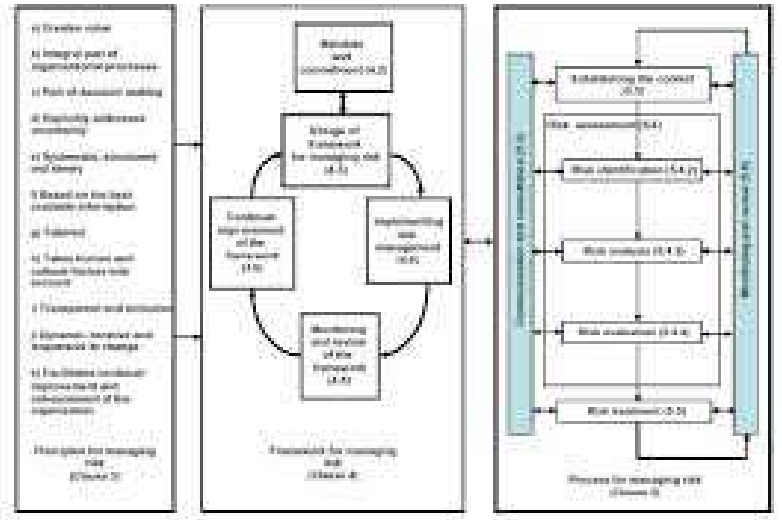
Data Governance in e-Science

Filipe Ferreira, Raquel Bairrão, Nuno Pradiante, 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**: Not all Data Governance (DG) concerns are properly addressed by DMP
- **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:

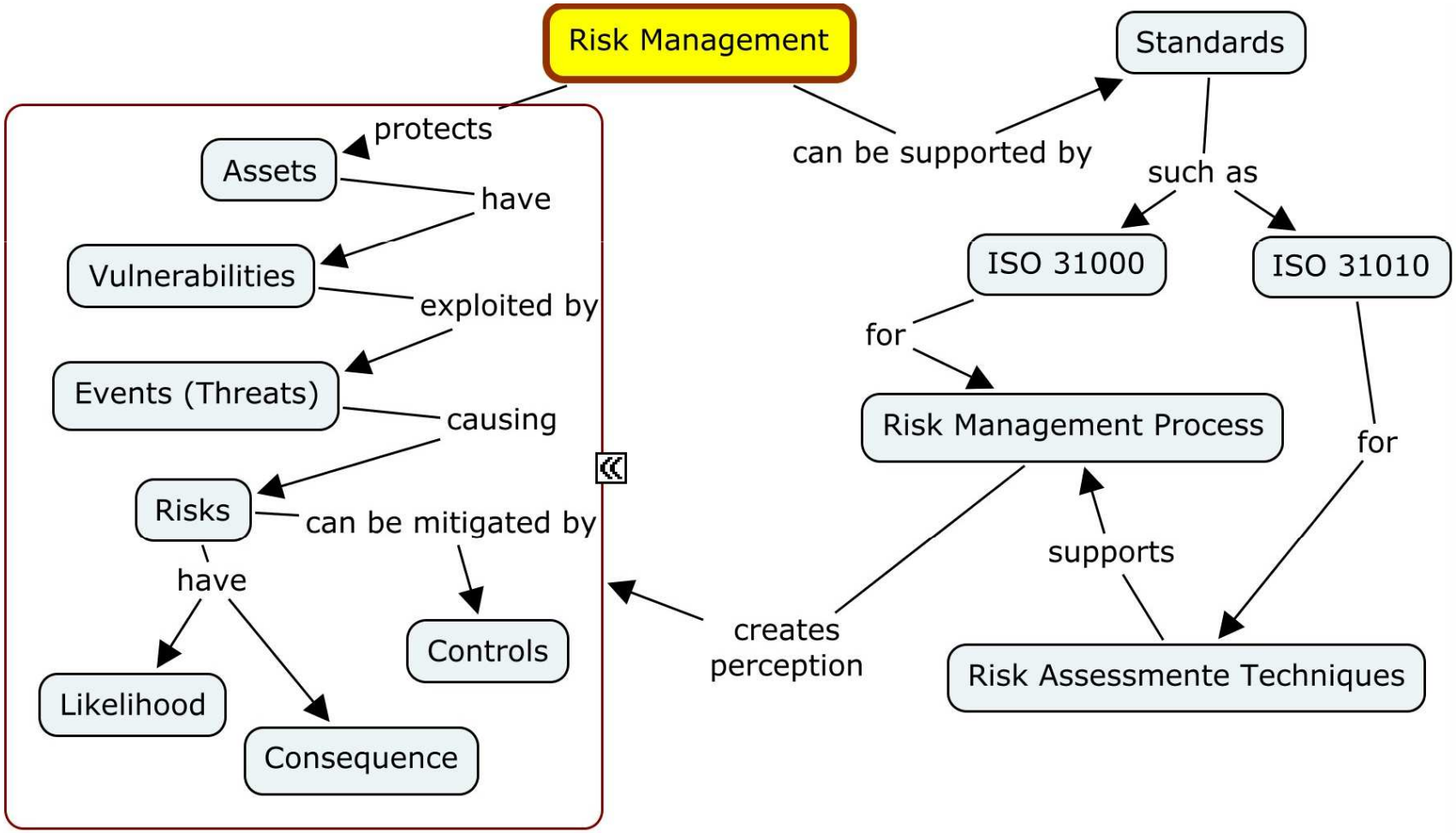
- Describe what and how data will be created, stored, preserved and shared in the project
- Guide researchers on how to reuse data in the project
- Records the project‘s decisions on Data Management

