secure
sba-research.org

Digital Preservation of a Process and its Application to e-Science Experiments

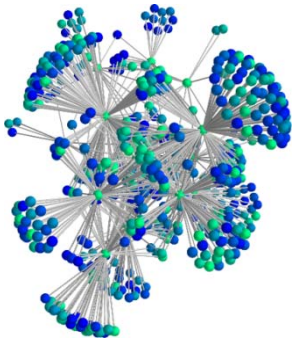
Stephan Strodl

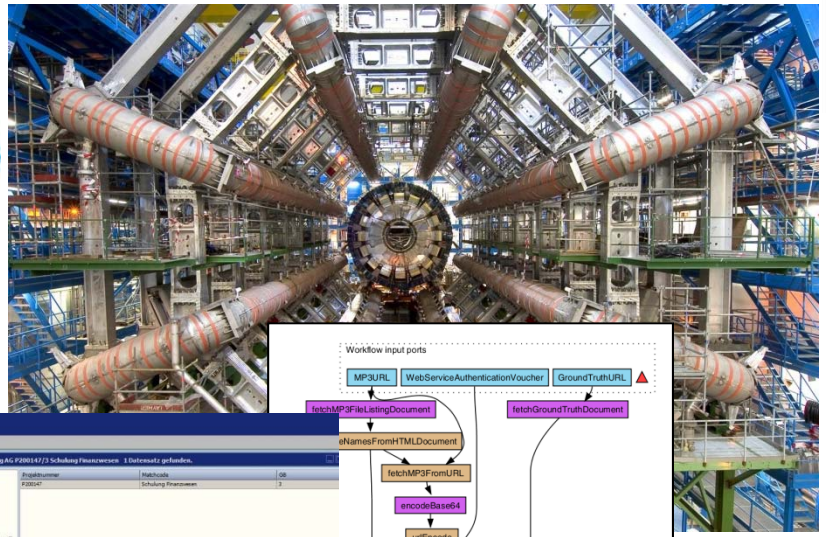
iPRES 2013

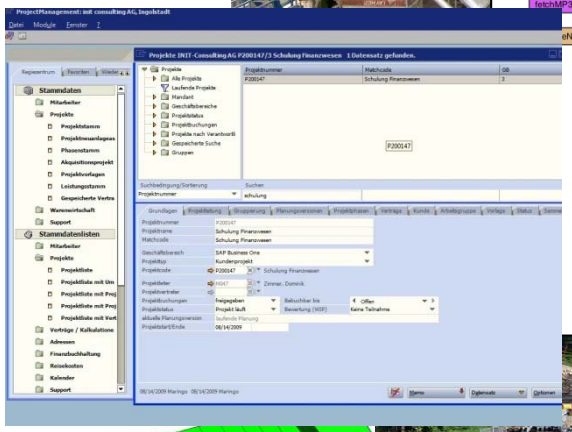
Lisbon, Portugal

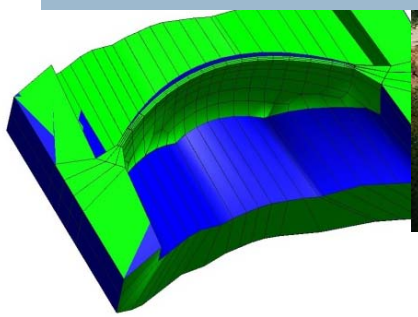
Beyond data

- Science
- E-Health
- Infrastructure
- Finance
- Insurance
- Aviation industry
- …



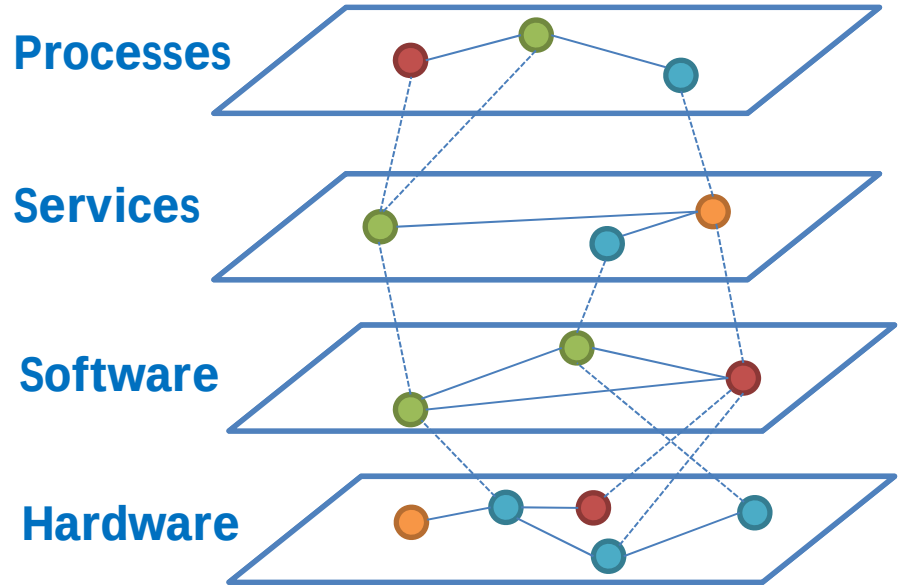






Process

- Data
- Activities
- Organisation
- SW-System
- Infrastructure
- Legal and contractual obligations
- …



