Emerging tools and issues in Digital Preservation: Virtualisation, Preservation and the TIMBUS project: a symphony in 4 parts

- Introduction to DP, Virtualisation and the Cloud
- Deeper dive into virtualisation versus emulation
- The TIMBUS Project
- Demonstrating TIMBUS products in action

William Kilbride, DPC
Perumal Kuppuudaiyar, Intel

Digital Preservation Approaches

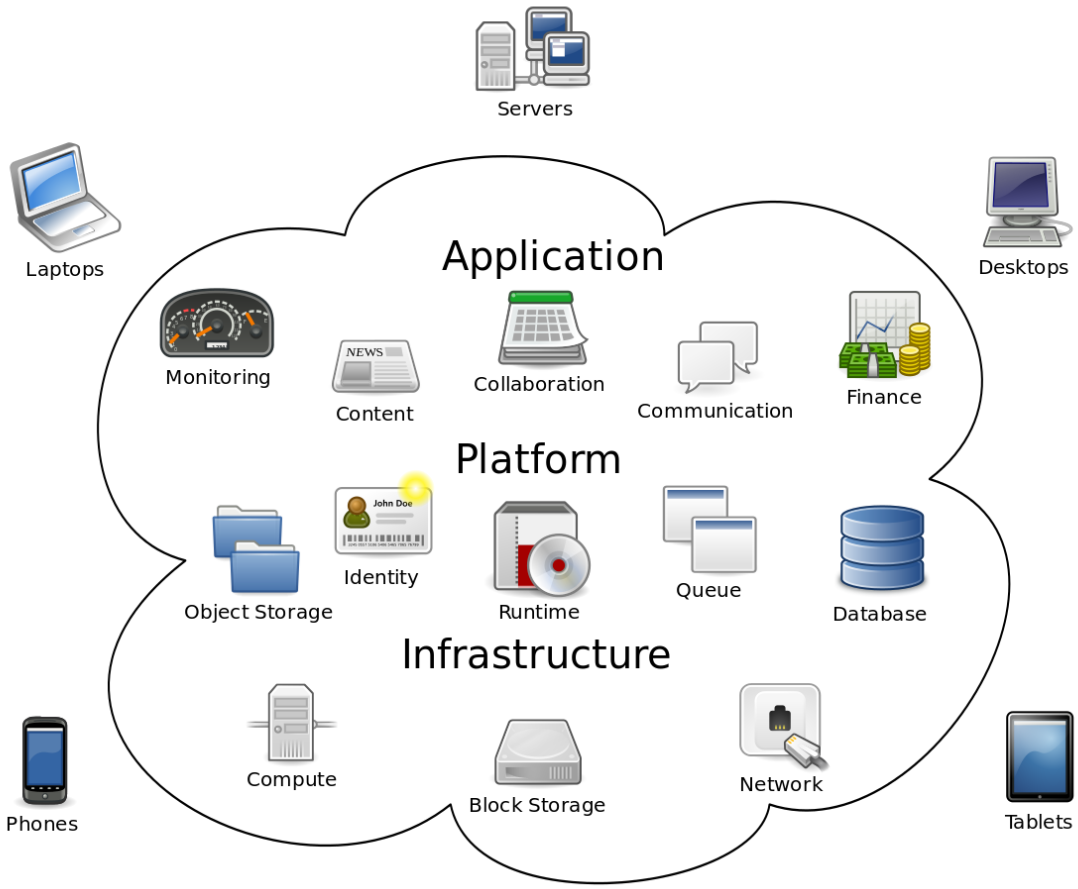
- Migration
  - Intervention at data layer to ensure information objects
  - Based on significant properties of content and performance
  - Quick start, low cost, ready quality assurance, focus on data/access
  - loss of authenticity, poor with complex objects
- Emulation
  - Intervention at software / OS layer to ensure operation of software
  - Based on significant properties of the environment and its behaviours
  - Slow start, high technical threshold, access less transparent
  - retains authenticity, geared towards complex objects

- Migration has done all the running in the last 10 years (20 years)

Cloud Computing Characteristics

- Scalable and Elastic
  - Services scale on demand to add or remove resources as needed
- Service based
  - The service could be considered "ready to use" or "off the shelf"
  - Offers IaaS, PaaS, and SaaS (soon will be LDPaaS)
- Economical
  - Services share a pool of resources to build economies of scale
  - Metered by Use : Pay-as- you- go
- Evolvability
  - Supports for migration and upgrades.
  - Services are configurable