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:



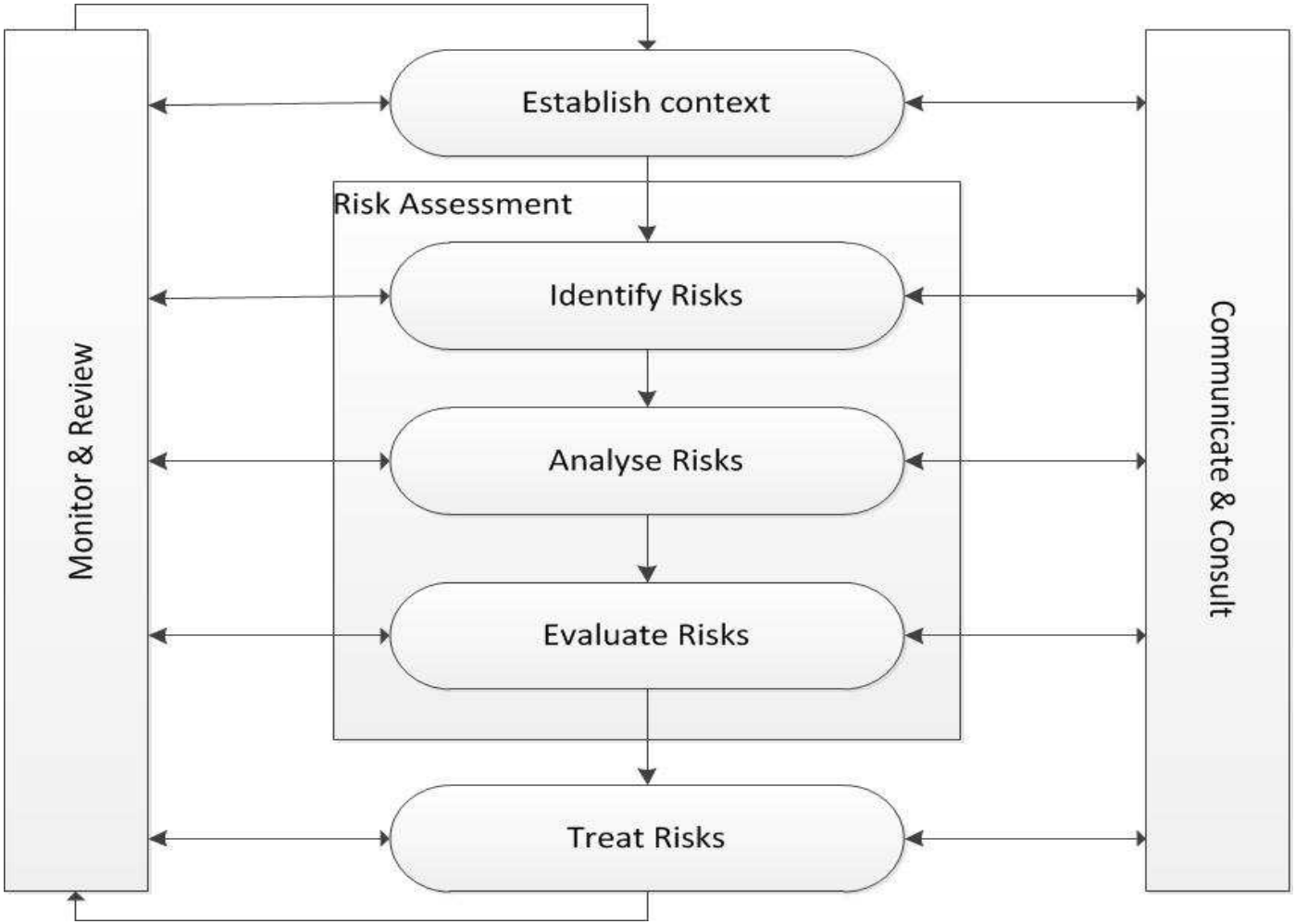
Thus, our hypothesis: A DMP should be complemented by a Risk Management Plan (RMP)!!!

