Data Management in Metagenomics:

A Risk Management Approach

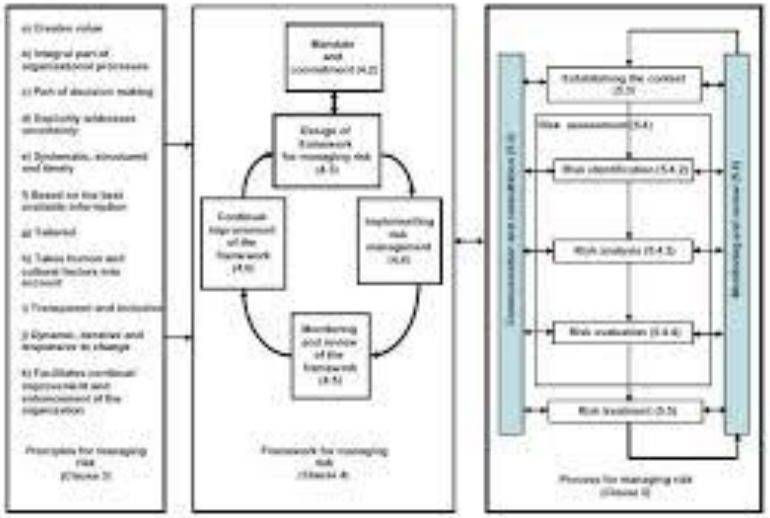
Filipe Ferreira, Miguel Coimbra, Raquel Bairrão, Ricardo Vieira, Ana Freitas, Luís Russo, José Borbinha
INESC-ID/IST (Portugal)

The Challenge:
Addressing Data Management (DM) concerns in e-Science

- **Fact**: **Data Management Plans** (DMP) already are being required by funding agencies…

- **Problem**: We claim actual guidelines for DMP can be limited and incomplete

- **Hypothesis**: We propose DM can be improved if the DMP can be complemented by a **Risk Management Plan** (RMP)



Key **Data Management** principles for a DMP:

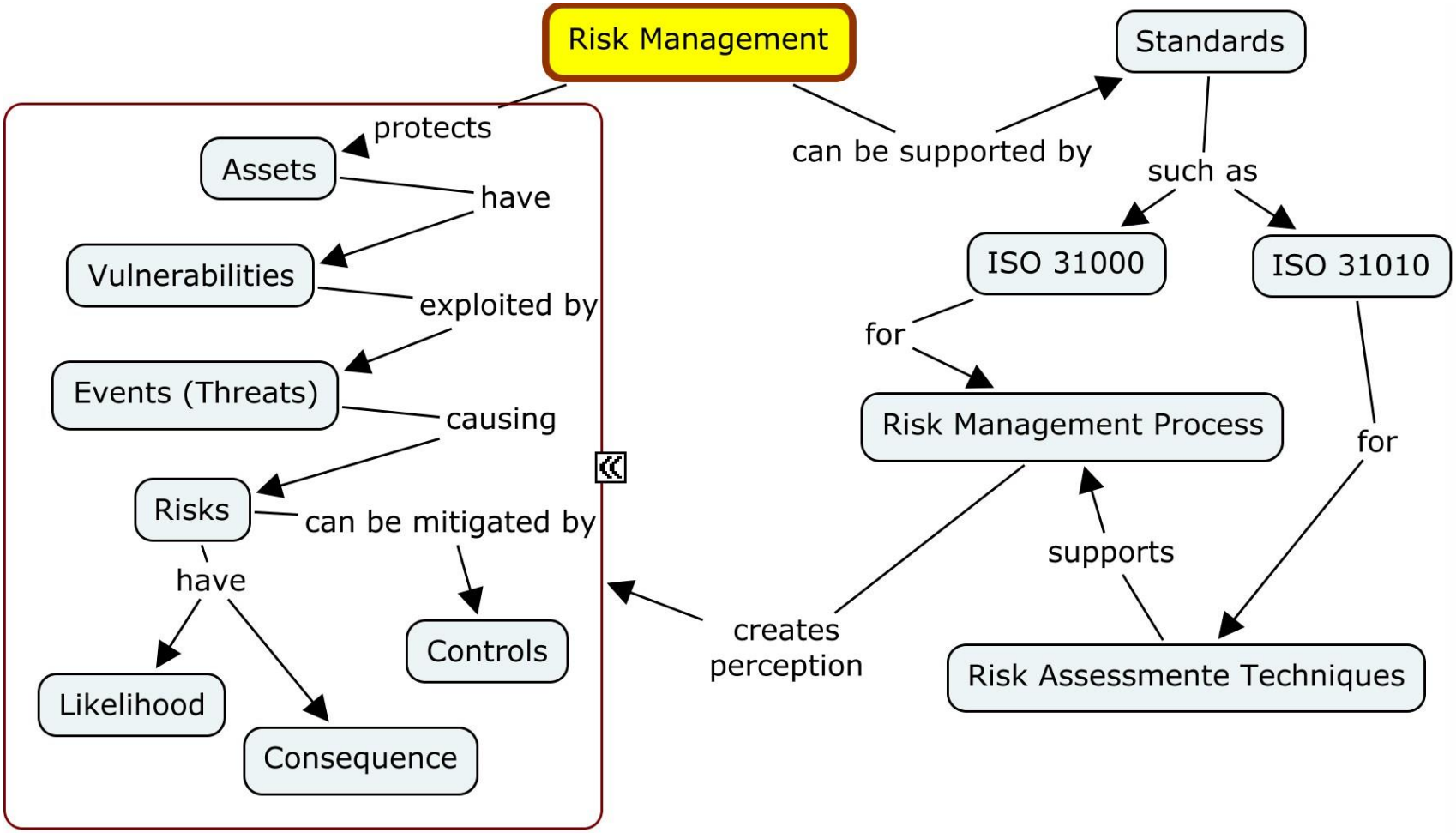
- Describe what and how data will be created, stored, preserved and shared in the project
- Guide researchers on how to reuse data in the project
- Records the project's decisions on Data Management