Cloud Computing

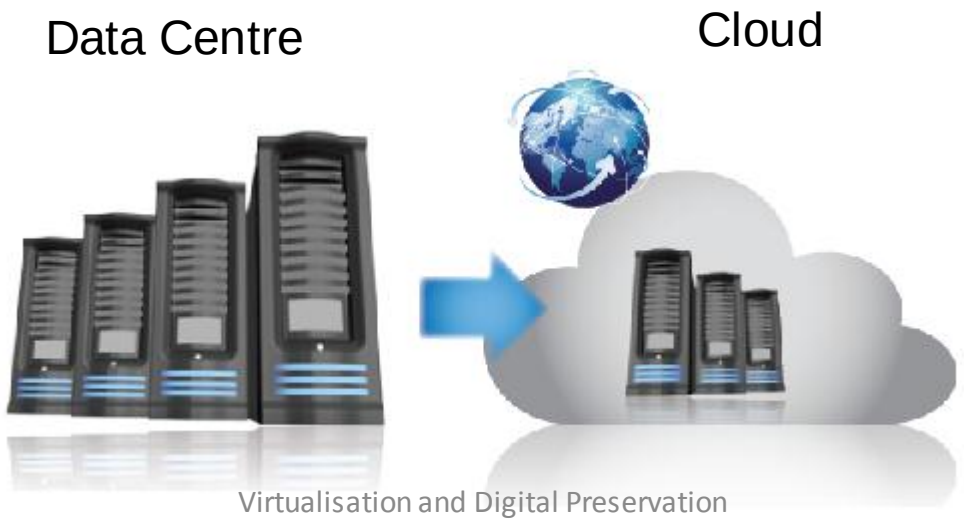


Cloud computing is the delivery of computing as a service rather than a product (source: Wikipedia)

Data Centre Vs Cloud

- Cloud is an **off-premise** form of computing that stores data on the Internet, whereas a data center refers to **on-premise** hardware that stores data within an organization's **local network**.

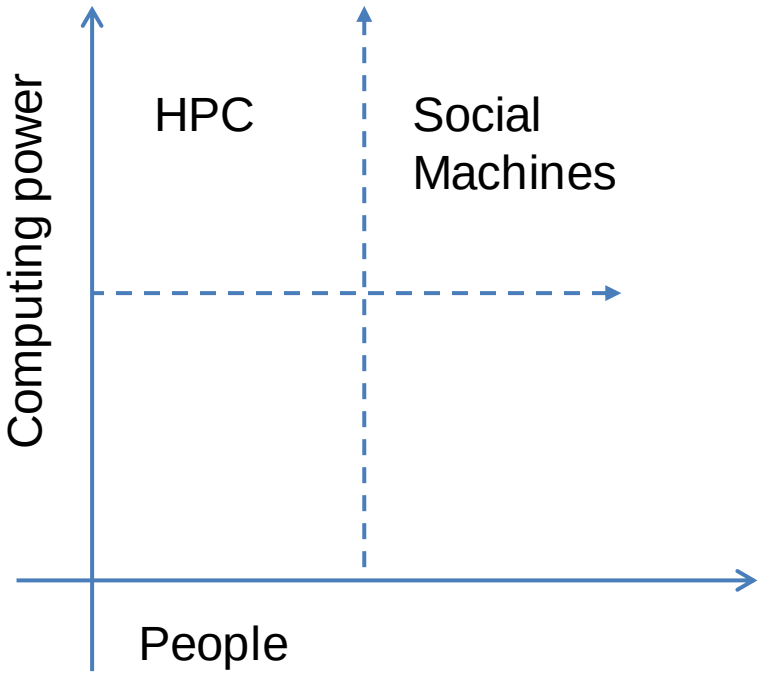
- While cloud services are **outsourced** to **third-party** cloud providers who perform all updates and ongoing maintenance, data centers are typically run by an **in-house IT department**.



Some observations …

1. **Big data complex data** as a metaphor for our future problems: does the cloud help?
2. Does the cloud make it **easier to engage** in digital preservation
3. Why would we ever put our **trust in the clouds**?
4. What will be **interface** between archives and producers
5. Does ‘preservation as a service’ change **development roadmap** (ie Cloud is for storage **and compute**)

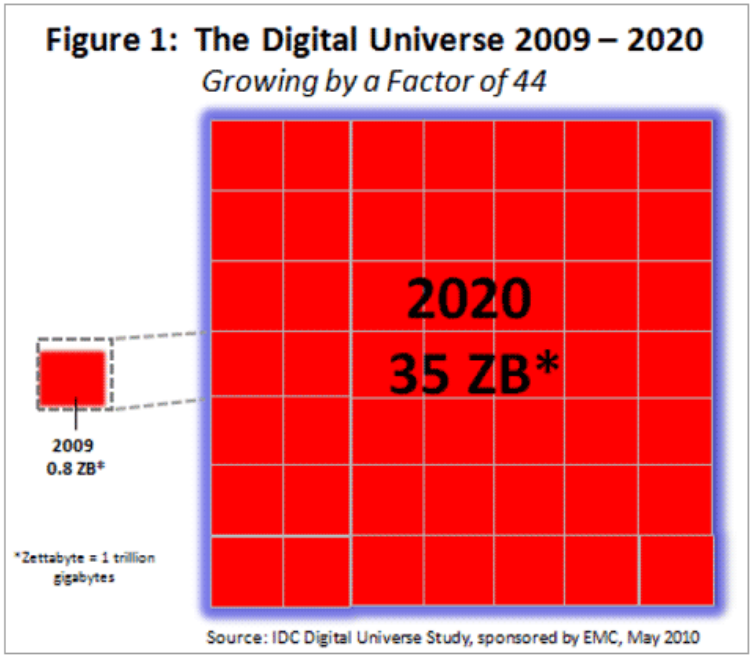
DP Futures



… it's not going to be about obsolescence so much as workflow and capacity

'Digital Universe' Nears A Zettabyte

May 4th, 2010 : Rich Miller



The Great Recession hasn't slowed the breakneck growth of the Digital Universe. In 2010 the volume of digital information created and duplicated in a year will reach 1.2 zettabytes, according to new data from IDC.

