R11	R12
Data Storage, Preservation and security	X	X	X	X	X	X		X	X			
Ethics and privacy										X	X	
Data Formats and Metadata					X	X		X	X			
Products of Research/Documentation					X	X						
Resourcing (Budget)							X					
Data Dissemination/sharing and licensing										X	X	
Data owners, stakeholders and Responsibilities												X

4.2.2 Risk Analysis

In order to perform the risk analysis, and calculate the risk level, of every risk identified, probability and consequence criteria were defined in table 7.

Risk probabilities (P) and consequences (C) obtained through the maximum value of the probability and consequence values from the specific riks represented in the previous section. Risk levels are obtained through P X C. The values of P and C are expressed in table 6:

Table 6. Values of likelihood and consequence of each risk.

Risks	**R1**	**R2**	**R3**	**R4**	**R5**	**R6**	**R7**	**R8**	**R9**	**R10**	**R11**	**R12**
Likelihood (L) Consequence (C)	0.5 9	0.5 9	0.3 9	0.3 9	0.3 7	0.3 7	0.3 7	0.1 5	0.1 7	0.1 7	0.3 7	0.1 5
Risk level (LXC)	4.5	4.5	2.7	2.7	2.1	2.1	2.1	0.5	0.7	0.7	2.1	0.5

Table 7. Likelihood and Consequence criteria and respective values.

	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 necessery.	**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 necessery.	**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 necessery.	**5**
0.7	Likely risk due to the presence of several serius problems and deficiencies, which take a considerable time and effort to mitigate.	High	High chance of endangering the workflow. Large changes are necessery.	**7**
0.9	Extremly 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 necessery.	**9**

4.2.3 Risk Evaluation

For the evaluation of risks, a risk matrix was developed table 8. From the matrix we conclude that R1 and R2 are the risks with a very-high priority, being the first ones treated. R3 and R4 have a high priority beginning treatment after R1 and R2. The risks R5, R6, R7, R9, R10 and R11 have a medium priority being the last ones treated. R8 and R12 have a low priority and need only to be controlled;

Table 8. MetaGen-FRAME's risk matrix.

Likelihood	0.9					
	0.7					
	0.5					R1, R2
	0.3				R5, R6, R7, R11	R3, R4
	0.1			R8, R12	R9, R10	
		1	3	5	7	9
		Consequence				

4.3 RISK TREATMENT AND CONTROL

Risk control measures should be implemented from the beginning of Miguel Coimbra's Master Thesis until its conclusion. For each control measure, in what concerns schedule and costs, it was considered a monthly cost for Miguel's work of 750€. The remaining costs belong to detailed services or are estimations. Time consumptions were estimated. The risk control measures are presented in table in the appendix, including the respective estimates and strategies for each control. No training is necessary, so no costs and timeframe are associated. The association between the risk controls and the designated risks is presented in table 9.

Table 9. Controls necessary, the respective strategies, state of implementation (I. – implemented, N.I. – not implemented) and the respective estimates for time and cost.

Nº	Designation	Strategy	State	Cost / Duration
C1	Use several backup systems in the local PC (local and remote), for example a system like shadow copy to store all the data and metadata	Consequence control	N.I.	To implement the shadow copy mechanism it would be necessary 1 day, 100€ for the disk and 25€ for the work.
C2	Implementation of syntactic and semantic verification mechanisms of the given inputs, which would alert automatically the user if the input didn´t had the correct format and content	Exposure control	N.I.	The implementation of syntactic and semantic verification mechanisms would take a week with a cost of 190 €
C3	Improve the security measures from NCBI, computational servers, local PC and communication channels; - better antivirus, traffic encryption, firewall	Exposure control	I.	-
C4	Keep all the software components up to date	Likelihood control	I.	-
C5	Backup systems for the NCBI, WS and computational servers	Consequence control	I.	-
C6	Access to other genome data bases such as the ones offered by the EMBL-EBI[10] like ENA[11] or other custom made in case of unavailability of the NCBI – utilization of the remaining third-party or custom WS using them as fall back or as a set	Consequence control	N.I.	To modify the workflow (Taverna) for different data bases and computational servers it would take about two weeks and 380€
C7	Access to other computational servers in case of unavailability of the current servers (fall back)	Consequence control	N.I.	-
C8	Create a replicated central storage to store, in real time, the execution results from the workflow using for example shadow copy in the local PC and computational servers	Consequence control	N.I.	Cloud solutions could be used or shadow copy. This would take two days and 50€ plus 120€ from dropbox service.

[10] http://www.ebi.ac.uk/

[11] http://www.ebi.ac.uk/ena/data/view/CP006584

Nº	**Designation**	**Strategy**	**State**	**Cost / time**
C9	Create a long term storage policy with a speciallized organization	Likelihood control	N.I.	1 month and 5000€.
C10	Anti-fire and earthquake measures in the NCBI and computational servers	Likelihood control	I.	-
C11	Emergency budget for financial changes in the NCBI organization or in case of abandonment of any project member	Consequence control	I.	-
C12	Insertion of alternative tools in the main workflow, to function in case of failure of the main tools used	Consequence control	N.I.	6 weeks and 1125€
C13	Keep all the software and hardware components up-to-date	Likelihood control	I.	-
C14	Usage of open-source tools and formats avoiding possible obsolesce of tools or formats of data	Likelihood control	I.	-
C15	Insertion of a new component in Taverna that would check the return value, and if this would be an error, a new tool would treat the error so the workflow wouldn't stop. The error would be wrote in a log, so the origin could be traced	Exposure control	N.I.	4 days and 100€
C16	Creation of alternative forms of documentation, for example physical documentation then digitalised and stored in a backup system	Consequence control	N.I.	1 month spread throughout the project is necessary and it would cost about 750€.
C17	Modify the formats used by the framework so that each output references the associated input data and output data (RDF style). This would lead to more interconnection between data elements	Exposure control	N.I.	1 week and 190€
C18	Define a data's confidentiality criteria to determine if in a given project or digital object, there is the possibility for sharing	Likelihood control	N.I.	3 days and 75€
C19	Obtain previous consent from the data's source or involved entities	Risk sharing	N.I.	1 day and 25€
Nº	Designation	Strategy	State	Cost / time
C20	Creation of a protocol defining the workflow execution properties or create additional metadata, creating stronger bounds between the biological results	Exposure control	N.I.	2 weeks and approximately 380€