Typical Sections in a DMP	Source of Guidelines/Recommendations			
	DCC	ANU	NSF (Eng.)	NSF (Bio.)
Ethics and privacy	X	-	-	X
Resourcing (Budget)	X	X	-	-
Data Dissemination/sharing and licensing	X	X	X	X
Data Storage, Preservation and security	X	X	X	X
Data owners and stakeholders	-	X	X	X
Responsibilities	-	X	X	X
Data Formats and Metadata	x	X	X	X
Products of Research/Documentation	-	X	X	X

DCC – Digital Curation Center
ANU – Australian National University
NSF – National Science Foundation (USA)

But we found Data Management is, a lot, about RISKs:

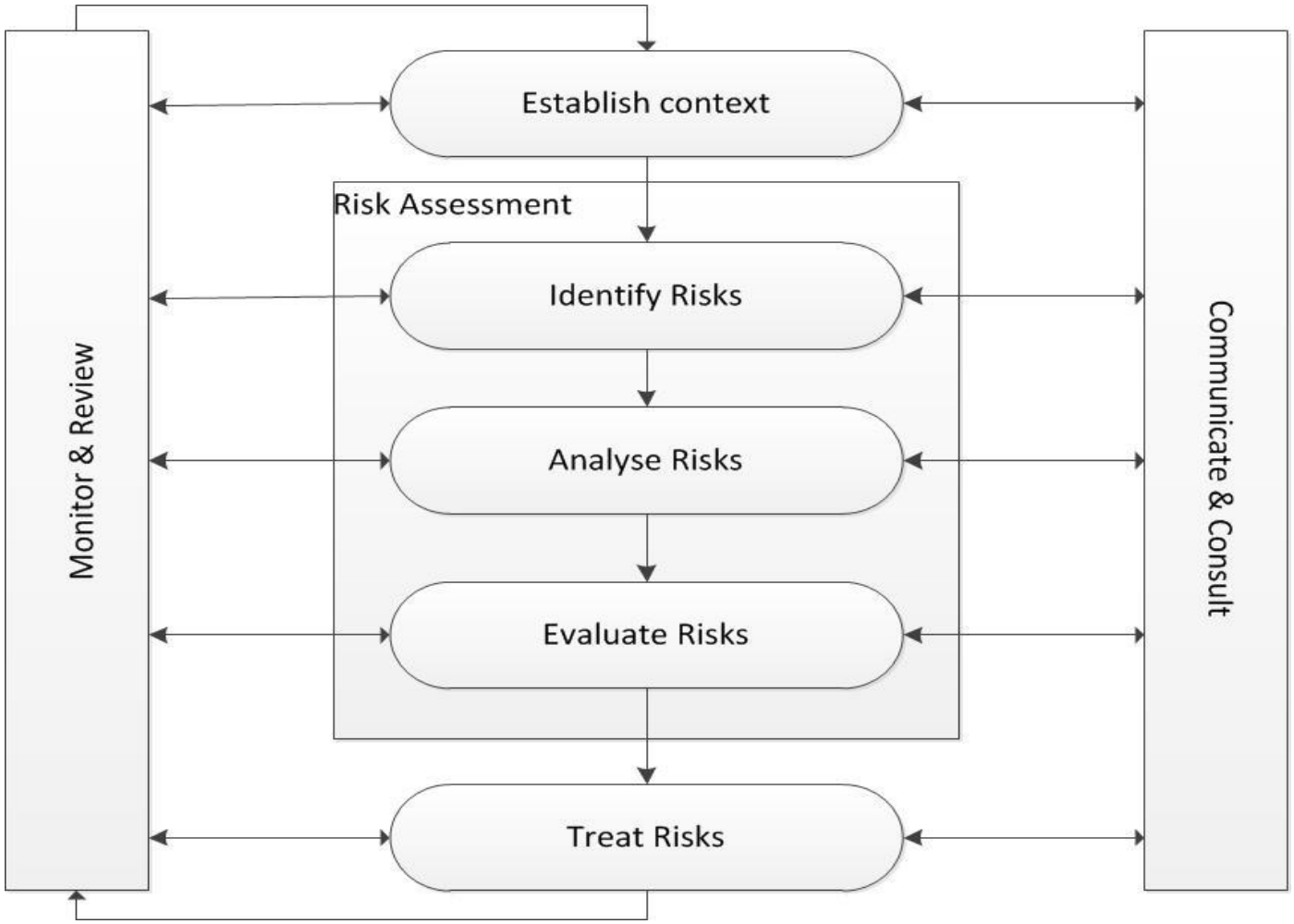
Usual concerns on the **creation, storing, preservation, sharing, reuse and deciding on data** might be about **mitigating risks**, which should be addressed as:



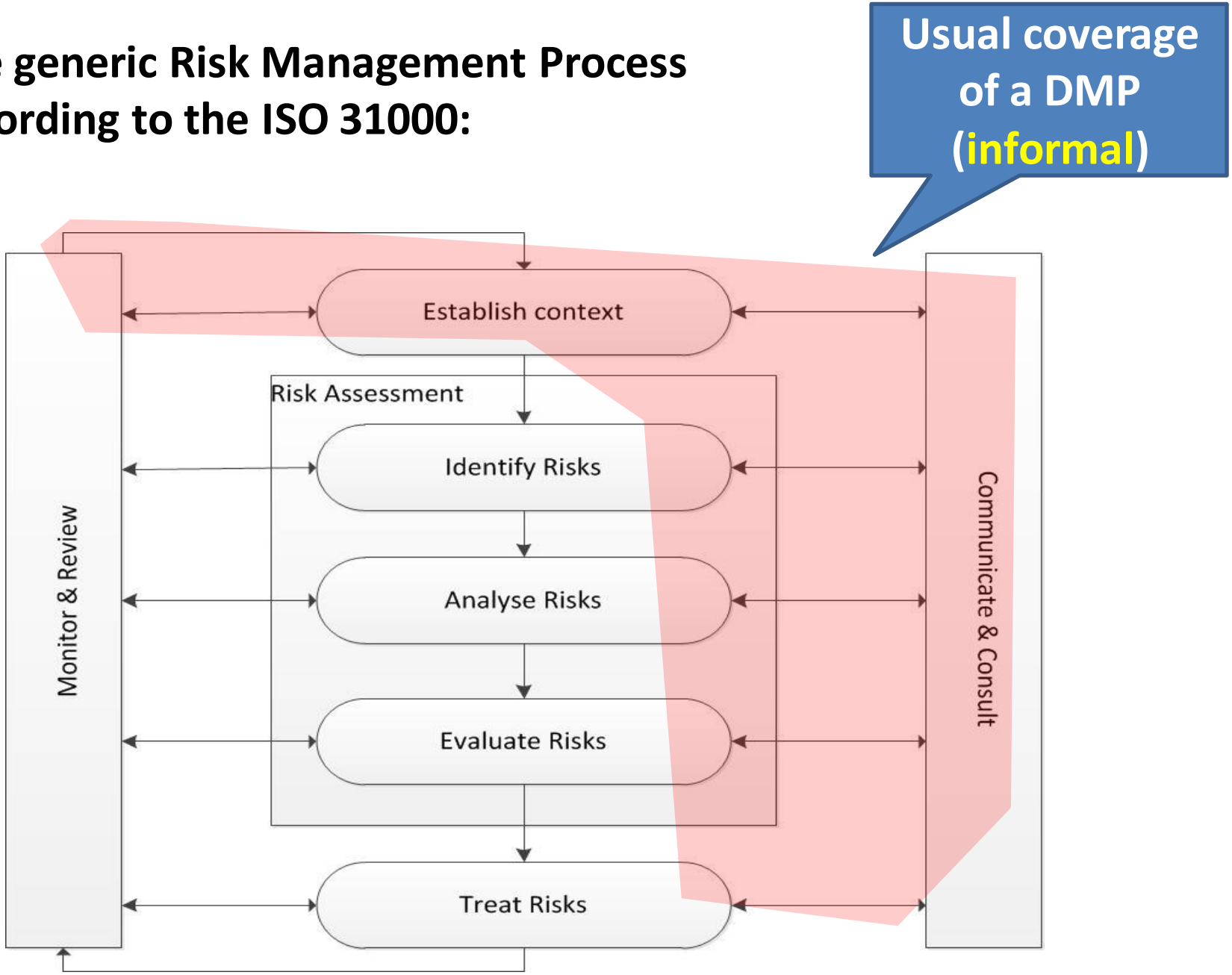
Thus, our hypothesis: A DMP should be complemented by a Risk Management Plan (RMP)!!!