Big data / complex data

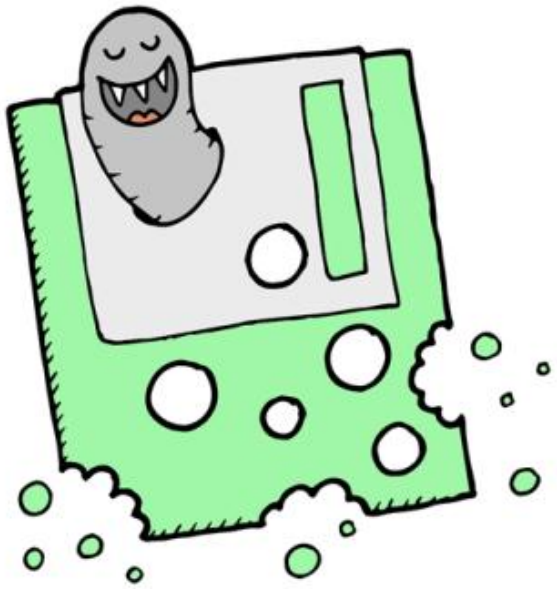
- Web archives
- Sound and vision
- Digitised content
- Email

Complex, vast, valuable, heterogeneous

Difficult to move

Difficult to access

Greater than the sum of its parts



Can the Cloud help core DP issues?

- Storage - yes
- Costs – maybe (maybe not)
- Skills – yes (in a narrow sense)
- Large scale migrations - yes
- Making emulation affordable?

Trust? Authenticity?

- Corporate Sustainability
- Policy and practice
- Effectiveness of preservation

...failure is not an option?

Trust1 – Corporate Sustainability

- Lock in to any service provider is a risk
- BRTF anyone? LIFE?
- http://4Cproject.eu/
- Need to model the costs in detail
- Need to understand business model of cloud providers

Unanswered question

How (expensive is it) to port material from one cloud provider to another?

(and does that change how we think about that?)

Trust2 – policy and practice

- Sensitivity review is really hard
- Highly (highly) risk averse
- Statutory requirements
- Copyright and preservation
- Need good mechanisms to tracking custody, document provenance etc
- Documented compliance to a whole range of standards
- Environmental impact
- And then there's the politics

Trust3 – effective practice

- What is success in digital preservation? OAIS?
- Evidence-based planning
- Audit and certification
  - ISO 16363
  - DSA
  - DIN
  - LOTAR …
- Succession planning
- Service dependency means service certification

Collecting the Cloud

- No question that cloud-based collections will interest memory institutions
- If collections are in the cloud already …
- How might a collection be transferred? What would ingest look like?
- Will everything ultimately become web archiving?
- What about time gates

Unanswered question

How will the interface between memory institutions and 'depositors' work?

(and does the Cloud change how we might think about that?)

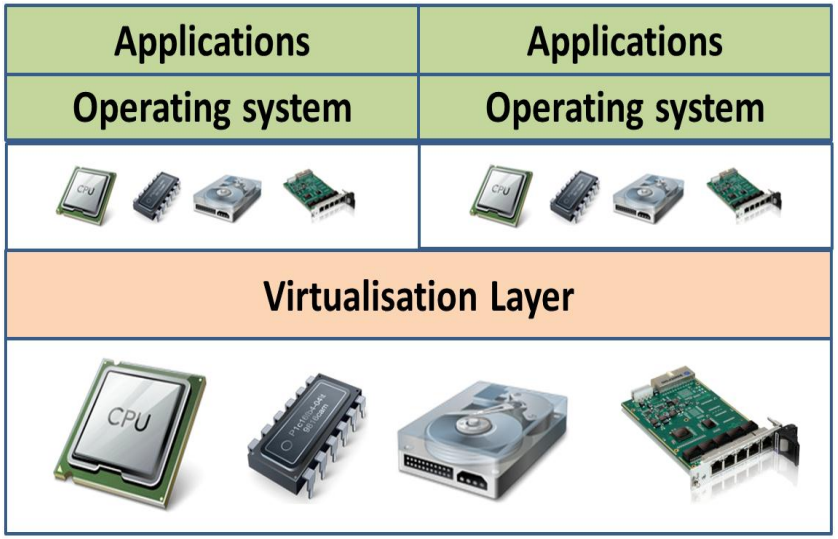
IBM 305 RAMAC (1956) with 50 x 24" discs holding 4.4Mb Leased by IBM for $35,000 pa

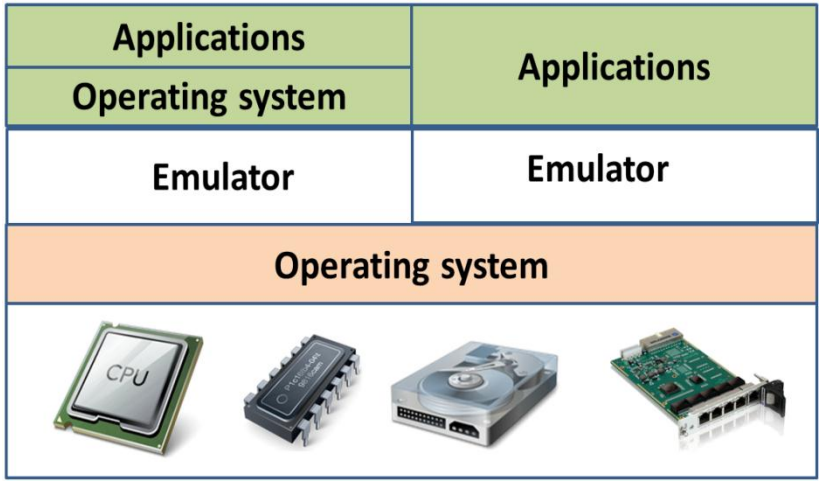
Development?