- Propose a DG approach for e-Science projects through the joint utilization of DMP and Risk Management Plan (RMP);

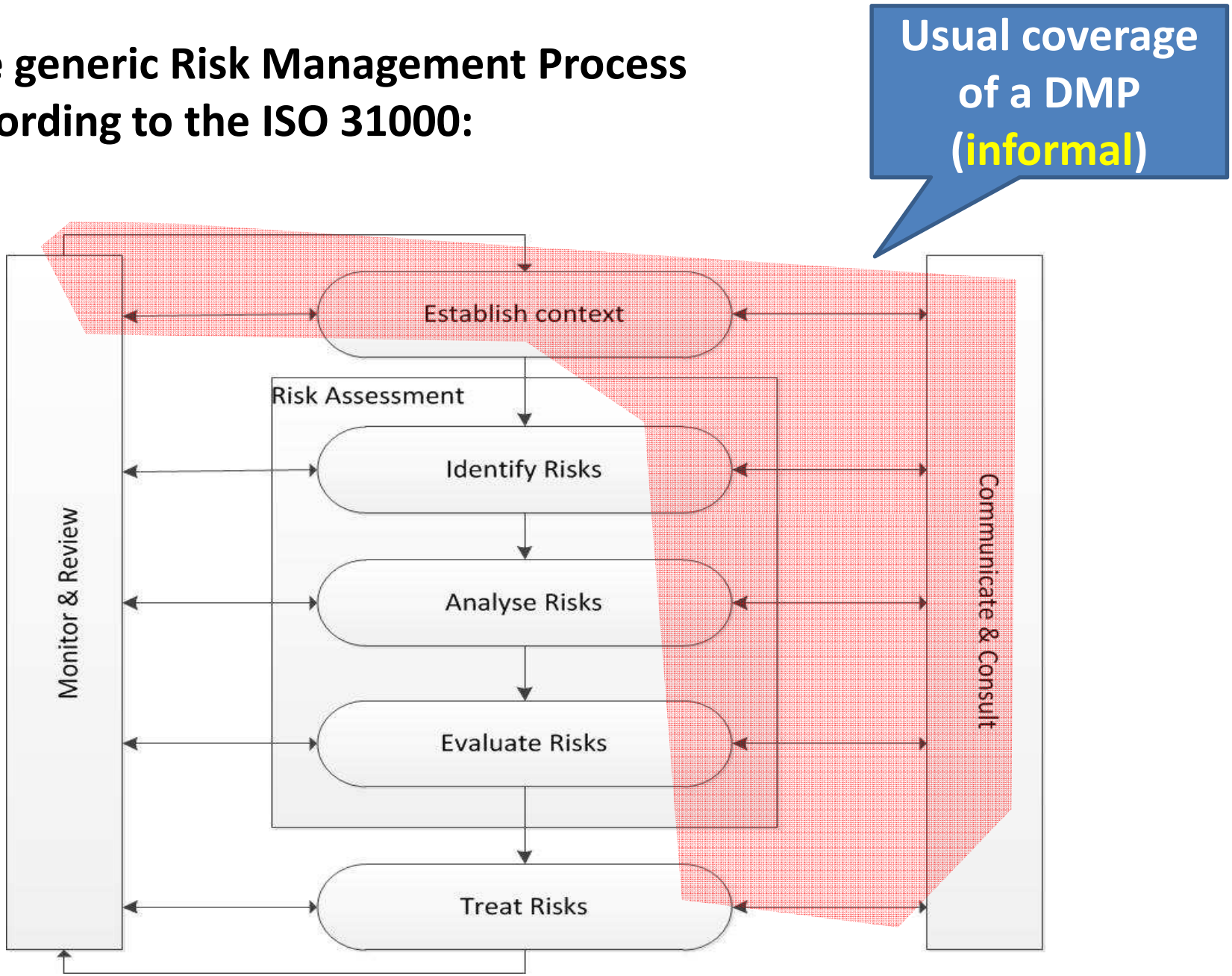
- Complement the DMP by a RMP that should support and justify the decisions and controls implemented by the DMP;

