IJDC | *Peer-Reviewed Paper*

Data Management in Metagenomics: A Risk Management Approach

Filipe Ferreira
Universidade de Lisboa, Portugal

Miguel E. Coimbra
Universidade de Lisboa, Portugal

Raquel Bairrão
Universidade de Lisboa, Portugal

Ricardo Viera
Universidade de Lisboa, Portugal

Ana T Freitas
Universidade de Lisboa, Portugal

Luís M.S. Russo
Universidade de Lisboa, Portugal

José Borbinha
Universidade de Lisboa, Portugal

Abstract

In eScience, where vast data collections are processed in scientific workflows, new risks and challenges are emerging. Those challenges are changing the eScience paradigm, mainly regarding digital preservation and scientific workflows. To address specific concerns with data management in these scenarios, the concept of the Data Management Plan was established, serving as a tool for enabling digital preservation in eScience research projects. We claim risk management can be jointly used with a Data Management Plan, so new risks and challenges can be easily tackled. Therefore, we propose an analysis process for eScience projects using a Data Management Plan and ISO 31000 in order to create a Risk Management Plan that can complement the Data Management Plan. The motivation, requirements and validation of this proposal are explored in the MetaGen-FRAME project, focused in Metagenomics.

Received 13 January 2014 | *Accepted* 26 February 2014

Correspondence should be addressed to Filipe Ferreira, Rua Alves Redol 9, 1000-029 Lisboa. Email: filipe.ferreira@ist.utl.pt

An earlier version of this paper was presented at the 9th International Digital Curation Conference.

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by the University of Edinburgh on behalf of the Digital Curation Centre. ISSN: 1746-8256. URL: http://www.ijdc.net/

http://dx.doi.org/10.2218/ijdc.v9i1.299
DOI: 10.2218/ijdc.v9i1.299

Introduction

eScience typically represents increasingly global collaborations of people and resources (Hey and Trefethen, 2003), using large scale infrastructures to process vast data sets, where data management and digital preservation concerns are addressed to mitigate the emerging risks to digital objects (David and Spence, 2003).

Within the data management and digital preservation concerns, the concept of a Data Management Plan (DMP) is used to represent the set of rules and good practices a project must follow regarding data management, according to the objectives of specific stakeholders – usually a funding organization (Fernandes et al., 2012). From another perspective, a DMP intends to 'protect' digital objects against several threats that exist in typical eScience workflows. As the mitigation of risks is the main goal of risk management, an opportunity arises for understanding how risk management can be used to enrich the DMP concept.

The motivation and validation for our proposal was the MetaGen-FRAME project (Coimbra, 2012), which we expect it can be used as a general framework for the field of bioengineering (a heterogeneous area comprising molecular biology, medicine and bioinformatics.). In this sense, we understand MetaGen-FRAME to be an eScience and Metagenomic project, focused on sequence analysis and genome annotation.

This paper is structured as follows: first, we outline the principles of eScience, scientific workflows, data management and DMPs; second, we introduce the concepts of digital preservation, risk management and the Risk Management Plan (RMP); third, we describe an analytical process for creating an RMP; fourth, we present the previous process' validation, based on the MetaGen-FRAME project; and finally, we present our major conclusions and some remarks about future work.

Data Management in eScience

eScience involves global collaboration and large datasets supported by an infrastructure (Jankowski, 2007). It is based on scientific workflows, allowing scientists to execute, reconfigure and rerun their analysis in a verifiable way (Braga and Digiampietri, 2008), typically involving many steps and vast datasets (Deelman and Chervenak, 2008).

Data management is an integral part of eScience. It allows researchers to produce higher quality data, increase the exposure of their research and protect data from being lost or misused (Fernandes et al., 2012). The scientific community has increasingly perceived concerns about data management, namely data's provenance, sharing, access and archival (Fernandes et al., 2012). As a result of these concerns, research funders have been increasingly requesting the inclusion of a DMP as part of the project proposals. A typical DMP describes how data will be created, stored and shared, with two purposes (Fernandes et al., 2012):

1. To guide researchers to reuse data,
2. To record the project's data management decisions.

Digital Preservation as a Risk Management Approach

Digital preservation aims to keep digital objects accessible over long periods of time (Rosenthal et al., 2005). For that, digital objects must be what they claim to be, implying trustworthiness and authenticity. Information provenance and traceability must be assured. Digital preservation environments might require scalability to face technology's evolution, which may be achieved through replacement of technological components, thus implying heterogeneity (Barateiro et al., 2010).