- Managing provenance and custody going to become much more important
- Assessing dependencies and version of services will become much more important
- Enables / extends alternative approaches to preservation
  - Migration on demand
  - Emulation / Virtualisation
  - Not just a welcome addition but a necessary solution

Virtualization Vs Emulation

- Virtualization
  - Virtualisation puts a layer between physical hardware and controls access to that machine.
  - Each guest machine (VM) that is built on top of the abstraction layer (hypervisor) is then provided access to the physical host's resources without modification.
  - The hypervisor act as a traffic cop by allowing certain amount of the physical resources to be used by the guests, as well as manages resource sharing when more than on guest system try to access the resources.

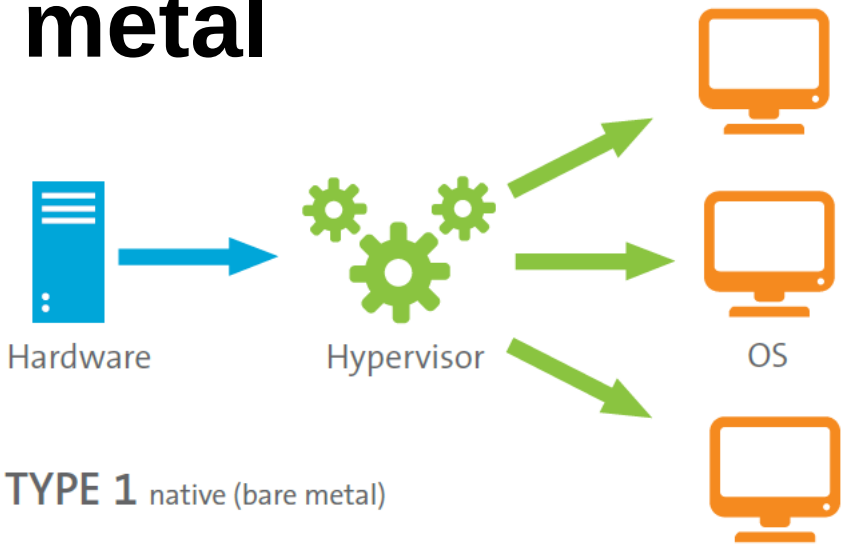




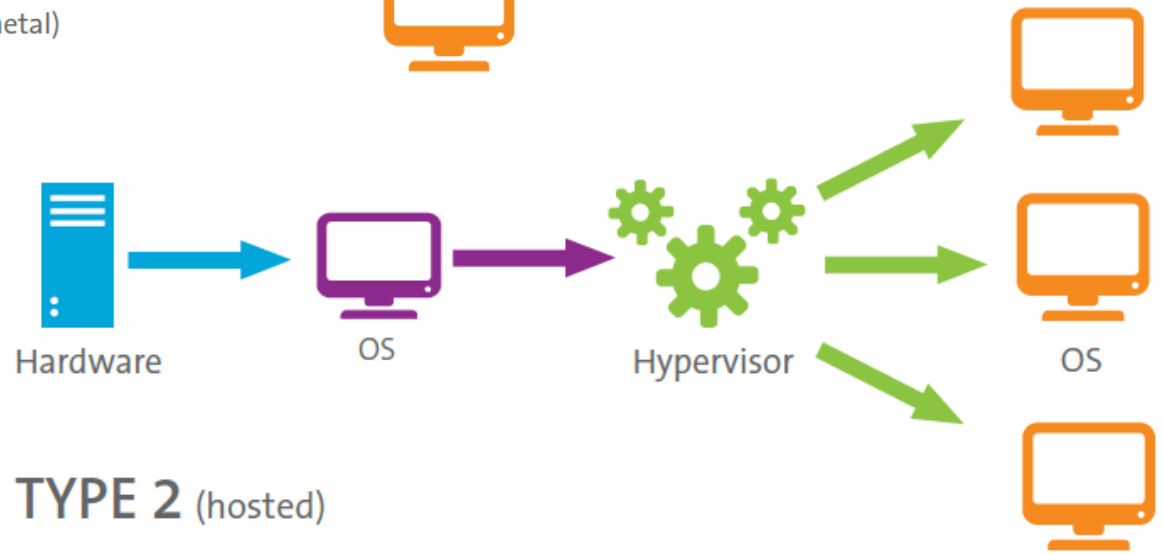
- Emulation
  - Duplication of functionality of systems, be it software, hardware parts, or legacy computer system as a whole, needed to display, access, or modify a certain contents.