- Concepts from [ISO 31000 – Risk Management Principles…]:
  - **Risk Management Plan** - specifies the approach, the management components and resources to be applied 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
  - **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
- Concluding: We can define the guidelines of a RM process for e-Science guided by the ISO 31000 !!!

The generic Risk Management Process according to the ISO 31000:

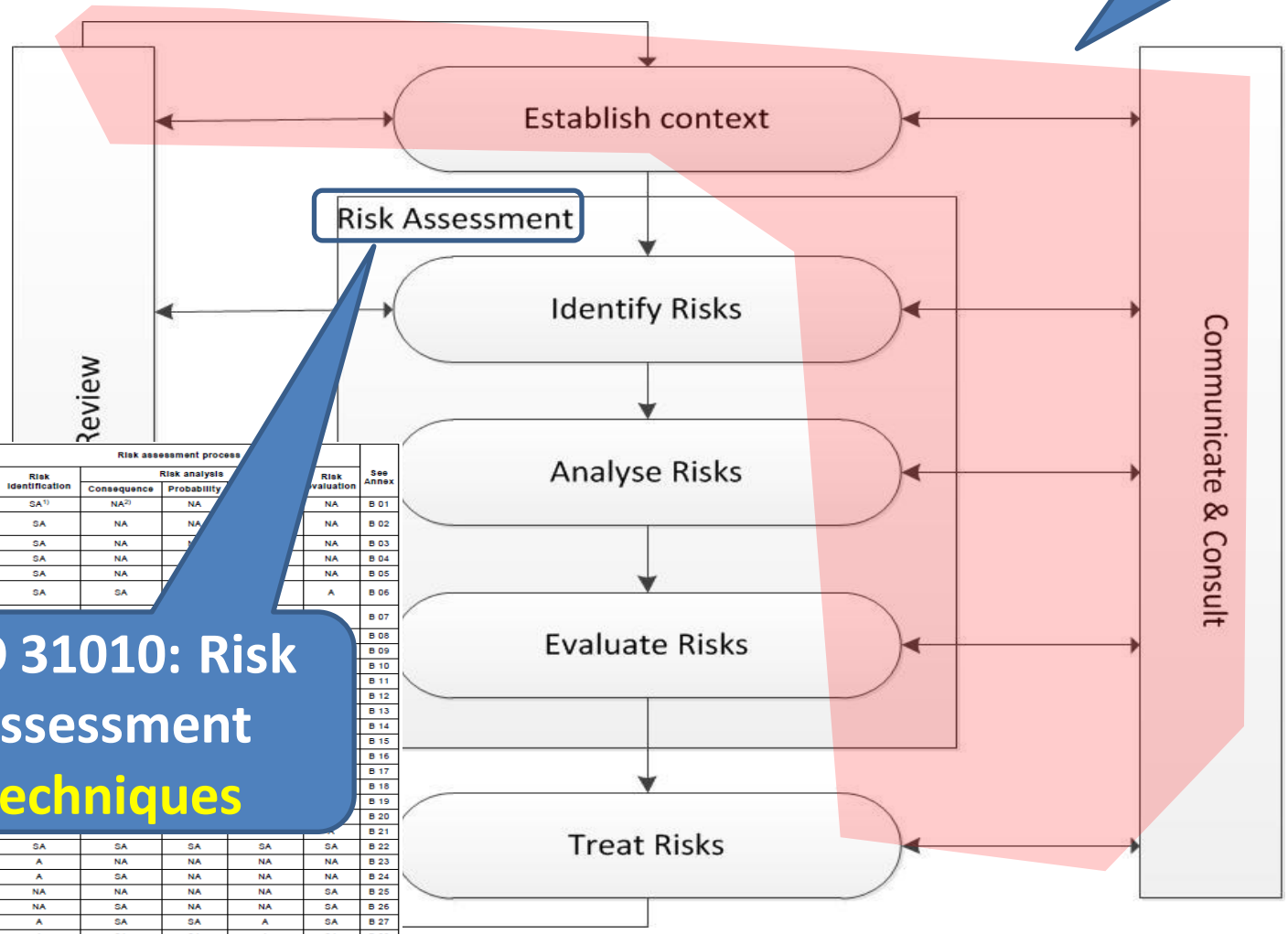


The generic Risk Management Process according to the ISO 31000:



The generic Risk Management Process according to the ISO 31000:

Usual coverage of a DMP (informal)