Risk Management Plan

Risk management identifies, assesses and mitigates risks to an acceptable level. It manages risks, i.e. the uncertainty associated with events which can affect assets (ISO, 2009c; Barateiro et al., 2010). A risk can be triggered by a positive event (an opportunity), or negative event (a threat) (Barateiro, 2012).

Standards, methods and tools for risk management vary with the market sector, type of business or organizational activities (Ramirez, 2008). There are standards that focus on defining the generic terminology, process, principles, methods and techniques, as well as specific domain standards (Ferreira et al., 2013a). ISO 31000 proposes a reference process to execute risk management properly (ISO, 2009a) as shown in Figure 1.

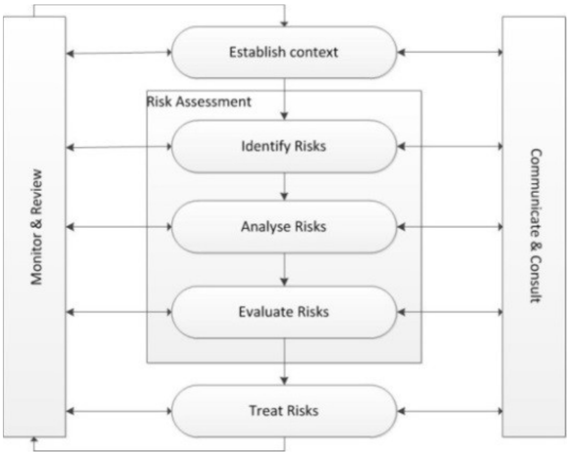

Figure 1. ISO 31000 risk management process.

The process proposes that the context, strategic objectives and risk criteria of risk management must first be defined (Barateiro et al., 2010). The next step is risk assessment, which is composed of three distinct phases: risk identification, which generates the list of risks (Barateiro et al., 2010); risk analysis, to consider the impacts and probabilities determining the risk level; and risk evaluation, to determine what risks need treatment or can be only controlled. The next step is risk treatment, where controls are designated to risks. Communication and review takes place throughout all previous stages.

An RMP defines the scope and process for the assessment and treatment of risks. The objective of this plan is to define a strategy for the management of risks to 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. A typical RMP comprises the following steps:

1. Introduction,
2. Planning,
3. Execution.

RMP and DMP Correlation

Each research project has specific purposes, resulting in specific policies and thus in different instances of a DMP. However, there are always common issues enabling the definition of generic DMP guidelines, as proved from a comparative analysis already performed (Ferreira et al., 2013b). Due to this fact we defined a set of typical sections to identify common risks, like ethical risks (Kaye et al., 2010), metadata preservation risks (Day, 2004) or data dissemination risks (Bimholtz and Bietz, 2003). Ferreira et al. (2013a) presents a more detailed list of typical risks related to eScience and its workflows.

Analysis Process for RMP Creation

In order to assure we can be efficient and effective in a systematic way, a generic process for the definition of an RMP for eScience projects should be possible. This process must be able to effectively create an RMP for an eScience project and also to align it with a corresponding DMP.

The process we propose, in response to the previous hypothesis, is based on good practices from the ISO 31000, a risk management reference. The process, as presented in Table 1, is based on three phases, each made of a set of steps, with expected results. See (Ferreira et al., 2013b) for further details.

Table 1. Proposal of analysis process for creation of RMP.