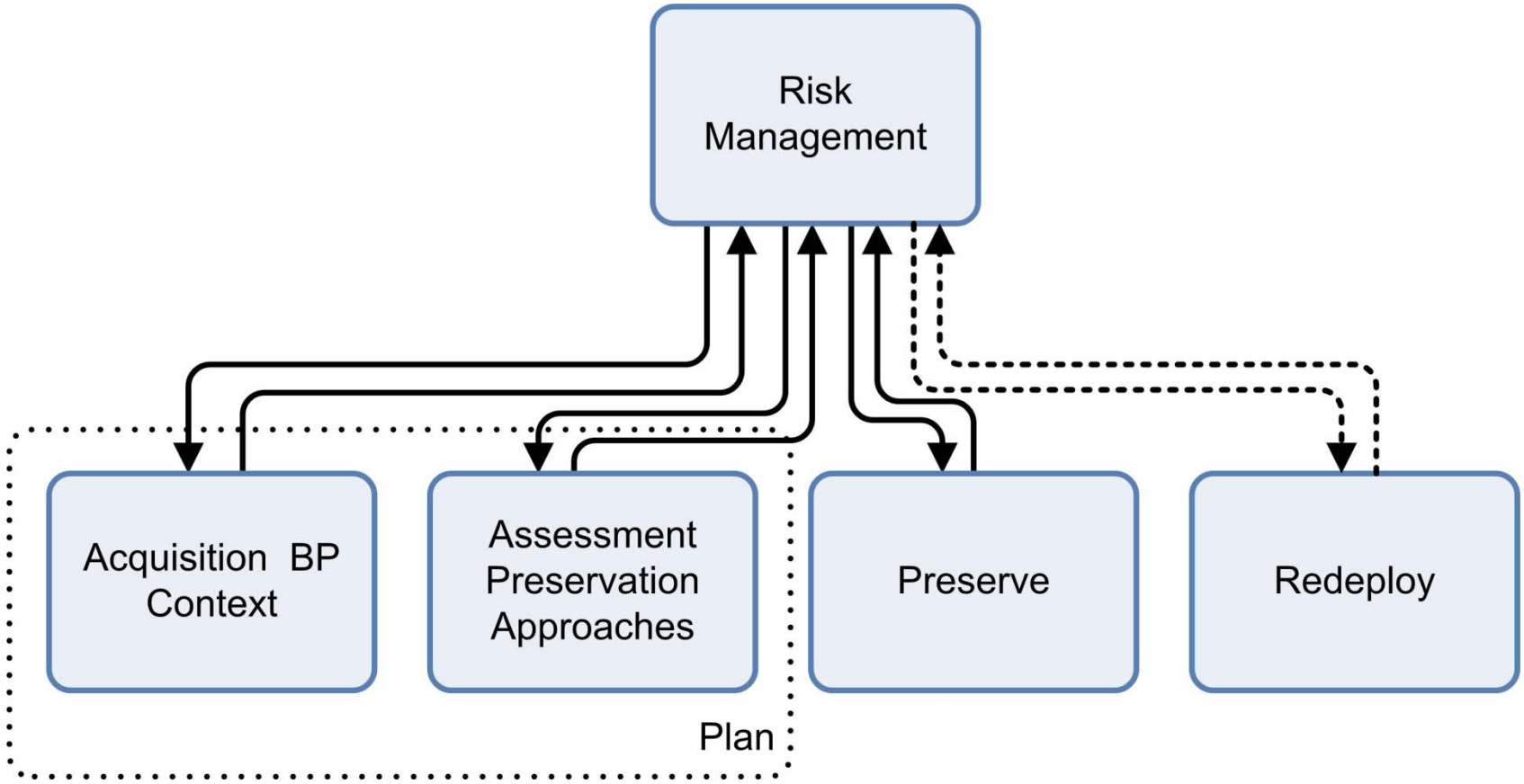
Long term resilience

- Obsolesces
  - Data
  - Software
  - Hardware
- 3rd party services
  - Cloud computing
  - Web Services
  - Customised software
- Documentation and specification

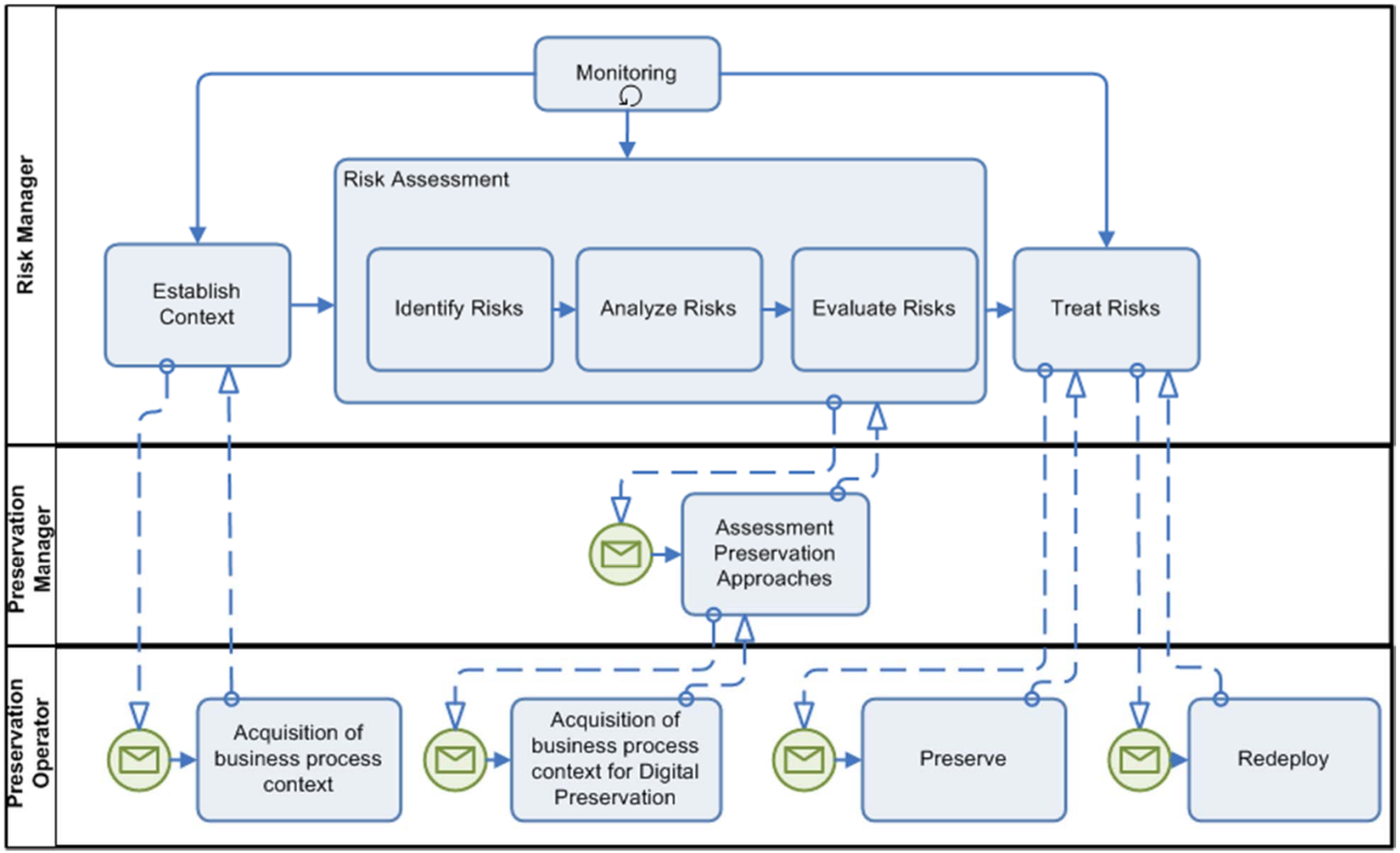
Process Framework

- Process to digitally preserve a process
- Framework for three phases:
  plan, preserve and redeploy
- Guidance for activities
- Flexibility, domain-independent
- Input, output, references, responsibilities
- Driven from Risk Management perspective