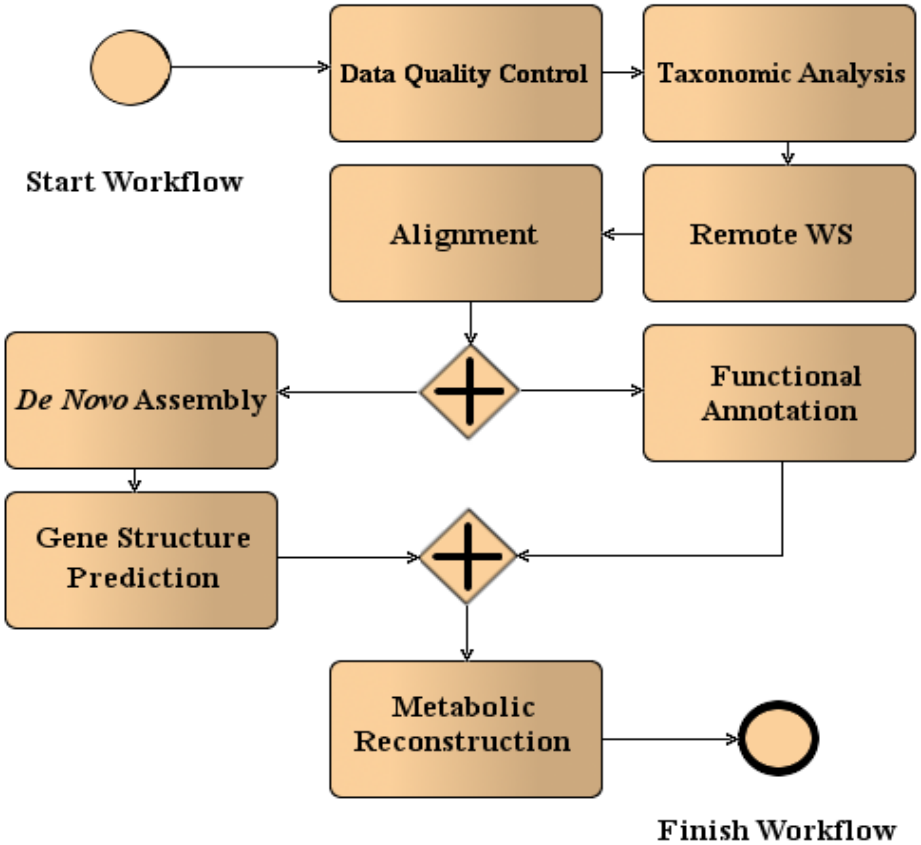


Tools and techniques	Risk assessment process					See Annex
	Risk identification	Risk analysis			Risk evaluation	
		Consequence	Probability			
Brainstorming	SA[1)]	NA[2)]	NA		NA	B 01
Structured or semi-structured interviews	SA	NA	NA		NA	B 02
Delphi	SA	NA			NA	B 03
Check-lists	SA	NA			NA	B 04
Primary hazard analysis	SA	NA			NA	B 05
Hazard and operability studies (HAZOP)	SA	SA			A	B 06
Hazard Analy Points (HA						B 07
Environ						B 08
Structur						B 09
Scenario						B 10
Busines						B 11
Root ca						B 12
Failure						B 13
Fault tre						B 14
Event tr						B 15
Cause a						B 16
Cause-a						B 17
Layer pr						B 18
Decision						B 19
Human re						B 20
Bow tie analy						B 21
Reliability centred maintenance	SA	SA	SA	SA	SA	B 22
Sneak circuit analysis	A	NA	NA	NA	NA	B 23