Phase	Step	Expected results
1. Introduction	1.1. Describe the project	Description of the workflow, tools, inputs, outputs
	1.2. Establish the goal and purpose of RMP	Identification of the goal of the RMP, the relation with the DMP and the intended audience
	1.3. Define the authority	Identify the authority financing the project and conducting the risk management analysis
	1.4. Establish the context	Identify the environment to which the RMP applies
2. Planning	2.1. Organizations and responsibilities – identify stakeholders	Description of all the stakeholders involved in the project

Phase	Step	Expected results
	2.2. Select Techniques	List of all the relevant techniques for risk assessment. Any technique from ISO/IEC 31010 (2009b) can be used in the risk assessment of a project, although there are several techniques more appropriate than others
3. Execution (Proceedings)	3.1. Identify assets, vulnerabilities and events	Identification of assets, vulnerabilities and events. Those concepts will ease the identification of risks and improve their understanding
	3.2. Identify risks	Identification of risks. Risks must be allocated according to the DMP sections (some sections might have no risks)
	3.3. Analyse risks	Calculation of the levels of risk
	3.4. Evaluate risks	Evaluation of risks through a risk matrix
	3.5. Treat risks	List of controls for each risk (also the handling strategy)
	3.6. Monitor and communicate risks	The process and rate of monitoring is defined according to each control and the total duration of the project

In order to validate the proposed method, a real case was used, Results are detailed in the next section, and can be compared with the expected results mentioned in Table 1.

Process Validation: The MetaGen-FRAME Project

The MetaGen-FRAME project is presented as a case study for validation of the proposed process. It is a metagenomics (Wooley et al., 2010) project whose practical results (Ferreira et al., 2013b) are presented in the next few sections according with the process phases presented in Table 1.

Phase One: Context

The MetaGen-FRAME is (Step 1.1) a metagenomics project with the goal to perform the analysis of large datasets of DNA sequences obtained by using Next Generation Sequencing (NGS) technologies to environment samples. In this particular case relatively-controlled environments are considered (possibly composed of several types of different bacteria, with each type being present in different quantities), whose chemical reactions may be influenced and enhanced. In general, metagenomics focus on the study of bacteria (prokaryotes). The samples origin can vary, ranging from an open environment, like the ocean, to a closed one like the human digestive system. The tools used for task execution are pre-selected. The project's main tasks are shown in Figure 2 in more detail.

The RMP goal and purpose (Step 1.2) is to describe how the MetaGen-FRAME risks are identified, analysed and evaluated, and also how the risk management activities are performed and monitored. This RMP also intents to complement the corresponding DMP, as it was stated before. The intended audience for this RMP is the project and management team.

Regarding the authority in the project (Step 1.3), the Fundação para a Ciência e a Tecnologia (FCT) was identified as the founder of the project. There is no official risk management authority in the risk management analysis of this project. The RMP is addressing an eScience environment, namely in metagenomics (Step 1.4).

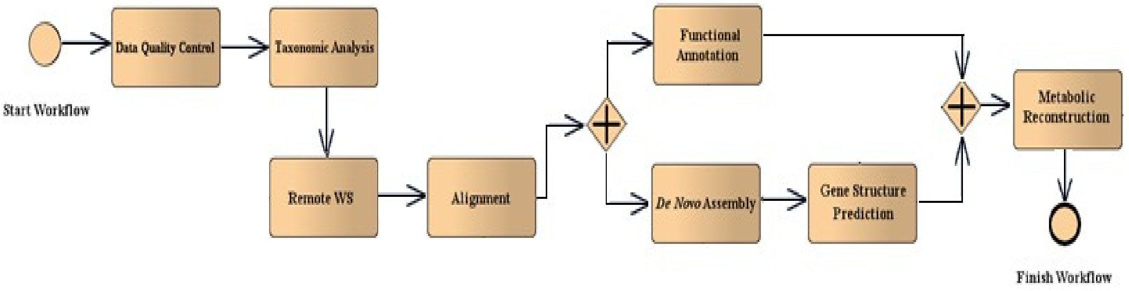

Figure 2. The MetaGen-FRAME workflow (detailed in Table 2).

Table 2. MetaGen-FRAME tasks using Taverna[1] (Coimbra, 2012), (Ferreira et al., 2013b).