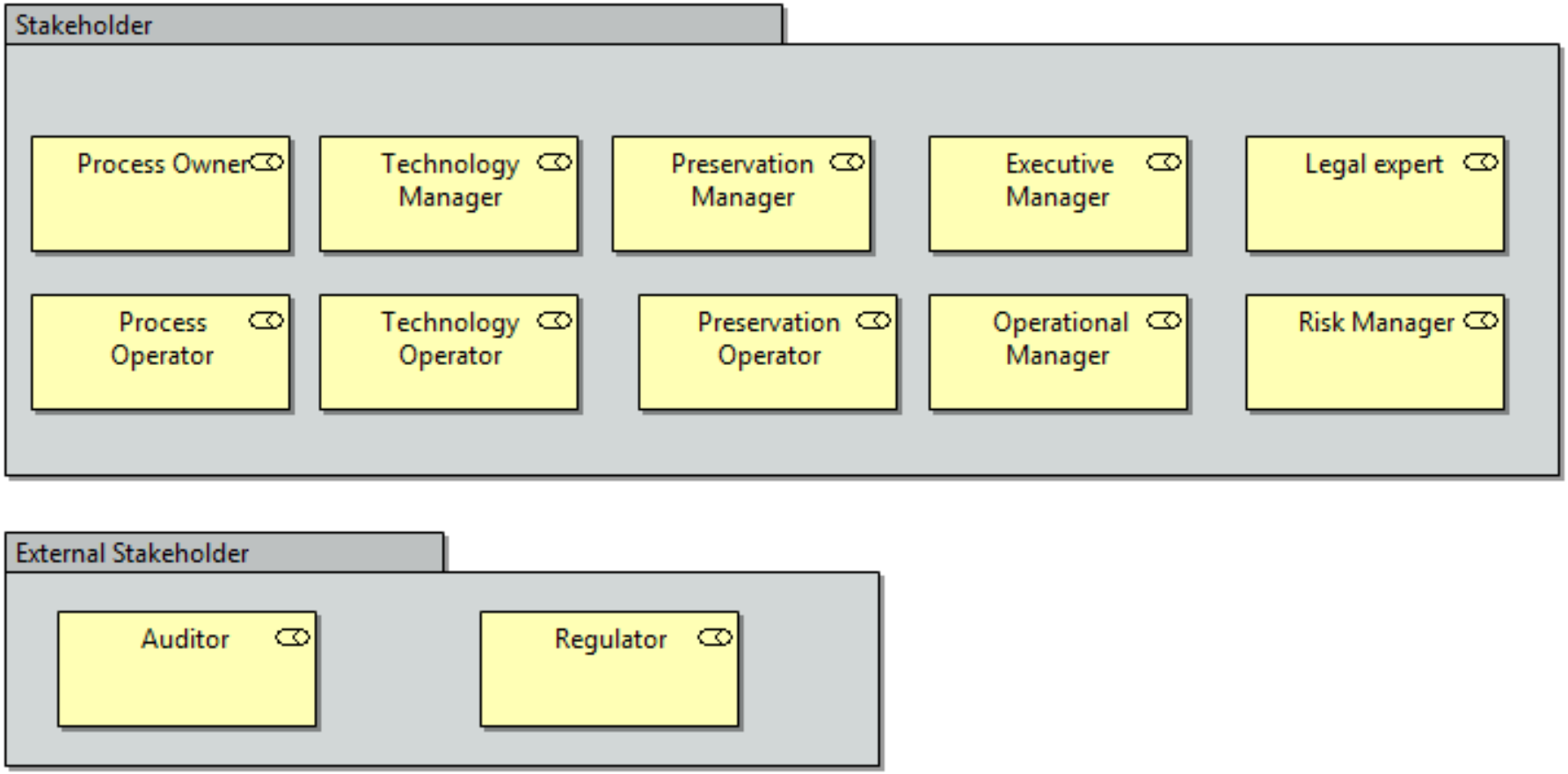
TIMBUS Process Framework



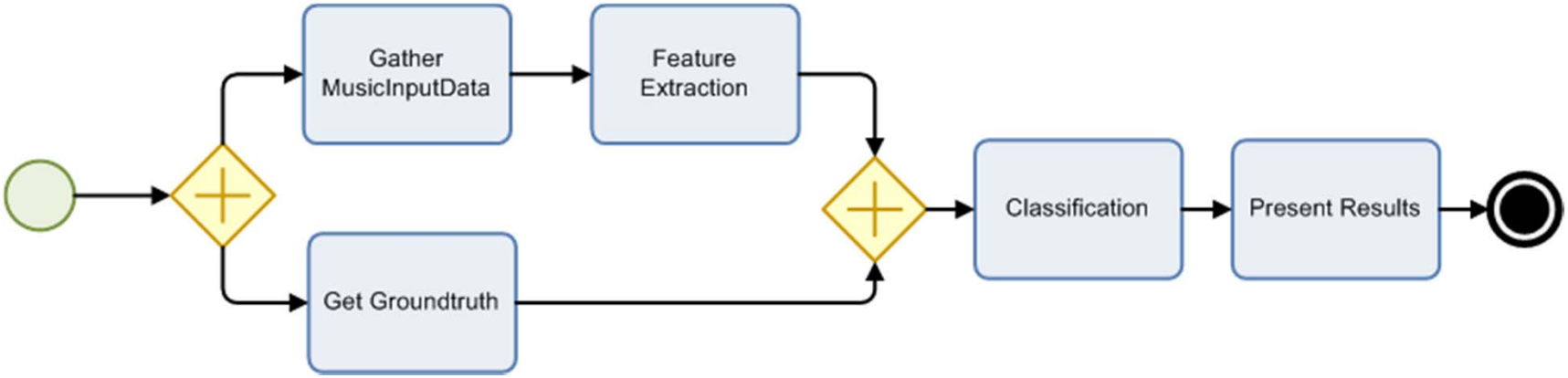
TIMBUS Process Framework



TIMBUS Stakeholder



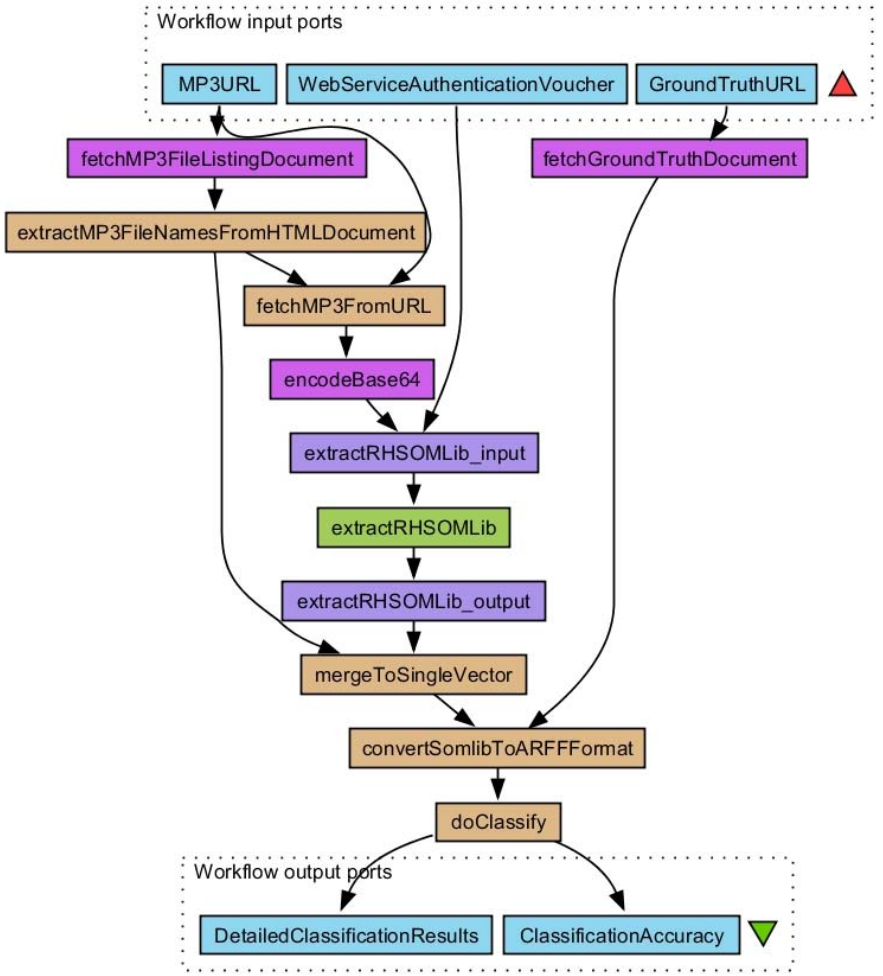
Music Classification Workflow



- Workflow: classification of music into predefined set of genres
- Learns a machine-learning model from given training data (i.e. data with manually assigned class/genre)