- Concluding: Tune the ISO 31000 method to develop a RMP for e-Science projects.

The generic Risk Management Process according to the ISO 31000:

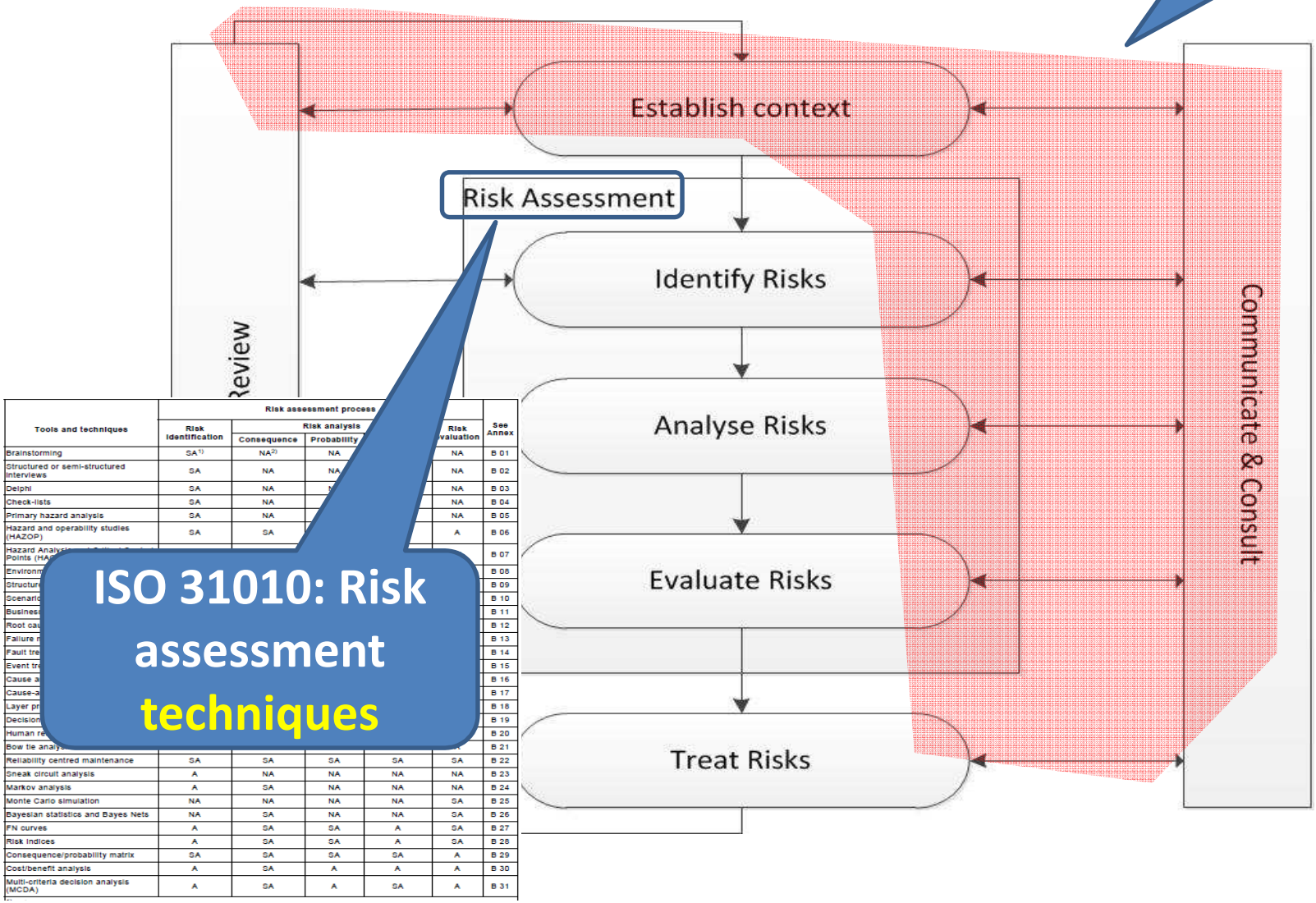


The generic Risk Management Process according to the ISO 31000:



The generic Risk Management Process according to the ISO 31000:

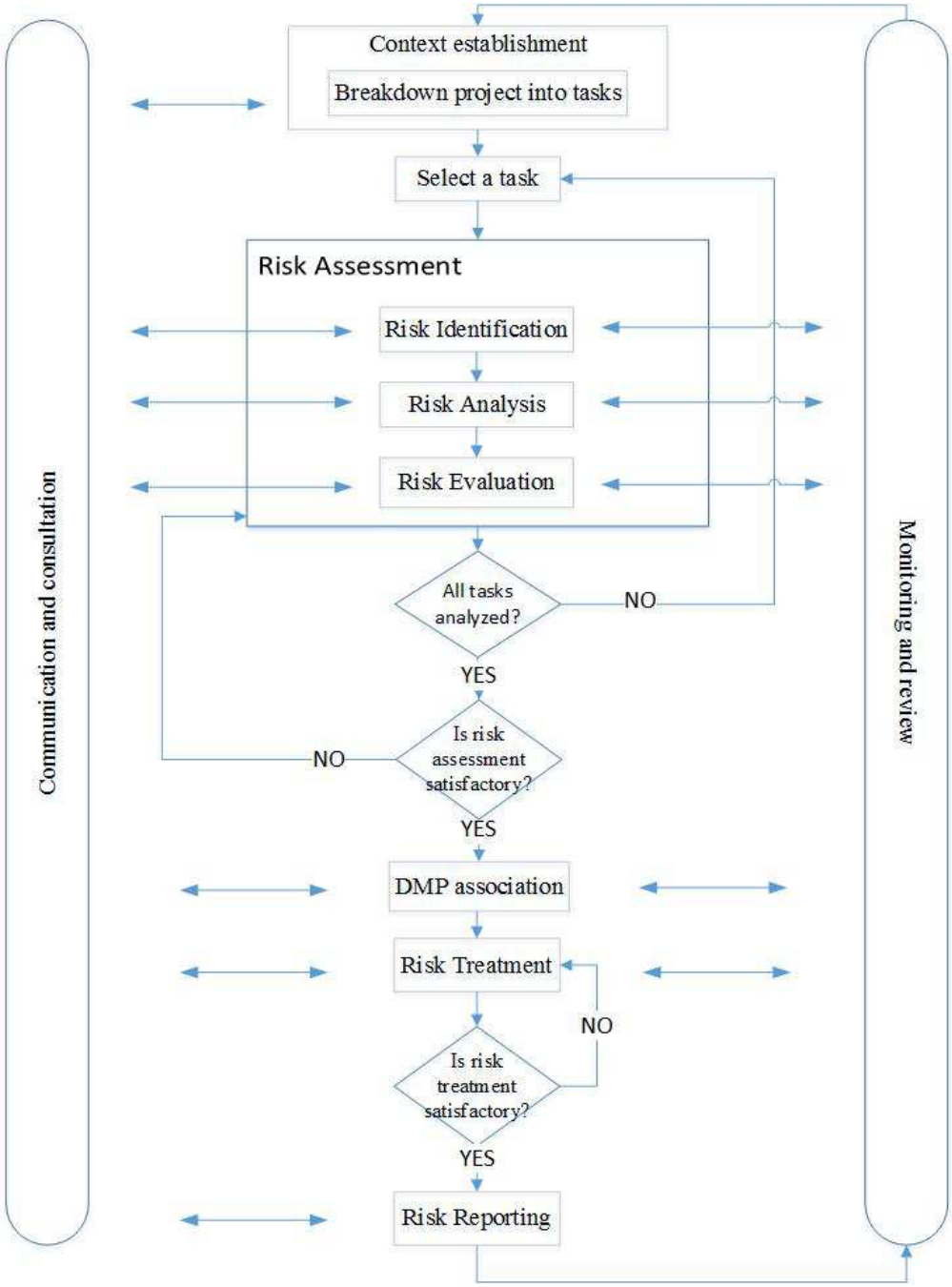
Usual coverage of a DMP (informal)