Task	Description
Data quality control	Before a data set is processed, the information needs to respect certain quality thresholds. This step may be local or remote. The tool used is NGS QC Toolkit[2]. The inputs are a text file with the sequences that are going to be analysed, a string with the format used by the previous file, a string detailing which sequence technology was used, and a variable to filter sequences by size. The output is a filtered version of the original data set, as well as statistics regarding the removed sequences
Analysis of taxonomy	This analysis determines the sample's microbial diversity, to determine the different organisms that are present and, if possible, their resolution levels (species, kingdom, etc). The tool used is MetaPhlAn[3], being a local task. The input is the filtered dataset produced previously, as well as a value which may represent a) the minimum percentage identity that a taxon (a group of one or more populations of organism(s)) needs to have to be considered valid; or b) the number of taxons to be returned as valid, in decreasing order of percentage identity. The output consists of several lists of organisms present in the sample, with respective resolutions and identity percentages

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

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

[3] MetaPhlAn: http://huttenhower.sph.harvard.edu/metaphlan/

Task	Description
Remote web service	A sequence of web services that use the NCBI[4] database. The web service sequence uses as an input the lists obtained in the former task and produces a set of corresponding NCBI IDs. Later in the web service sequence, the NCBI is consulted using the IDs and returns a list of sequences associated to the existing taxonomic results, in .fasta format
Alignment	Establishment of an order between the sequences by comparison with the sequences obtained previously. This step uses a parallel version of TAPyR[5] mapper and is performed locally. It receives as an input the former list of sequences and generates as outputs a set of aligned sequences in .SAM format, a set of non-aligned sequences in a .fasta file, a set of aligned sequences also in a .fasta file
Functional annotation	The set of consensus sequences are submitted to a functional annotation procedure. This may be a local or remote task. It is composed of two steps, starting with a separate execution of the NCBI BLAST program and then feeding its results in .xml format to the default tool Blast2GO[6]. It receives as an input the .fasta file with alignment sequences produced in the alignment task and produces image and texts identifying the main genes and components that were found to be associated to the aligned reads
De novo assembly	Sample identification by reconstruction. MetaVelvet[7] is the default program. This task may run locally or remotely on a more powerful infrastructure. As an input, it receives the set of non-aligned sequences and as an output it returns contigs (junctions of several sequences)
Gene structure prediction	This is used to obtain information about the sample's genes and to find out if genetic structures are present. One tool that can execute this step is BG7[8]. This is a local task. As an input, it receives the set of contigs generated in the de novo task and the output contains information regarding predicted genes in the following formats: .gff, .gbk, .tsv and .xml
Metabolic reconstruction	The aim was to produce results associated with the sample's metabolism. Due to technical constraints, this task was implemented implicitly by the result display from the Functional Annotation and Gene Structure Prediction steps

Phase Two: Planning

In Phase 2, the stakeholders involved, and their responsibilities, were identified in a responsibility assignment (RACI) chart in Table 3 (Step 2.1). The techniques used in the stages of risk assessment were taken from the ISO 31010, which were (Step 2.2):

- **Risk identification:** Check lists, brainstorming, SWIFT, FMEA/FMECA, HRA;
- **Risk analysis:** SWIFT, FMEA/FMECA, HRA, decision tree analysis;
- **Risk evaluation:** A risk matrix.

[4] NCBI: http://www.ncbi.nlm.nih.gov/
[5] TAPyR - Tool for Alignment of Pyrosequencing Reads: http://www.tapyr.net/
[6] Blast2GO: http://www.blast2go.com/b2ghome
[7] MetaVelvet: A short read assembler for metagenomics: http://metavelvet.dna.bio.keio.ac.jp/
[8] BG7 Bacterial genome annotation system: http://bg7.ohnosequences.com/

Table 3. RACI chart of the MetaGen-FRAME project (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 risk management	R	I		
Definition of the acceptable levels of risks	C	R	I	
Risk Management Plan acceptance	I	R	C	
Control's efficiency and effectiveness monitoring	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

Phase Three: Execution (Proceedings)

For Step 3.1, the following types of assets were identified in the MetaGen-FRAME project:

- Data (**A1**);
- Tools (**A2**) such as Taverna, Blast2GO, NGS QC Toolkit, BG7, MetaPhlAn, TAPyR;
- Computational servers (**A3**);
- Databases (**A4**);
- Local personal computer (PC) (**A5**);
- Web-services (**A6**); and
- Workflow/Tasks (**A7**).