Markov analysis	A	SA	NA	NA	NA	B 24
Monte Carlo simulation	NA	NA	NA	NA	SA	B 25
Bayesian statistics and Bayes Nets	NA	SA	NA	NA	SA	B 26
FN curves	A	SA	SA	A	SA	B 27
Risk indices	A	SA	SA	A	SA	B 28
Consequence/probability matrix	SA	SA	SA	SA	A	B 29
Cost/benefit analysis	A	SA	A	A	A	B 30
Multi-criteria decision analysis (MCDA)	A	SA	A	SA	A	B 31

ISO 31010: Risk assessment techniques

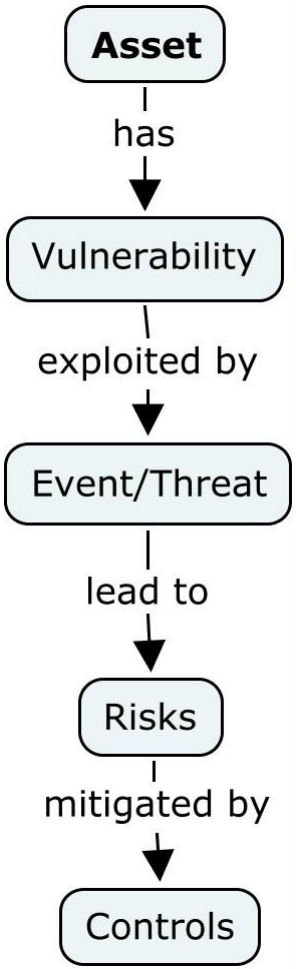
Motivation and Validation:

- **(Area) Metagenomics**: The study of populations of microorganisms (metagenomes = samples containing the DNA obtained from uncultured microorganisms)
- **(Project) MetaGen-FRAME:** Design of an open and flexible framework with several bioinformatics modules to study environments composed by multiple types of bacteria



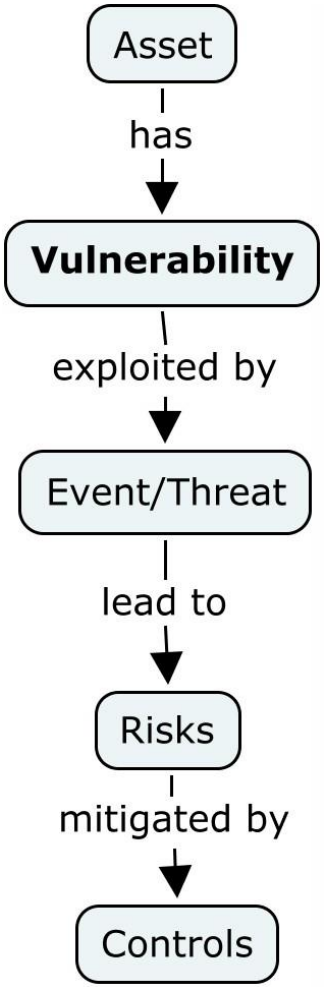
Assets in MetaGen-FRAME:

- Computational servers
- Web-services
- Local personal computers (PC)
- Software Tools (e.g. Taverna, Blast2GO, NGS QC Toolkit, BG7, MetaPhlAn, TAPyR)
- Databases / Data sets
- Workflow / Tasks



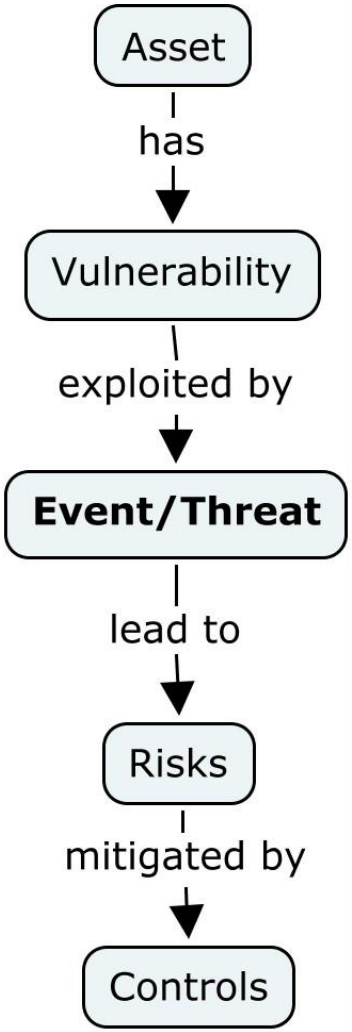
Main Vulnerabilities of MetaGen-FRAME:

- Unreliable storage hard drives
- Poor debugging capabilities of Taverna
- Lack of a long storage policy
- Bandwidth bottleneck
- Security breaches
- Missing criteria about data confidentiality
- Uncertainty on economic/organizational breakdowns (long-term project continuity...)



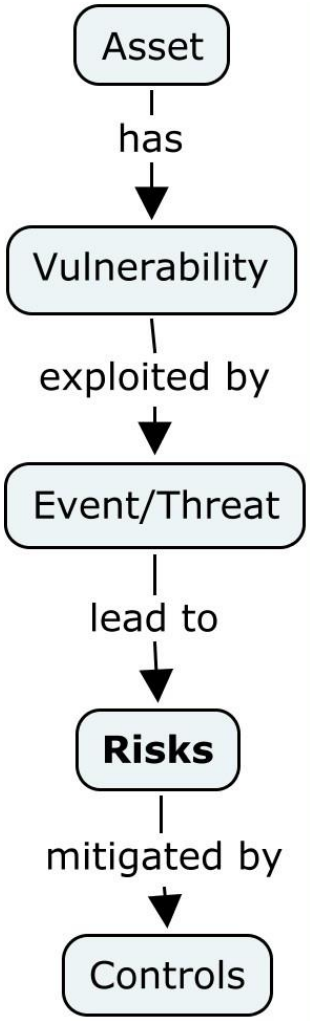
Potential Events in MetaGen-FRAME:

- Typing of wrong/incomplete inputs by the human operator
- Financial, legislative or organizational changes
- Hacker attacks
- Natural disasters
- Tool discontinuation
- Hardware failures (unexpected hardware failure of unknown cause and duration)
- Hardware maintenance (scheduled hardware failure with expected duration)



Examples of Risks in the MetaGen-FRAME:

R1 - Accidental change or deletion of digital objects
R2 - Insertion of wrong input values: Ex. 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 that are filtered in the data quality control task, therefore influencing all the following results...
R3 - Change of the external WS, National Center for Biological Information (NCBI), computational servers or local PC used causing their unavailability or failure
R4 - Loss of information due to communication failures
R5 - Loss of information and data traceability due to a media fault, compromising the workflow's recreation
R6 - Loss of metadata denying the representation of the output information to the user via Taverna
R7 – Lack of financial or legal requirements to preserve data
R8 - Obsolesce of the tools used in the workflow or in the NCBI or local PC
R9 – Ocurrence of an error that cannot be explained, leading to repetition of the whole experiment
R10 - Sharing of confidential data
R11 - Difficulties sharing the information and the workflow's execution details in other future scenarios
R12 – Stakeholder's lack of involvement