ISO 31010: Risk assessment techniques

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	N		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 Analysis and Critical Control Points (HA						B.07
Environm						B.08
Structure						B.09
Scenario						B.10
Business						B.11
Root cau						B.12
Failure r						B.13
Fault tre						B.14
Event tre						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

Profiled method for e-Science:



Typical roles and responsibilities in a research project:

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

Typical roles and responsibilities in a research project:

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

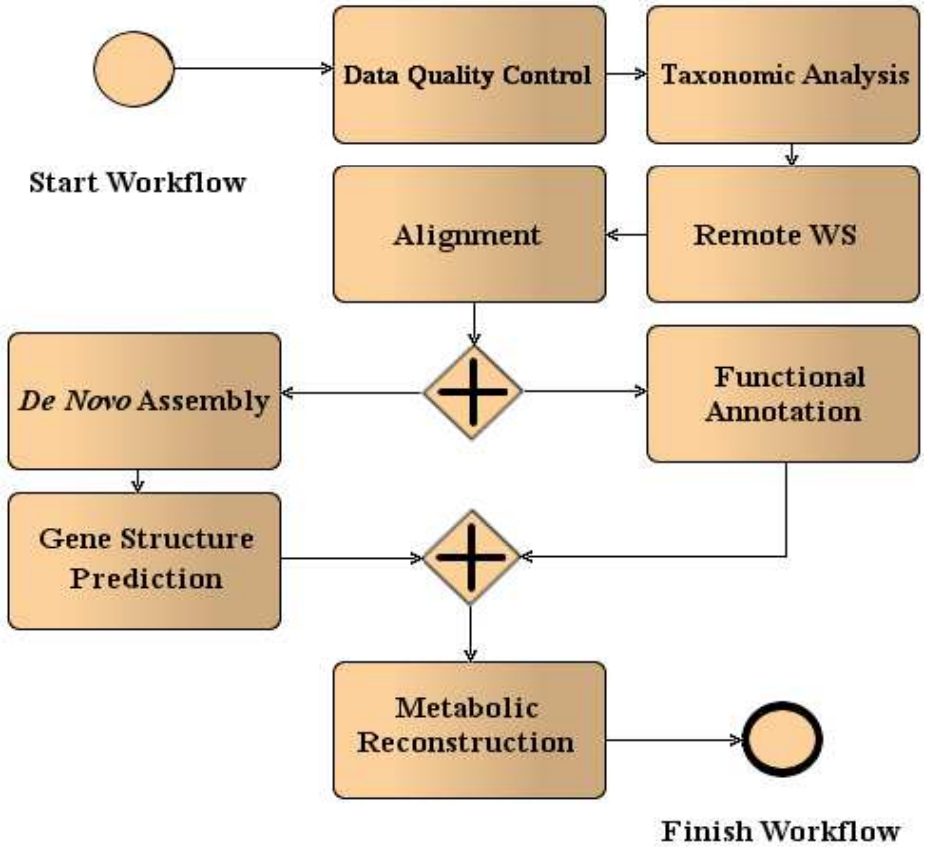
Risk Expert: This is where we identify an opportunity for expert librarians/data curators…

Proposed **skills for risks expert**:

- **Core:**
  - **Data Management:** know data Management principles
  - **Risk Management:** know **PROCESSES** to how to identify, analyze, evaluate and treat risks
  - **Security:** background on breaches that threatens data
  - **Metadata:** Know how to produce, collect and secure metadata
  - **Advocacy, copyright and intellectual property rights:** to mitigate risks concerning data dissemination
- **Complementary**
  - **Technical skills:** determine technology or infrastructure risks and controls
  - **Data value:** assess the value of the data objects worth protecting
  - **Domain Knowledge:** Engineering or scientific domain knowledge

An Example – The MetaGen-FRAME project:

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

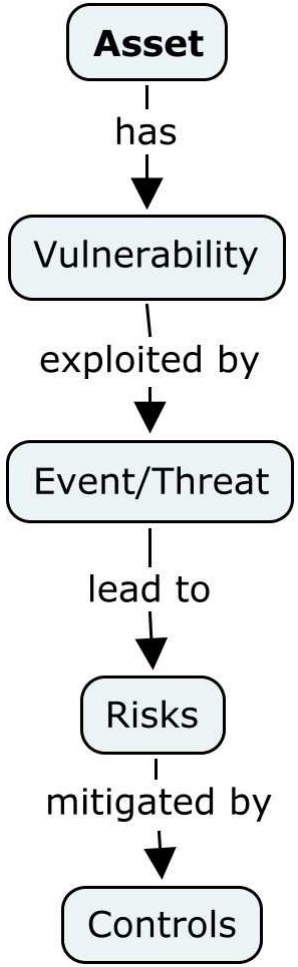


Stakeholders and roles in the MetaGen-FRAME project involved in this Risk Assessment:

- **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 vulnerabilities identified in the collaboration network…
- **Risk Expert (responsible for performing the RM analysis):**
  - Eng. Filipe Ferreira

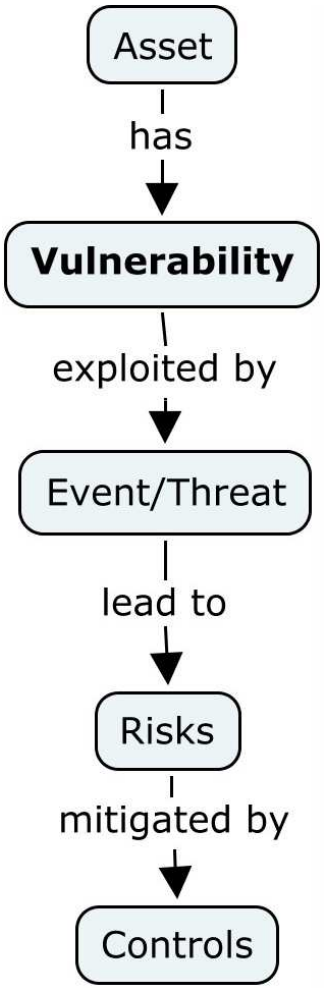
Assets in MetaGen-FRAME:

#	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



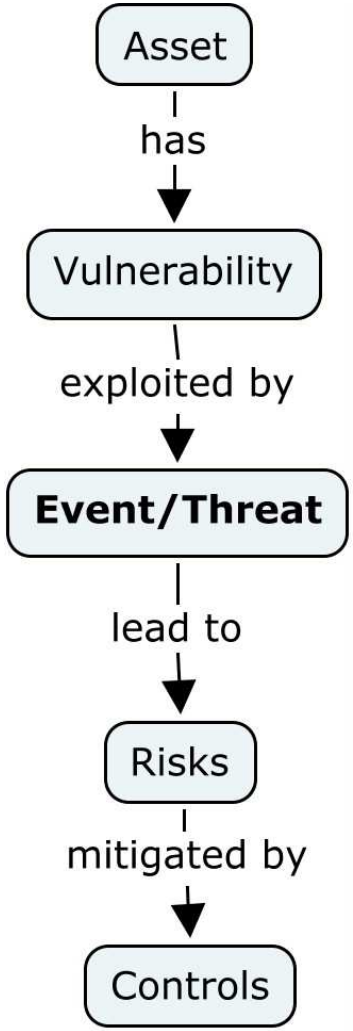
Main Vulnerabilities of MetaGen-FRAME:

#	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



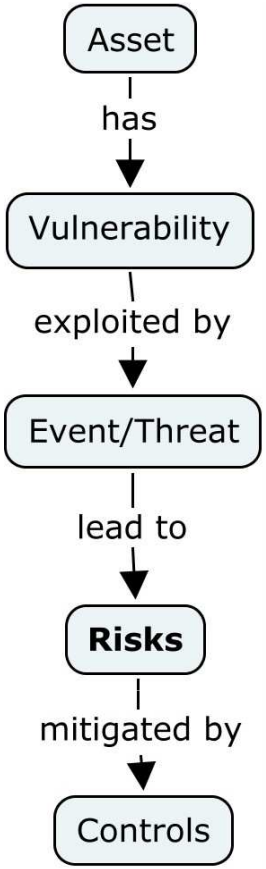
Potential Threatening Events in MetaGen-FRAME:

ID	Events
E0	Abandonment of a stakeholder
E1	Creation and utilization of new technologies that increase substantially the quantity of data and metadata generated
E2	Data managed by biologist with no DM experience
E3	Data used inappropriately
E4	Errors on search, access and delivery of preserved data
E5	Financial, legislative or organizational changes
E6	Hacker attack
E7	Hardware obsolesce
E8	Human errors
E9	Infrastructure failure
E10	Infrastructure maintenance