The corresponding vulnerabilities and the events that can exploit them and affect the assets are expressed in Table 4 and Table 5.

Table 4. List of vulnerabilities of MetaGen-FRAME project.

ID	Designation
V1	Unreliable storage hard drive in the local PC. PCs and external hard drives are useful for short term storage, but inadequate for long term storage
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 formation in security
V3	Poor debugging capabilities of Taverna. In the case of failure, it is problematic if the SWMS does not provide debugging capabilities that show the failure cause
V4	Lack of syntactic and semantic verification mechanisms to check the initial inputs given by the human operator
V5	Lack of a long storage policy
V6	Communication channel overload (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

Table 5. List of events of MetaGen-FRAME project.

Event	Designation
E1	Media units, databases, web services, computational servers or network failures
E2	Media units, databases or computational servers maintenance
E3	Hacker attacks to the infrastructures or communication channels
E4	Natural disasters (fires, floods, earthquakes)
E5	Insertion of wrong or incomplete input values by the human operator
E6	Tool discontinuation and lack of support
E7	Financial, legislative or organizational changes in the organization running the databases used, leading to changes on the digital preservation policies
E8	Sharing of information without consent
E9	Project's abandonment from a stakeholder
E10	Sudden workflow interruption

Regarding risk identification (Step 3.2), the MetaGen-FRAME project extrapolates several types of information from the dataset it receives as an 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. In general, projects that deal with DNA sequences involve important issues associated with the secrecy and storage of data, as it will potentially convey information that is important to the project owner or entity's activity. An example of such an activity is the process of analysing and enhancing biomass decomposition, fuel refinement, crude extraction, among others.

Such processes may constitute 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 in Table 6.

Table 6. Identified risks with the respective assets, vulnerabilities and events.

Risks		Assets	Vulnerabilities	Events
R1:	Accidental change or deletion of digital objects	A1, A7	V4	E5, E10
R2:	Insertion of wrong input values: One example is the introduction of a 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, A7	V4	E5
R3:	Change of the external web services, National Center for Biological Information (NCBI), computational servers or local PC used causing their unavailability or failure	A3, A4, A5, A6, A7	V1, V2	E1, E2, E3, E4
R4:	Loss of information due to communication failures	A1, A7	V6	E3, E10
R5:	Loss of information and data traceability due to a media fault, compromising the workflow's recreation	A1, A2, A7	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, E10
R7:	Lack of financial or legal requirements to preserve data	A1	V7	E7
R8:	Obsolesce of the tools used in the workflow or in the NCBI or local PC	A2, A7		E6
R9:	Occurrence of an error that cannot be explained, leading to repetition of the whole experiment	A2, A7	V3	E6, E10
R10:	Sharing of confidential data	A1	V8	E8
R11:	Difficulties sharing the information and the workflow's execution details in other future scenarios	A1, A2, A7	V8	E8
R12:	Stakeholder's lack of involvement	A1		E9

As the RMP intents to complement a DMP, the risks presented in Table 6 must be allocated to the generic sections of the DMP (Ferreira et al., 2013b), leading to the distribution expressed in Table 7. As it can be seen, all the DMP sections have at least one risk associated, showing that DMP sections can be used to categorise the risks found, and that there are risks associated with every DMP section considered.

Table 7. Relation between the typical sections of a DMP and the identified risks.

Section	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

To perform risk analysis (Step 3.3) and calculate the level of every risk, likelihood and consequence criteria were defined according to the criterion from a very low priority (0.1) to a very high priority (0.9). Risk levels are obtained by multiplying the risk's likelihood and consequence. The likelihood, consequence and respective risk levels of each risk are presented in Table 8.

Table 8. Values of likelihood, consequence and risk levels of each risk.

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

For the evaluation of risks (Step 3.4), a risk matrix was developed (see Table 9). 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, thus being the last ones treated. R8 and R12 have a low priority and need only to be controlled.