Hypervisor Types

Bare metal

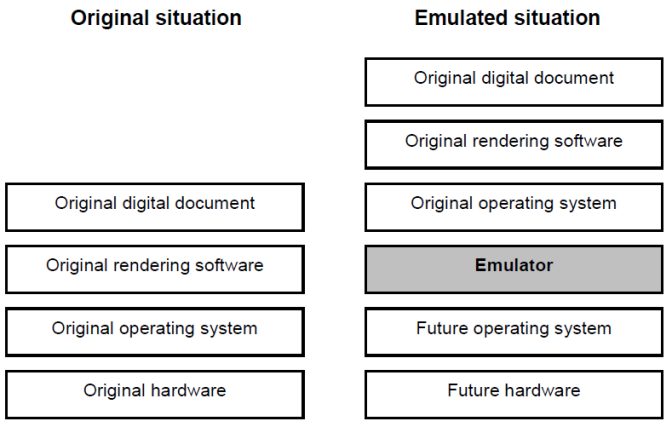


Hosted

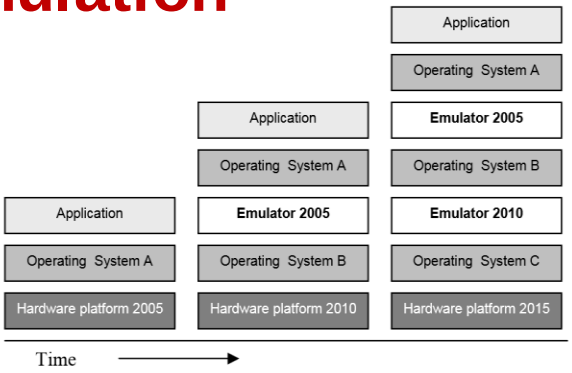


Emulation strategies for LDP

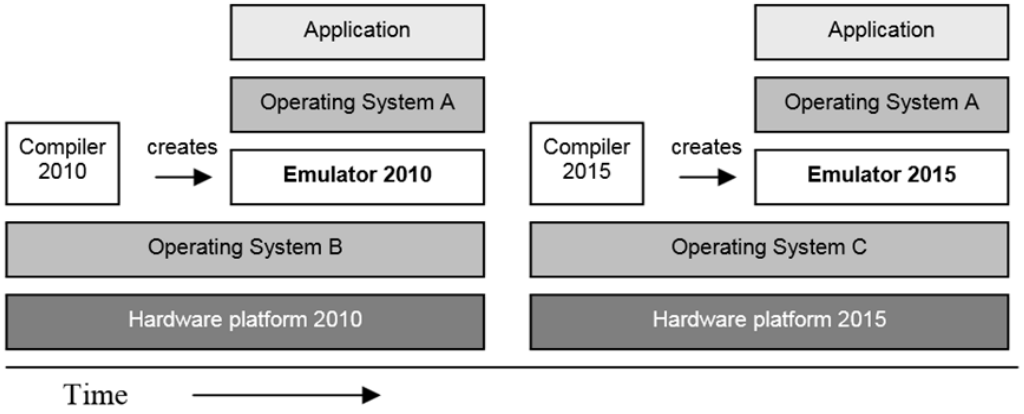
1. SW Emulation for HW



2. Stacked Emulation



3. Migrated Emulation over time

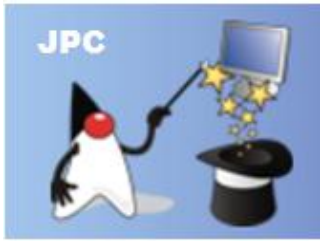


Source: (Jeffrey van der Hoeven)

Examples

- Hypervisors

- Emulators



DIOSCURI

Summary

- Virtualisation and emulation are two important technique of today's IT business.
- Future technology development will be focused more towards the benefit of emerging cloud computing.
- Virtualisation and emulation are looked on as potential enablers for preserving complex business environments for continued access regardless of technology changes over time.

Stop for questions!

- Next up, the TIMBUS project and some tools you can (ie will be able to) try out …

TIMBUS

Digital Preservation for TIMeless BUsiness processes and Services.

TIMBUS

TIMELESS BUSINESS

Digital preservation for timeless business processes and services

- timbusproject.net/
- info@timbusproject.net
- https://twitter.com/timbus_project

- April 2011 – March 2014

- co-funded by the European Union
  FP7/2007-2013
  under grant agreement no. 269940

The TIMBUS Consortium

- SAP – Lead partner (NI, CH)
- Intel (Ireland)
- Software Quality Systems (Germany)