- Predicts genre for previously unseen data

Music Classification Workflow

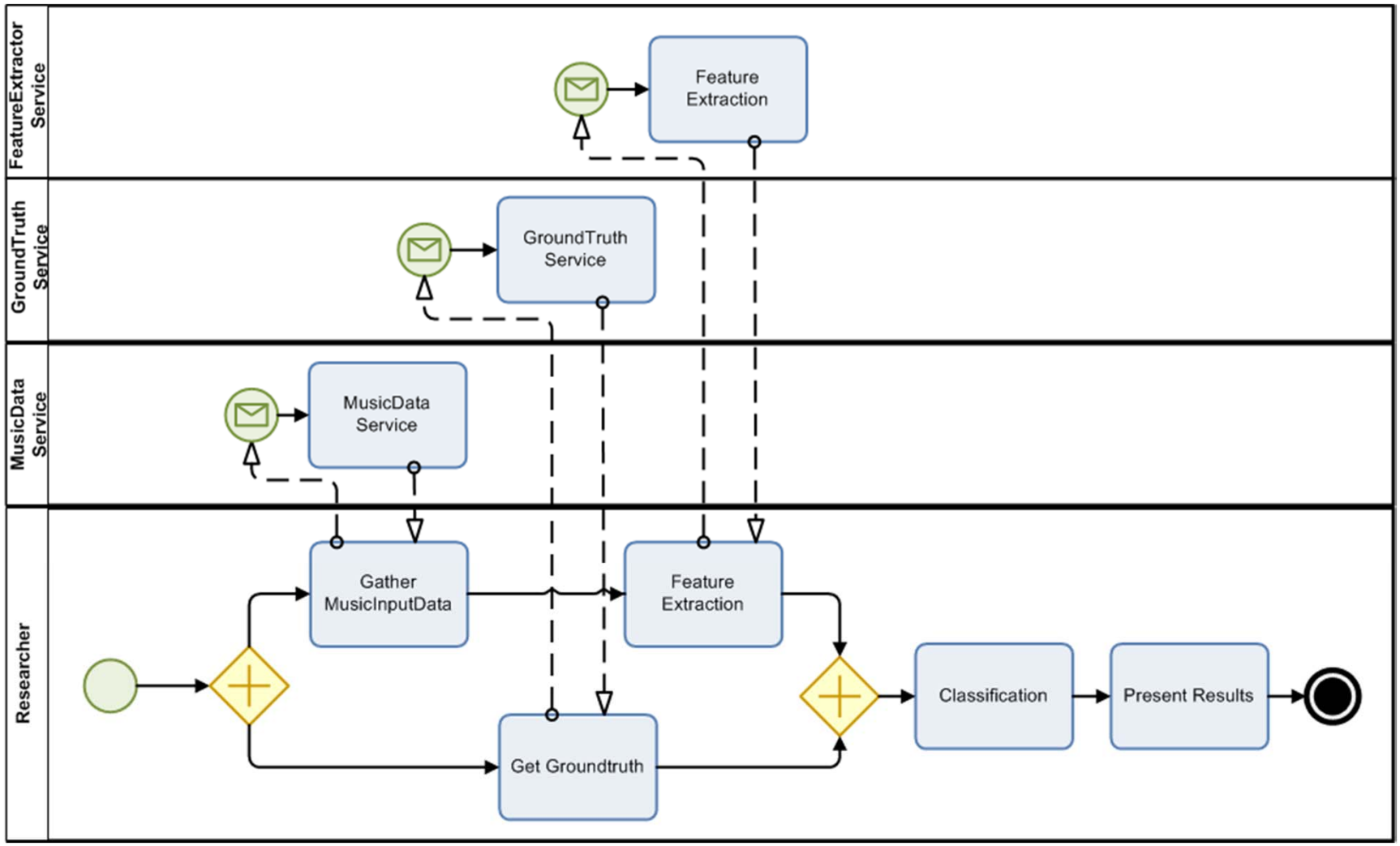
- Researcher at a research institution
- Input: Music (e.g. MP3 format)
- Training data, i.e. music for which the genre is known a-priori
- Output: Classification of music, e.g. into genres
- Technical infrastructure
  - External and internal services