For risk treatment (Step 3.5), risk control measures were identified and are presented in Table 10. The controls use different strategies to mitigate the risk by reducing specifically the consequence of the risk, the exposure of the vulnerability, the likelihood of the event or sharing the risk with other entities.

Table 9. 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				

Table 10. Required controls, respective strategies (Barateiro, 2012) and the risks each control applies. RC = Reduce Consequence, RE = Reduce Exposure, RL = Reduce Likelihood, SR = Share Risk.

Number	Designation	Strategy	Risk(s)
C1	Use several backup systems (local and remote) in the local PC, for example, using a system like shadow copy to store all the data and metadata	RC	R1, R4, R5, R6, R7
C2	Implement syntactic and semantic verification mechanisms to alert if inputs don't have the correct format and content	RE	R2
C3	Improve the security measures in NCBI, computational servers, PCs and communication channels (e.g. better antivirus, encryption, firewall)	RE	R3, R4, R5
C4	Keep all the software and hardware components up to date	RL	R8
C5	Backup systems for the NCBI and computational servers	RC	R3, R5
C6	Access other genome databases, such as the ones offered by the EMBL-EBI[9] like ENA[10], in case NCBI becomes unavailable	RC	R3, R7
C7	Access other computational servers if current ones fail (fall back)	RC	R3
C8	Create a replicated central storage to store, in real time, the execution results from the workflow using, for example, shadow copy	RC	R1, R3, R5, R6

[9] The European Bioinformatics Institute: http://www.ebi.ac.uk/

[10] The European Nucleotide Achieve (ENA): http://www.ebi.ac.uk/ena/data/view/CP006584

Number	Designation	Strategy	Risk(s)
C9	Create a long term storage policy with a specialized organization	RL	R1, R4, R5, R6
C10	Implement anti-fire and earthquake measures in the NCBI and hosting of the computational servers	RL	R3
C11	Allocate an emergency budget for financial changes in the NCBI organization or in case of abandonment of any project member	RC	R7, R12
C12	Insert alternative tools in the workflow, if the main ones fail	RC	R3
C13	Use open-source tools and formats	RL	R8
C14	Add a new component in Taverna to check the return value, and if this would be an error, a new tool would treat it by avoiding workflow termination. The error would be registered, to trace its origin	RE	R9
C15	Create of alternative forms of documentation, for example, physical documentation which can be digitalised and stored in a backup system	RC	R5
C16	Modify formats used by the framework so each output references the associated input (RDF style), leading to interconnected data elements	RE	R5, R11
C17	Define data confidentiality criteria for sharing data	RL	R10
C18	Obtain previous consent from the data's source/entities	SR	R10
C19	Create a protocol defining the workflow execution properties for achieving stronger bounds between the biological results	RE	R11

All of the risks and controls for the MetaGen-FRAME project need to be monitored (Step 3.6) in a monthly basis until the end of the project. For each review the risks and control's effectiveness must be communicated to all involved stakeholders.

Conclusions and Future Work

Risk management has applications in different areas and projects. In eScience, collaboration and data management are crucial. New challenges and risks are raised and must be assessed so the project's data can be preserved and reused. To answer these dilemmas, DMPs are developed before a research project takes place. The paper tries to give added value to the DMP concept using risk management principles so that the emerging risks surrounding eScience projects and digital preservation can be met. It is becoming necessary to understand how risk management can enhance DMPs. In order to assess that, we propose a process that uses jointly DMP and ISO 31000 to create an RMP.

Our motivation for the proposed process resides in the field of metagenomics with the MetaGen-FRAME case study. In the case study risk management analysis:

1. 12 risks were identified;
2. All the risks were successfully analysed;
4. All the risks were successfully evaluated with the determination of which risks need treatment or only control;
5. Risk treatment and control measures were found for each risk;
6. All the typical DMP sections were complemented by the RMP, as there were risks allocated to each section.

This validation is also achieved through the compliance of several evaluation metrics (Ferreira et al., 2013a). In future work, we intend to use these results to generalize the process for the development of a DMP and an RMP in bioengineering, leading to a better curation in the same domain.

Acknowledgements

This work was supported by national funds through the 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.