Industry

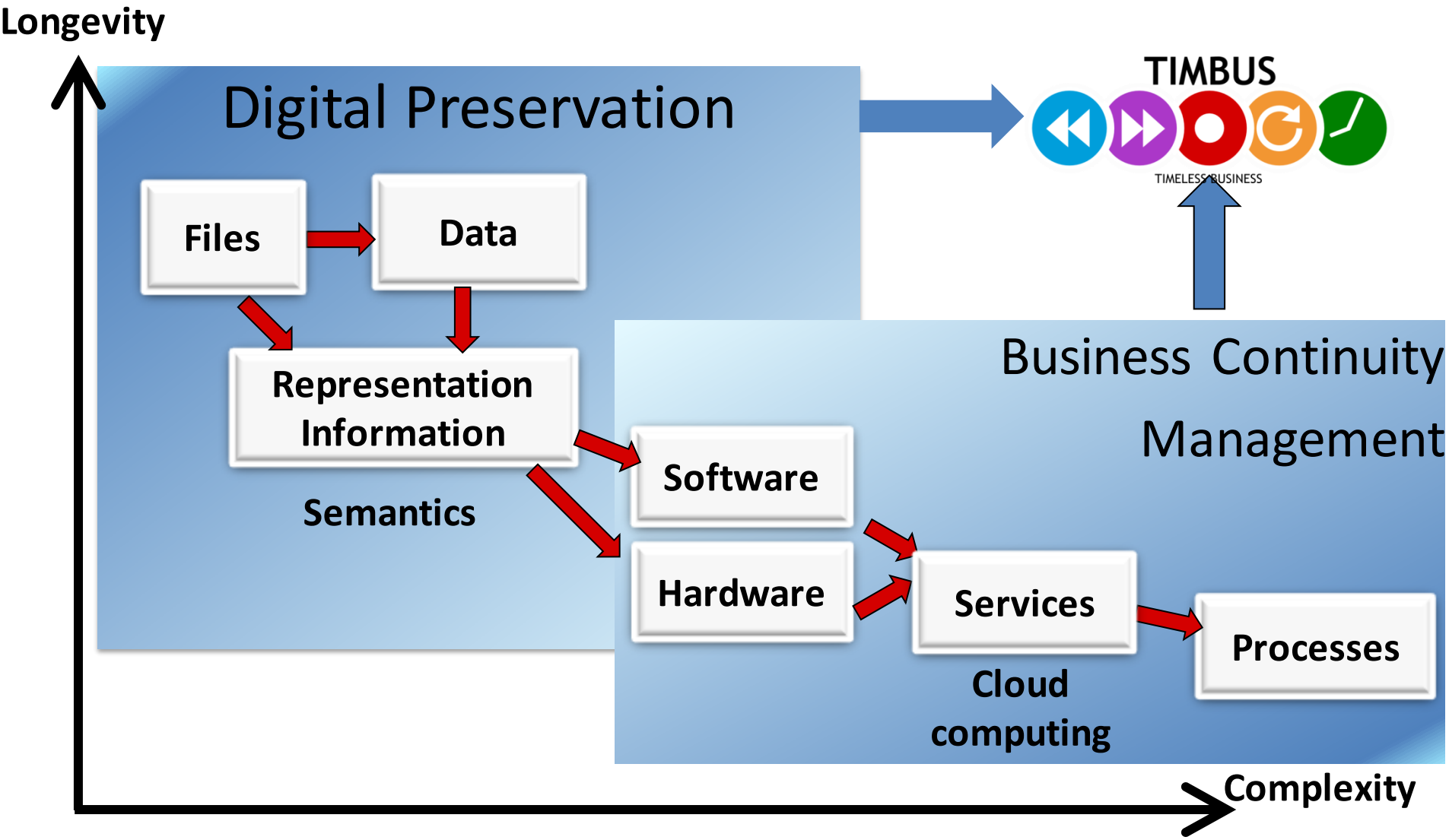
- Digital Preservation Coalition (UK)
- INESC – ID (Portugal)
- Karlsruhe Institute for Technology (Germany)
- Laboratório Nacional de Engenharia Civil (Portugal)
- Münster University (Germany)

Research

- Caixa Magica Software (Portugal)
- Secure Business Austria (Austria)

SMEs

A Preservation Continuum



Topics

- Motivation
- Objectives
- TIMBUS Process
- Architecture
- Context Model
- Owl Ontologies
- Tools
- Demo