Music Classification Workflow

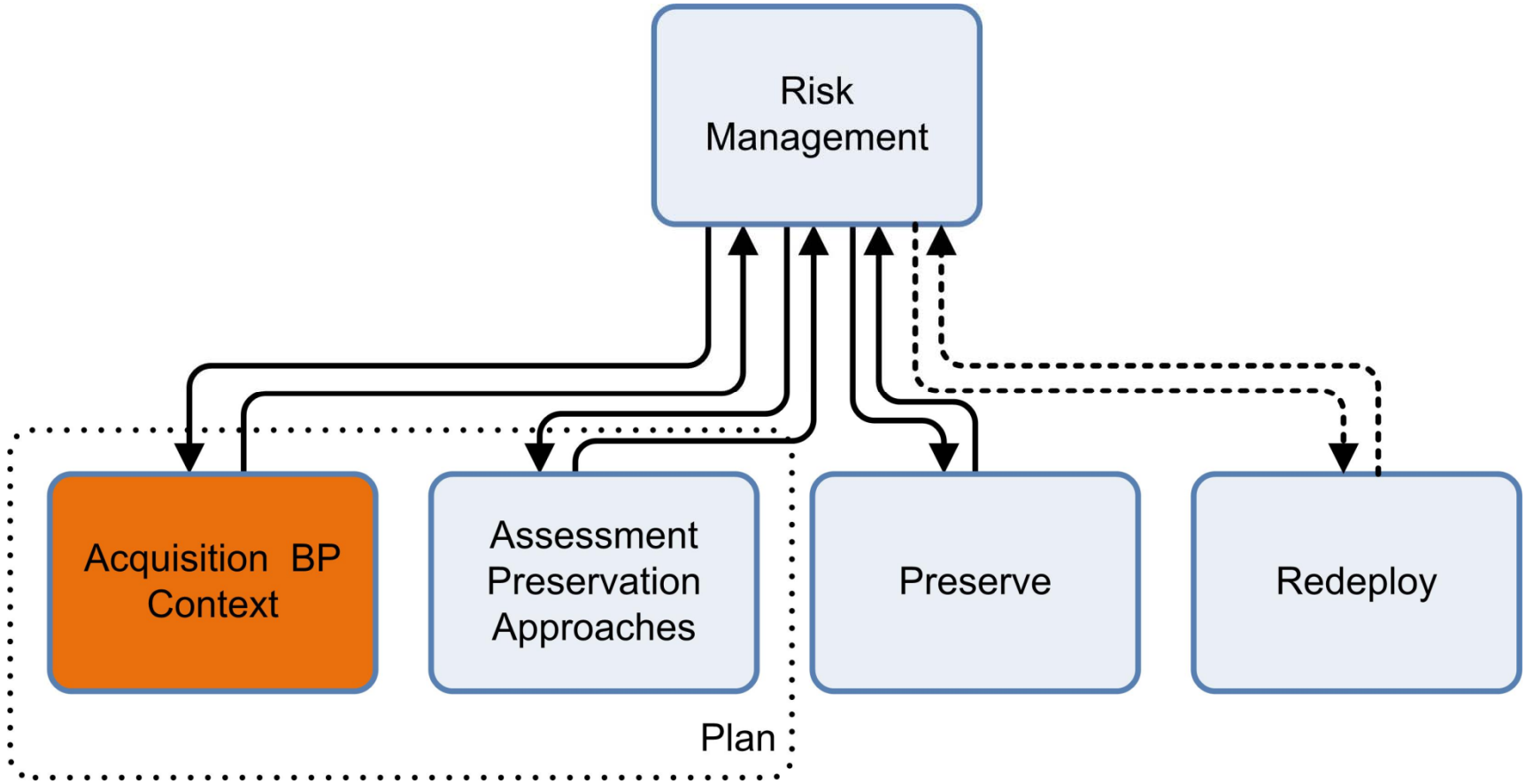
- Technical implementation
  - 3 External services:
    - MusicData Service
    - GroundTruth Service
    - FeatureExtractor Service
  - Internal services
    - Classifier Service
  - Taverna Workbench 2.4.0
    - Linux/Windows
  - Java 1.7