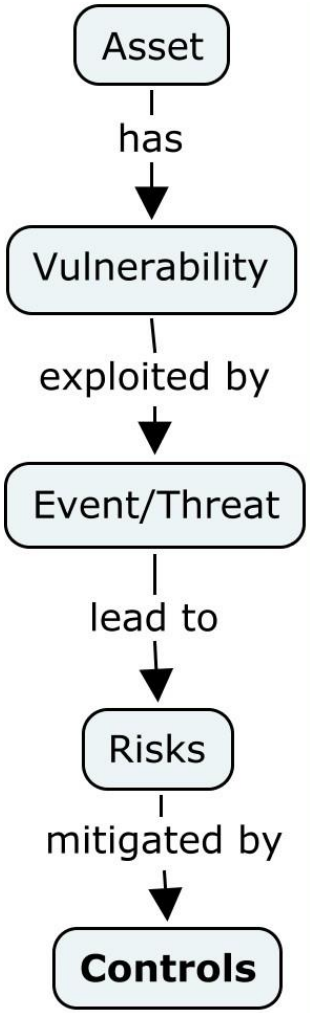


BTW, an overview of the relations we found between these risks and the typical sections of a DMP:

Section in the DMP	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

Possible controls for the MetaGen-FRAME:

Nº	Description of the Control
C1	Use several backup systems (local and remote) in the local PC, for example a system like shadow copy to store all the data and metadata
C2	Implementation of syntactic and semantic verification mechanisms to alert if the input don't have the correct format and content
C3	Improve the security measures in NCBI, computational servers, PCs and communication channels (better antivirus, encryption, firewall)
C4	Keep all the software and hardware components up to date
C5	Backup systems for the NCBI and computational servers
C6	Access other genome data bases such as the ones offered by the EMBL-EBI like ENA in case NCBI becomes unavailable
C7	Access to other computational servers if current ones fail (fall back)
C8	Create a replicated central storage to store, in real time, the execution results from the workflow using for example shadow copy
C9	Create a long term storage policy with a speciallized organization
C10	Anti-fire and earthquake measures in the NCBI and hosting of the computational servers



Reporting (an example of one possible technique): Risk Matrix in the MetaGen-FRAME (after an assessment by the project's experts)

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				

Stakeholders and roles from the MetaGen-FRAME project involved in this Risk Assessment:

- **Project sponsor (funding agency):**
  - FCT (Fundação para Ciência e Tecnologia)
- **Group Leader (project PI – consult, communicate risks and ultimate decision making):**
  - Prof. Ana Freitas
- **Project Manager (context establishment and RM decision proposal):**
  - Prof. Luís Russo
- (representative of the local) **Scientific team (definition and programing of the scientific workflows; implementation of local technical risk controls)**
  - Eng. Miguel Coimbra
- **Risk Owner (control and monitoring of the identified risks)**
  - … assumed by different partners (organizations) or people, depending of the vulnerbilities identified in the collaboration network…
- **Risk Expert (responsible for performing the RM analysis):**
  - Eng. Filipe Ferreira

Conclusions:

- **What is the actual thinking** - In e-Science, DM is crucial to govern the lifecycle of the data, so we all agree projects must have a DMP…
- **What was our analysis** - The DMP is not enough effective addressing risks in a formal and structured way!!!
- **What is our proposal** – The DMP must be complemented by a RMP considering the main state of the art in RM!!!
- **What we had proved** - The application of the generic RM principles to a specific e-Science project was possible and relevant…
- **What we are still doing** – We are developing a generic method for the creation of a RMP in e-Science:
  - Identify the common classes of stakeholders, assets, vulnerabilities, events/threats, risks and relevant controls
  - recommend the most suitable techniques for each step/phase, considering the domain (e-Science)

Examples of methods/techniques from the MetaGen-FRAME project (the generalization to e-Science still to be better validated…)

Step	Objective and Possible Method/Technique to Recommend
Risk Identification	Identify assets, vulnerabilities, threats and risks • Application of guidewords (**HAZOP technique**) • Application of what-if scenarios (**SWIFT technique**) • Application of check-lists (**Brainstorm sessions**)
Risk Analysis	Quantify risk likelihood and consequence (**SWIFT technique**)
Risk Evaluation	Rank risks according to their likelihood and consequence (**SWIFT technique** to produce **Risk matrix**)
Risk Treatment	Determine risk controls (**SWIFT technique**)

Thank you!!!
Questions?