Risks identified in the MetaGen-FRAME:

ID	Risks	Event	Assets
R0	Copyright infringement by sharing confidential information	E14	A2
R1	Copyright infringement by claiming ownership of data sets produced by others	E16	A2
R2	Difficulties sharing the information and the workflow's execution details in other future scenarios due non successful extrapolation of data's metadata	E12	A2, A7
R3	Difficulties sharing the information and the workflow's execution details in other future scenarios due to errors on search, access and delivery of preserved data	E4	A2, A7
R4	Difficulties sharing the information and the workflow's execution details in other future scenarios due to data being used inappropriately	E3	A2, A7
R5	Inapt, incomprehensible or incomplete data, metadata and documentation by being managed by inexperienced personal	E2	A2, A3



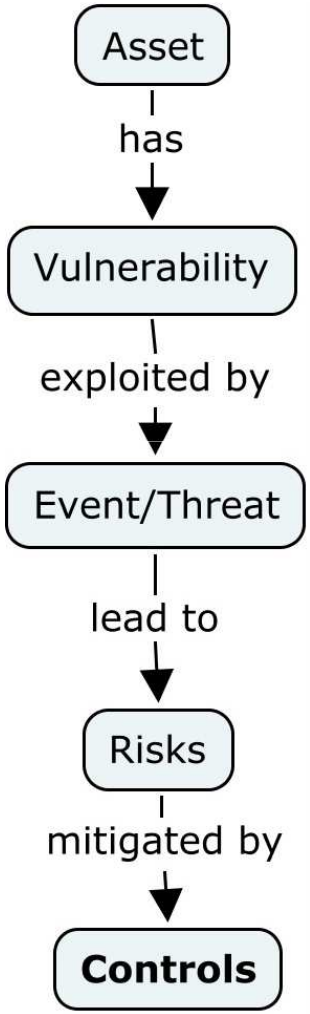
Likelihood	Consequences: 1	3	5	7	9
0.9					
0.7				R4 R26	R7 R14
0.5		R21	R19	R2 R3 R11 R24 R28	R6 R12 R17 R30
0.3			R0 R1 R8	R9 R20 R22 R23 R25	R5 R10 R13 R15
0.1				R27 R29	R16 R18

Association of the Risks to the Sections of a standard DMP

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

Examples of Controls for the MetaGen-FRAME:

ID	Controls	Type	Vulnerabilities, Risks, Events
C0	Anti-fire and earthquake measures in the NCBI and computational servers	Consequence Likelihood	R27 E11
C1	Backup system	Exposure Consequence Likelihood	V9 R10, R11, R12, R13, R14, R15, R16, R17, R18, R19, R22, R23, R24, R25, R26, R27, R28, R29 E3, E4, E6, E8, E9
C2	Create a long term storage policy (recovery management Plan)	Exposure Consequence	V6 R9, R12, R13, R14, R15, R16, R17, R18, R22, R23, R24, R25, R26, R27, R28, R29, R30 E11,
C3	Create a protocol defining the workflow execution properties or create additional metadata, creating stronger bonds between the biological results	Exposure Consequence	V12 R2, R3, R4, R7, R19
…	…	…	…



Overall Risk Reporting:

Risk Category	Risk	Control
Financial, organizational or workflow (strategic)	R9	C2, C3, C5, C6
Legal	R0, R1, R10, R20	C4
Data	R2, R3, R4, R5, R6, R10, R11, R12, R13, R14, R15, R16, R17, R18, R19	C1, C2, C11, C12, C14
Operational (Hardware/software)	R19, R22, R23, R24, R25, R26, R27, R28, R29, R30	C0, C1, C7, C8, C9, C10, C13, C14, C15
Staff/stakeholders (human)	R7, R8, R21	C6

Conclusions:

- **What is the actual thinking** - In e-Science, DG and DM are crucial to govern the lifecycle of the data, so it is nowadays accepted projects must have a DMP…

- **What was our analysis** - The DMP is not enough for addressing DG concerns 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!!!
  - Data managers with skills in risk management are required for this purpose!!!

Thank you!!!
Questions?

Filipe Ferreira -> filipe.ferreira@ist.utl.pt