References

Barateiro, J., Antunes, G., Freitas, F., & Borbinha, J. (2010). Designing digital preservation solutions: A risk management based approach. *The International Journal of Digital Curation*, *(5)*1, 4-17. doi:10.2218/ijdc.v5i1.140

Barateiro, J., (2012). A risk management framework applied to digital preservation. PhD Dissertation. Universidade Técnica de Lisboa, Instituto Superior Técnico.

Bimholtz, J.P., & Bietz, M.J. (2003). Data at work: Supporting sharing in science and engineering, GROUP 03. Paper presented at the 2003 International ACM SIGGROUP Conference, Sanibel Island, Florida, USA.

Braga, R.S., & Digiampietri, L.A. (2008). Automatic capture and efficient storage of eScience experiment provenance. *Concurrency and Computation: Practice and Experience, 20*(5), 419-429. doi:10.1002/cpe.1235

Coimbra, M. (2012). Metagenomic frameworks. Project Report. Universidade Técnica de Lisboa, Instituto Superior Técnico.

David, P.A., & Spence, M. (2003). *Towards institutional infrastructures for eScience: The scope of the challenge.* Oxford Internet Institute. Retrieved from http://sydney.edu.au/about/leadership/vc/e-science.pdf

Day, M. (2004). *Preservation metadata initiatives: Practically, sustainability, and interoperability*. University of Bath, UK. Retrieved from http://www.ukoln.ac.uk/preservation/publications/erpanet-marburg/day-paper.pdf

Deelman, E., & Chervenak, A. (2008). Data management challenges of data-intensive scientific workflows. In *8th IEEE International Symposium on Cluster Computing and the Grid (CCGRID '08)* (pp. 687-692). doi:10.1109/CCGRID.2008.24

Fernandes, D., Bakhshandeh, M., & Borbinha, J. (2012). Survey of data management plans in the scope of scientific research. Inesc-ID, Timbus Timeless Business. Retrieved from http://www.inesc-id.pt/ficheiros/publicacoes/8410.pdf

Ferreira, F., Coimbra, M., Vieira, R., Proença, D., Freitas, A.T., Russo, L.N.S., & Borbinha, J. (2013a). *Risk aware data management in metagenomics.* Paper presented at the Inforum 2013 Conference, Évora, Portugal.

Ferreira, F., Coimbra, M., Vieira, R., Bairrão, R., Freitas, A.T., Russo, L.N.S., & Borbinha, J. (2013b). A risk management plan in metagenomics. Inesc-ID, Technical Report. Retrieved from http://www.inesc-id.pt/ficheiros/publicacoes/9356.pdf

Hey, T., & Trefethen A. (2003). The data deluge: An eScience perspective. In F. Berman, G. Fox, & A. J. G. Hey (Eds.), *Grid Computing: Making the Global Infrastructure a Reality* (pp. 809-824). Chichester, UK: Wiley.

ISO. (2009a). Risk management: Principles and guidelines. ISO FDIS 31000:2009. Geneva, Switzerland.

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

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

Jankowski, N.W. (2007). Exploring eScience: An introduction. *Journal of Computer-Mediated Communication, 12*(2), 549-562. doi:10.1111/j.1083-6101.2007.00337.x

Kaye, J., Boddington, P., Vries, J., Hawkins, N., & Melham, K. (2010). Ethical implications of the use of whole genome methods in medical research. *Europe Journal of Human Genetics, 18,* 398-403. doi:10.1038/ejhg.2009.191

Ramirez, D. (2008). Risk management standards: The bigger picture. *Information Systems Control Journal, 2008*(4). Retrieved from http://www.isaca.org/Journal/Past-Issues/2008/Volume-4/

Rosenthal, D.S.H., Robertson, T., Lipkis, T., Reich, V., & Morabito, S. (2005). Requirements for digital preservation systems: A bottom-up approach. *D-Lib Magazine, 11*(11). doi:10.1045/november2005-rosenthal

Wooley J.C., Godzik, A., & Friedberg, I. (2010). A primer on metagenomics. *PloS Computational Biology, 6*(2), e1000667. doi:10.1371/journal.pcbi.1000667