Music Workflow



Planning Phase

TIMBUS Process Framework

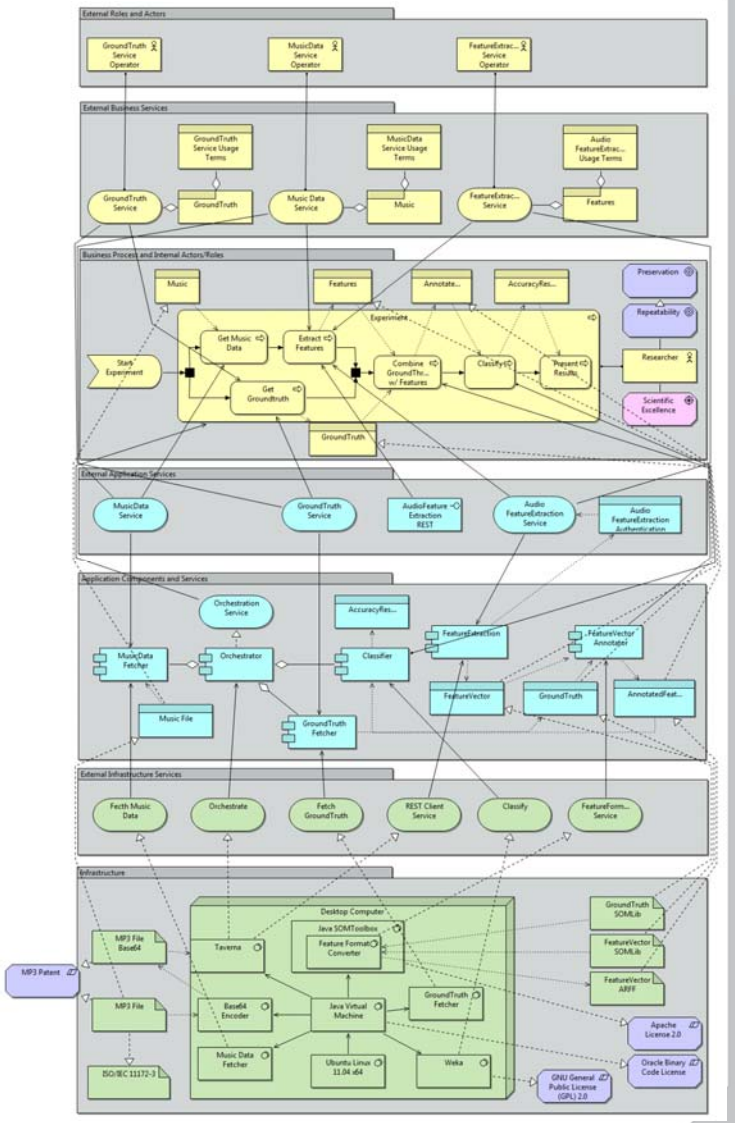


Acquisition business process context

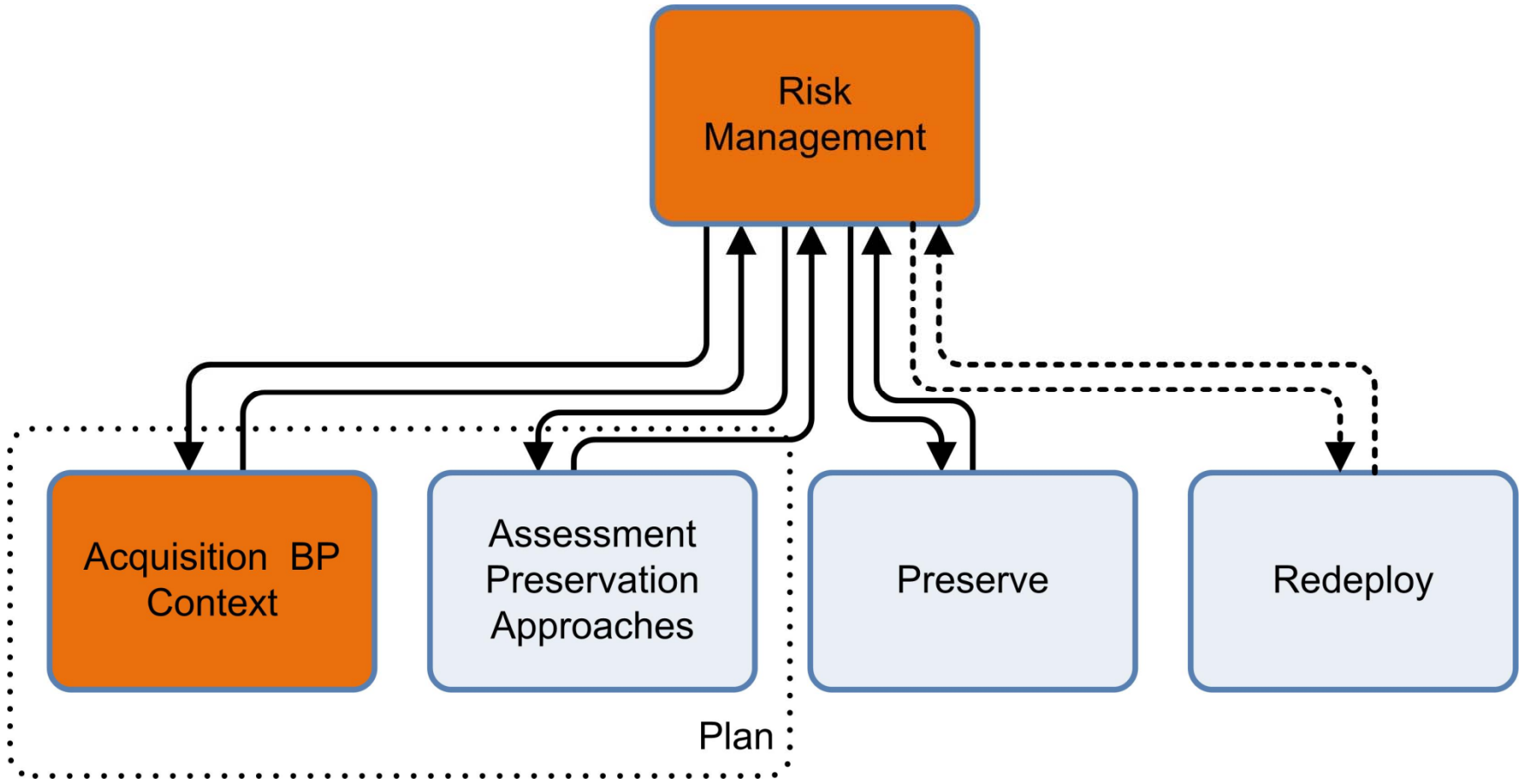
- Business process specification
  - BPMN, Taverna model, scientific papers
- Identify stakeholders
  - Internal
    - Researcher
    - Research institution
  - External
    - MusicData Service Operator
    - GroundTruth Service Operator
    - FeatureExtractor Service Operator
- Context acquisition

Context Model and Views

- External business service
- Workflow
- Business objects
- Motivation
- Policies
- Software Services
- Files and formats
- Licences and patents



TIMBUS Process Framework



Risk Management

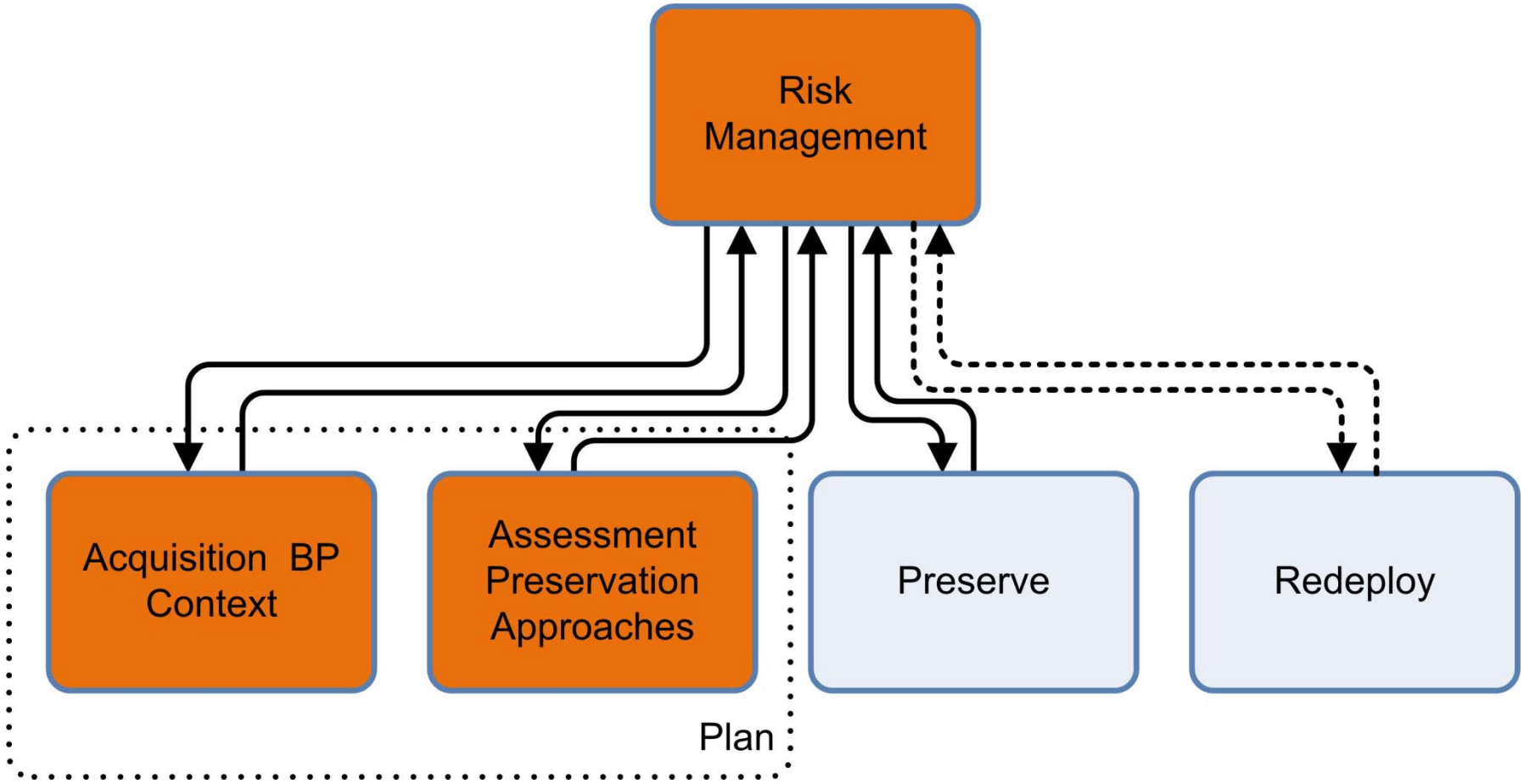
- Motivations
  - Trust
  - Share and reuse
  - Reproduction of results
- Risks
  - Use external services
  - Distributed documentation
  - Lack of documentation of experiments
  - Loss of experiment results
  - Technological dependencies