The association between the risk controls and the risks is presented in table 10:

Table 10. Association between the risks and risk control measures.

Risk	**R1**	**R2**	**R3**	**R4**	**R5**	**R6**	**R7**	**R8**	**R9**	**R10**	**R11**	**R12**
Control	C1 C2 C8 C9	C1 C2	C3 C5 C6 C7 C8 C10 C12	C1 C3 C9	C1 C3 C5 C8 C16 C17 C9	C1 C8 C9	C1 C6 C11	C4 C13 C14	C15	C18 C19	C17 C20	C11

4.4 RISK MONITORING, CONTROLLING, AND REPORTING

Risk exposure on the MetaGen-FRAME Project, namely the control measures already implemented and the respective risks, will be reviewed monthly during the life of the project. At these reviews new risks will be identified and assessed, existing risks will be reviewed, progress on agreed actions will be assessed, and new actions will be allocated where required. The effectiveness of the risk process will be reviewed to determine if changes to the approach, tools or techniques are required. Where process changes are agreed, this Risk Management Plan will be updated and reissued to document the revised process. In what concerns reporting, a list containing the top risks should be performed in order to review adequately the major risks and their respective treatment measures. This list must also be reviewed monthly, in order to update it, containing any new risks that may appear.

4.5 CONCLUSION

Our motivation for the poposed process resides in the field of Metagenomics with the MetaGen-FRAME use case. In the use case RM analysis: (i) twelve risks were identified (ii) all the risks were sucessfully analysed (iii) all the risks were sucessfully evaluated with the determination of which risks need treatment or only control (iv) risk treatment and control measures were found for each risk (v) all the typical DMP sections were complemented by the RMP, as there were risks allocated to each section. This validation is also acheived through the compliance of several evaluation metrics [10].

APPENDIX A: REFERENCES

1. Coimbra, M. (2012) Metagenomic Frameworks, Project Report, Universidade Técnica de Lisboa, Instituto Superior Técnico.
2. *Risk Management - Principles and guidelines*. ISO FDIS 31000:2009. Geneva, Switzerland : ISO
3. ISO (2009) Risk management - Risk assessment techniques. ISO IEC 31010:2009. Geneva, Switzerland
4. ISO (2009) Risk management – Vocabulary. ISO Guide 73:2009. Geneva, Switzerland
5. Braga, R. S. (2007) Automatic capture and efficient storage of e-Science experiment provenance, *Wiley InterScience.*
6. Deelman, E., Chervenak, A. (2008) Data Management Challenges of Data-Intensive Scientific Workflows, *USC Information Sciences Institute.*
7. Vermaaten, S., Lavoie, B., Caplan, P. (2012) Identifying Threats to Successful Digital Preservation: the SPOT Model for Risk Assessment, *D-Lib Magazine.*
8. Fernandes, D., Bakhshandeh, M., Borbinha, J. (2012) Survey of data management plans in the scope of scientific research, Inesc-ID, *Timbus Timeless Business.*
9. The Australian National University (ANU), "ANU Data Management Manual - Managing Digital Research Data at the Australian National University (version 10.09.17)," The Australian National University, Canberra, Australia, 2010.
10. Ferreira, F., Coimbra, M., Vieira, R., Proença, D., Freitas, A. T., Russo, L., N., S., Borbinha, J. (2013) Risk aware Data Management in Metagenomics, Inforum Conference.

APPENDIX B: KEY TERMS

The following table provides definitions [4], [5], [6], [7], [8] for terms relevant to the Risk Management Plan.

Term	Definition
Risk	Effect of uncertainty on objectives;
Asset	Anything of value to the organization;
Risk Management	Coordinated activities to direct and control an organization with regard to risk
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, analyzing, evaluating, treating, monitoring and reviewing risk
Risk Management Plan	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
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
Probability	Measure of the chance of occurrence expressed as a number between 0 and 1, where 0 is impossibility and 1 is absolute certainty
Consequence	Outcome of an event affecting objectives
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
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
Digital Preservation	Long term maintenance of the accessibility of a digital object
Data Management	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	Document that describes what data will be created, collected, stored, managed and disseminated during a project
E-Science	Global collaboration in key areas of science and the next generation of infrastructure that will enable it
Scientific Workflow	Means by which 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
Metagenomics	Study of populations of microorganisms, namely metagenomes

APPENDIX C: ABBREVIATIONS GLOSSARY

The following table provides the abbreviations used in the Risk Management Plan.

Abbreviation	Term
RM	Risk Management
RMP	Risk Management Plan
DP	Digital Preservation
DM	Data Management
DMP	Data Management Plan
WS	Web Service
NCBI	National Center for Biological Information
EMBL-EBI	European Bioinformatics Institute
ENA	European Nucleoite Archive
BLAST	Basic Local Alignment Search Tool
RDF	Resource Description Framework
SAM	Sequence Alignment/Map
XML	Extensible Markup Language