Comparison of Cloud Providers

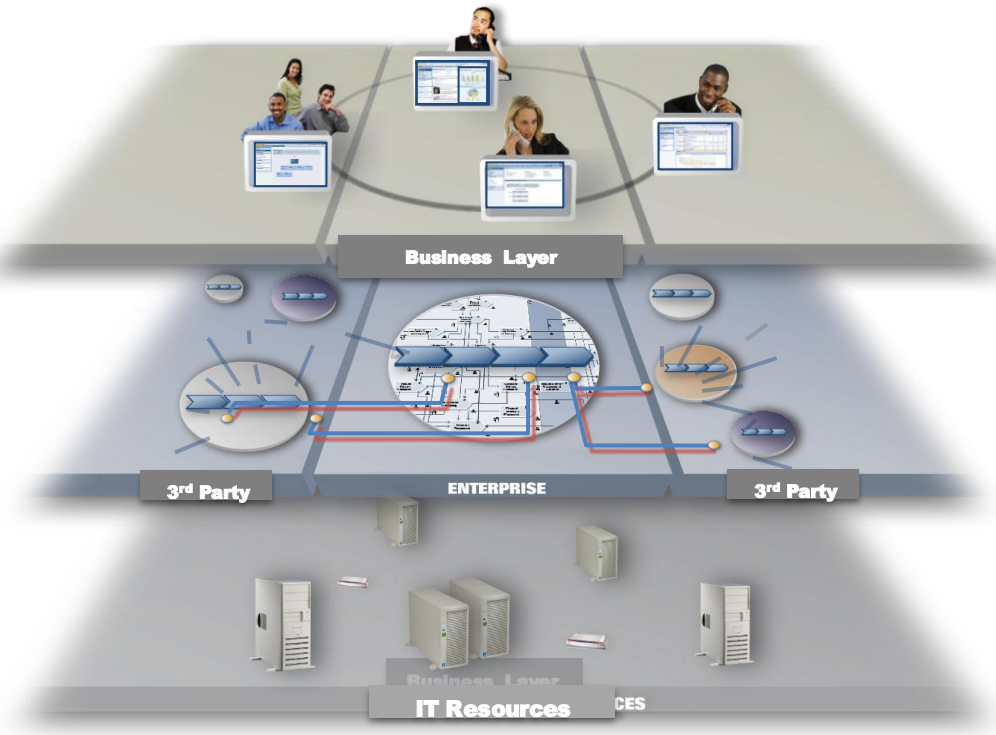
TIMBUS targets this

Provider / Service	Data Integrity	Reliability	Scalability	Retention & Portability	Availability	Data Ownership	Preservation Functionality	Total 3-year Costs
Amazon S3	Limited Checksums	Average	Almost unlimited	Not easy to move	Average	Similar to others	None	Medium
Amazon Glacier	No checksums	Multiple tape copies	Almost unlimited	Not easy to move	Lower since on tape	Similar to others	None	Low
Google Cloud Storage	No checksums	Average	Almost unlimited	Not easy to move	None in contract	Contract concerns	None	Medium
Tessella's Preservica	Checksums & CRC	Multiple cloud copies	Same as S3	Multiple Providers	Multiple cloud copies	Similar to others	Developed for this	?
VISI ReliaCloud	Limited Checksums	Average	Cannot support MHS	Somewhat limited	Average	Similar to others	Some claimed	High
SDSC Cloud Storage	Automatic Verification	Average but no tapes	Almost unlimited	Claimed to be easy	Concerns about disks	Similar to others	None	Medium
DuraSpace DuraCloud	User run checksums	Multiple cloud copies	Same as S3	Multiple Providers	Same as S3	Similar to others	Some claimed	Medium
IBM SmartCloud	Limited Checksums	Average	Unknown	Unknown	None in contract	Similar to others	None	?
FugiFilm Permivault	Custom plans	On-site and cloud	Almost unlimited	Somewhat limited	On-site copy	Similar to others	Limited	Low
FugiFilm Permivault Client	Custom plans	Cloud only	Almost unlimited	Somewhat limited	Average	Similar to others	Limited	Low
Code 42 CrashPlan Pro	Limited Checksums	Average	More limited than S3	Somewhat limited	Average	Similar to others	None	Low

TIMBUS Why

PRIMARY MOTIVATIONS

- Declining popularity of centralized in-house businesses.
- Increasing popularity of SaaS, PaaS, (*aaS), and IoS.
- Requirement for dependability.



Evidence under Litigation

Protect Knowledge

Enable Diagnosis

Business Continuity

Regulatory Compliance

Objective

- Establish efficient and effective processes and methods for DP of business processes
- Make sure to address technical, but also socio-technical and legal issues
- Aim at processes and methods that are intelligent and do not rely solely on human guidance
- Framework to guide Validation and Verification Process
- Identification of security aspects of use cases