Risk Management

Consequences

- Missing validation of results in the future
- Loss of trust on existing research data
- Loss of expert knowledge
- Loss of scientific results
- Loss of scientific reputation
- Financial loss

TIMBUS Process Framework



Assessment Preservation Approach

- Creation of Preservation Plan
- Identification and evaluation of different preservation approaches
- Preservation requirements
  - use case scenarios
    - re-execution, re-use
  - significant properties
- Abstraction and generalisation
  - process logic

Define Preservation Plan

- Identification, evaluation and comparison of preservation plans
- Preservation plan
  - Acquisition procedure
  - Preservation procedure
  - Redeployment procedure
  - Verification and Validation procedure
- Process → orchestration of services
- Combination of preservation strategies
- Preservation of dependencies and relationships

Preservation and supporting strategies

- Metadata and documentation
- Migration
  - File formats
  - Storage media
  - Alternative services
    - Open source service
    - In-housing of services
- Emulation

Preservation and supporting strategies

- Virtualisation
  - Clone
  - Build
- Re-build of SW systems
- Mock-up of software system
- Software Escrow

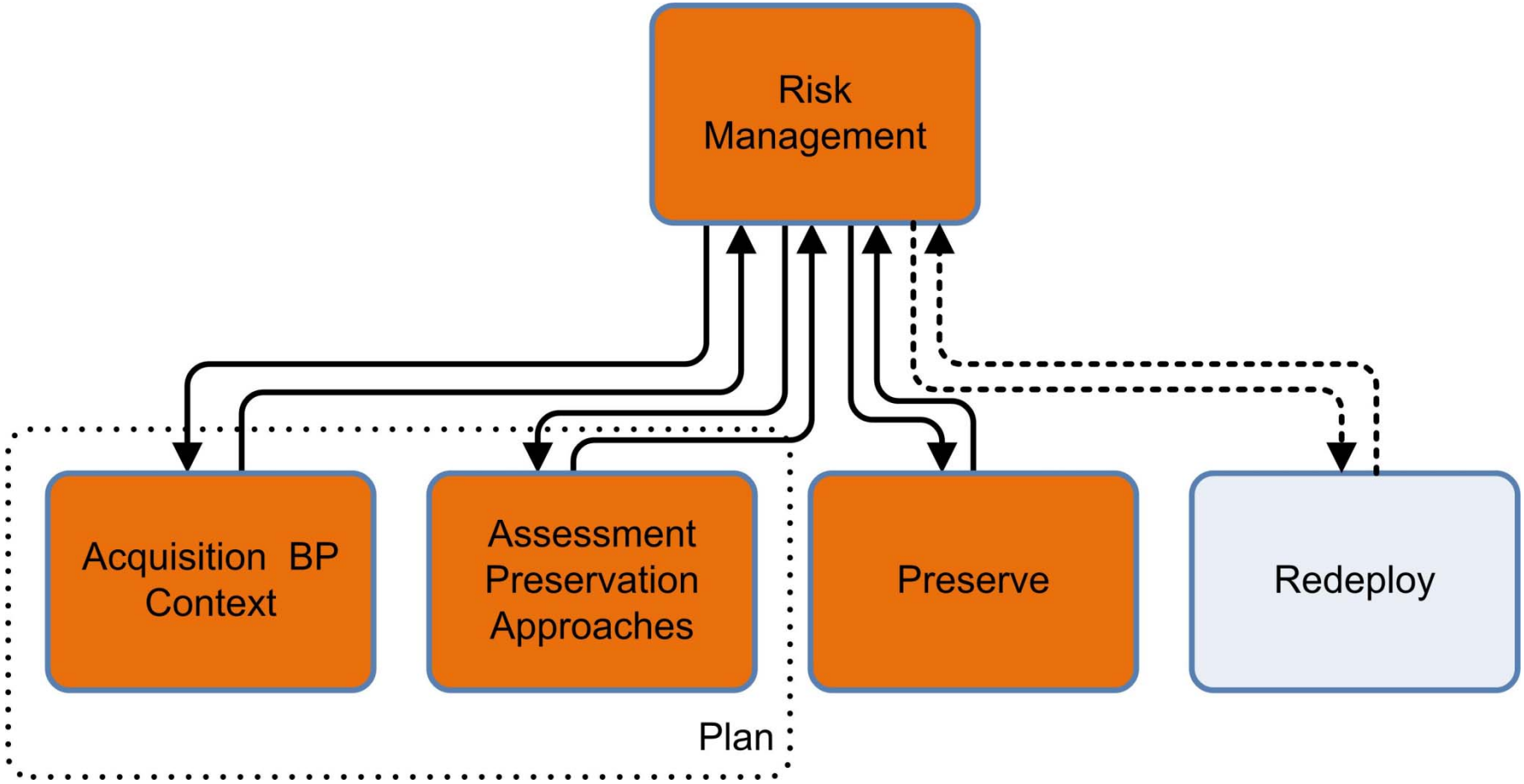
Define Preservation Plan

- Client
  - Virtualsiation
- Music Data Provider
  - Migration to local service
- FeatureExtractor Service Provider
  - Mock-up
  - Escrow
- GroundTruth Service Provider
  - Mock-up
  - Escrow
  - Migration to local service
- Documentation
  - Migration

Evaluate and Analyse

- Virtulisation of client → Oracle VirtualBox
- Music Files → local service → Apache Sever
- Feature Extractor → mockup service
- GroundTruth → local service → Apache Sever
- Migration of documentation
  - PDF to PDF/A using Adobe Acrobate 9.5.4
  - HTML

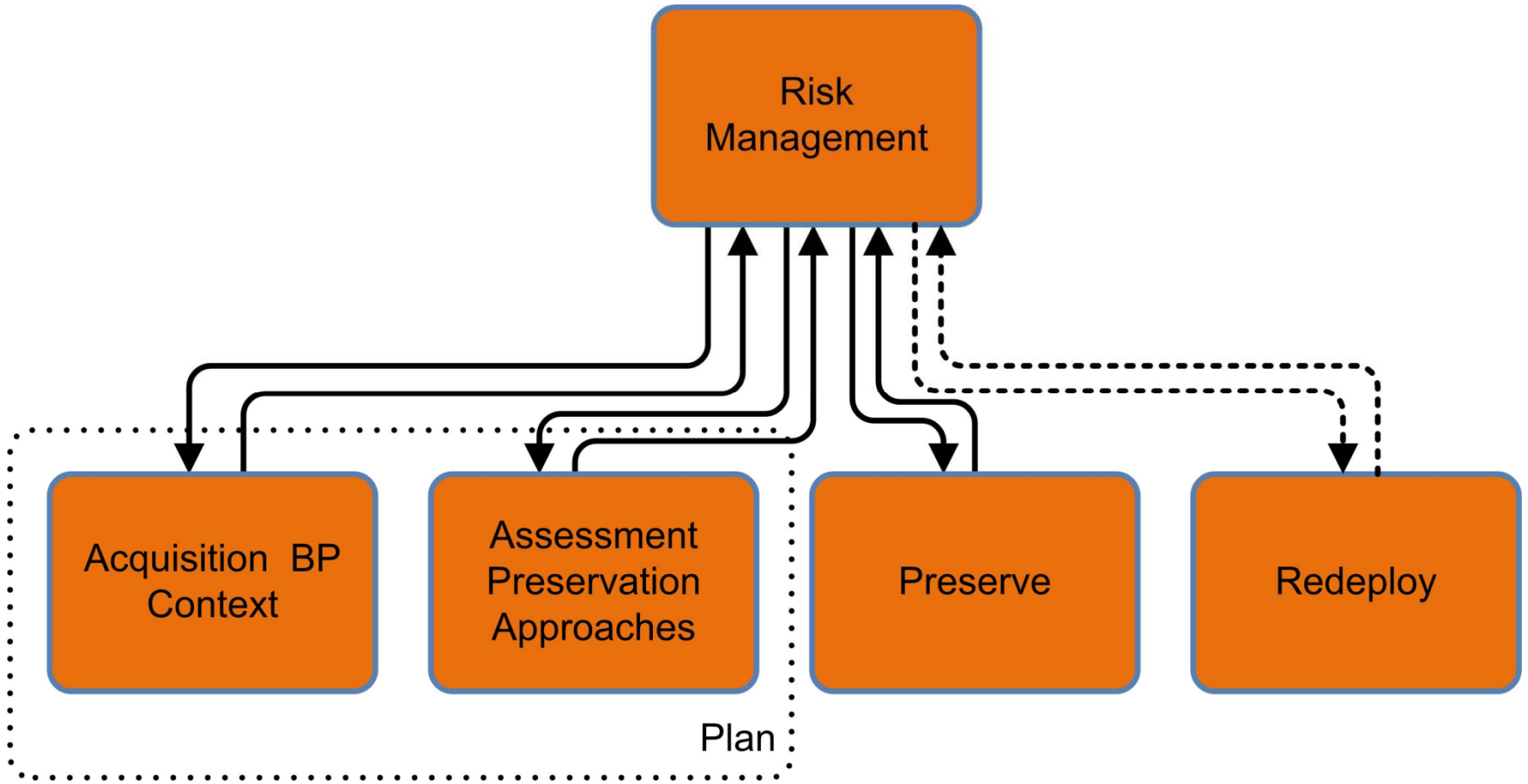
TIMBUS Process Framework



Preserve

- Execution of preservation plan
  - Creation of VM
  - Migration of service (GroundTruth, MusicData Service)
  - Mock-up of external service
  - Migration of documentation
- Acquisition of process data
- Validation and verification
  - Run test experiments, capture measurements
  - Test process instances
- QA checks
  - Testing of workflow
  - Quality checks

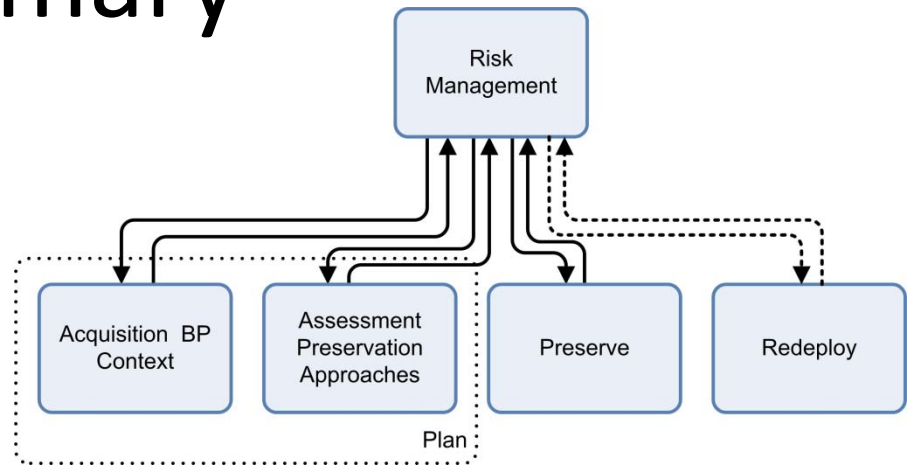
Redeploy


Risk Management
Acquisition BP Context
Assessment Preservation Approaches
Preserve
Redeploy
Plan

Redeploy

- Redeployment planning
  - Capturing of redeployment environment
  - Gap analysis
    - technical, organisational and legal gaps
- Re-execution of preserved process
  - Different redeployment approaches
  - VM-Player (e.g. VirtualBox)
  - Emulation-as-a-Service (e.g. bwFLa emulator for VMs)
- Validation of redeployment

Summary



- Process framework
- Flexibility, domain independent
- Guidance for implementation
- Technical and organisational implementation
- References and tool support
- Control mechanism