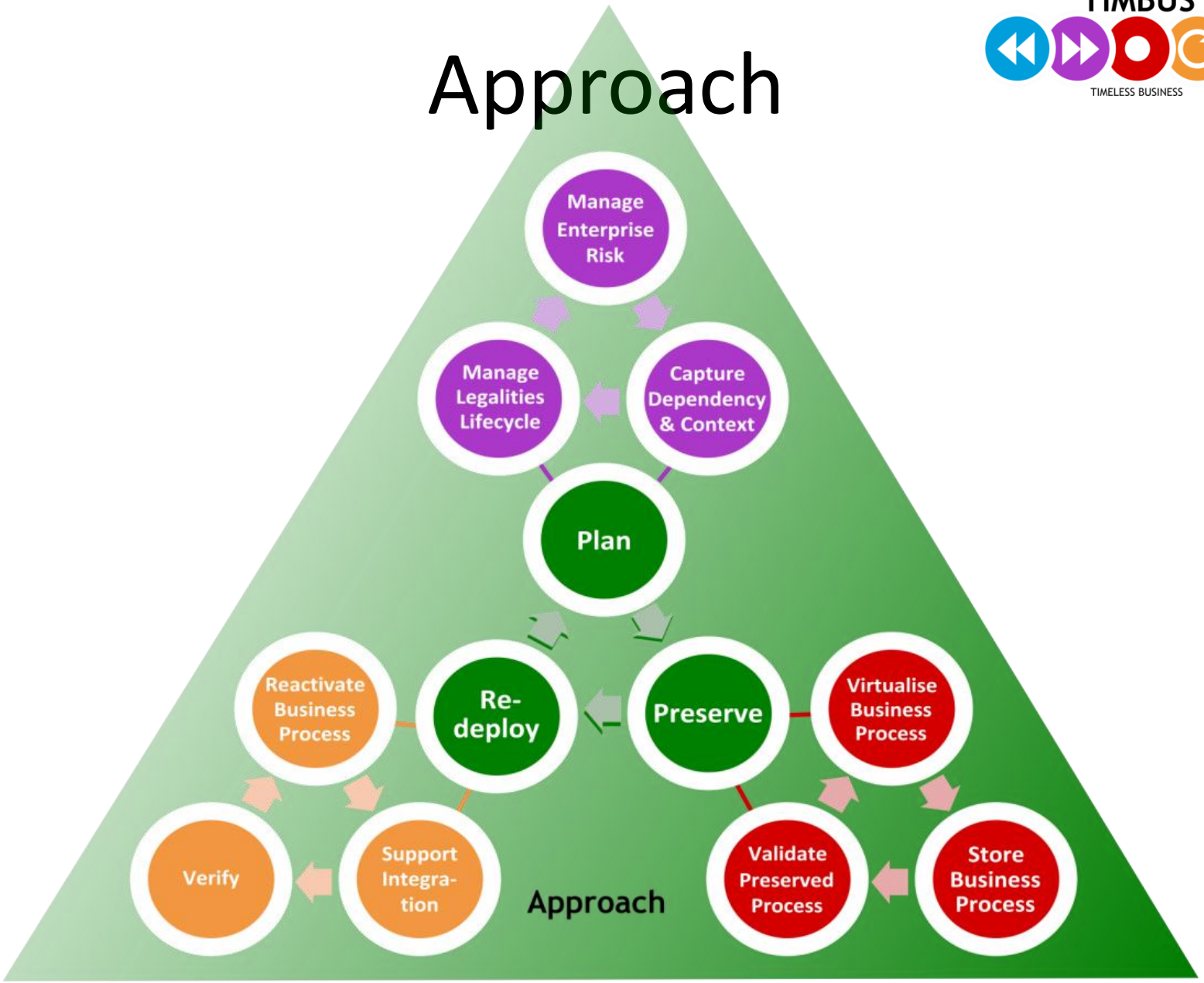
TIMBUS Approach

- Establish **activities**, **processes** and **tools** to ensure continued access to business processes and supporting services and infrastructure.

- Align **preservation** with **enterprise risk management (ERM)** and **business continuity management (BCM)**.

- Explore DP from a BCM perspective.

Approach


Manage Enterprise Risk
Manage Legalities Lifecycle
Capture Dependency & Context
Plan
Reactivate Business Process
Re-deploy
Preserve
Virtualise Business Process
Verify
Support Integra-tion
Approach
Validate Preserved Process
Store Business Process

TIMBUS Innovations

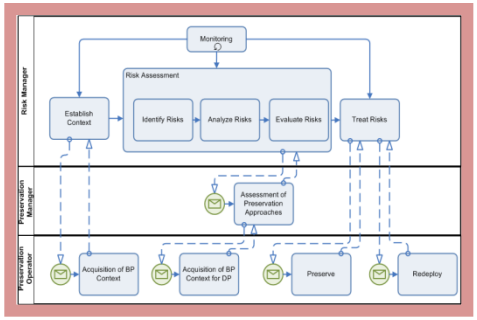
- **PLANNING**
  - Intelligent enterprise risk management
  - Service Dependency Analysis
  - Legalities Lifecycle Management
  - Business Process Context Capture

- **PRESERVATION**
  - Business Process Virtualization and Storage
  - Validation of the preserved process

- **REDEPLOYMENT**
  - Business Process reactivation / exhumation
  - Integration Support
  - Verification

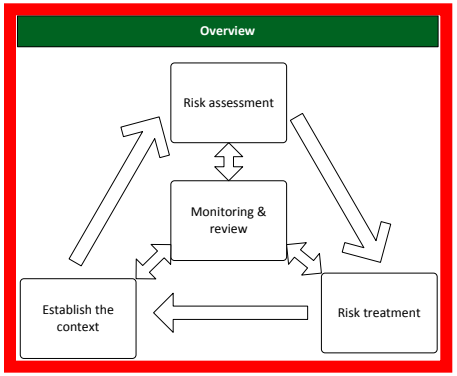
Process Flow

Risk process



Software ESCROW

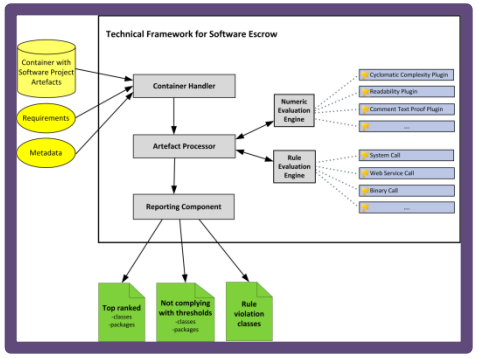
TIMBUS Process Framework



Risk Management

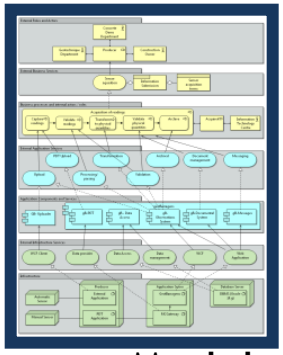
Context model

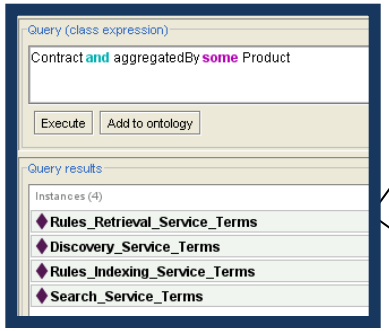
Reasoning



Technical Framework for ESCROW

Context and dependencies

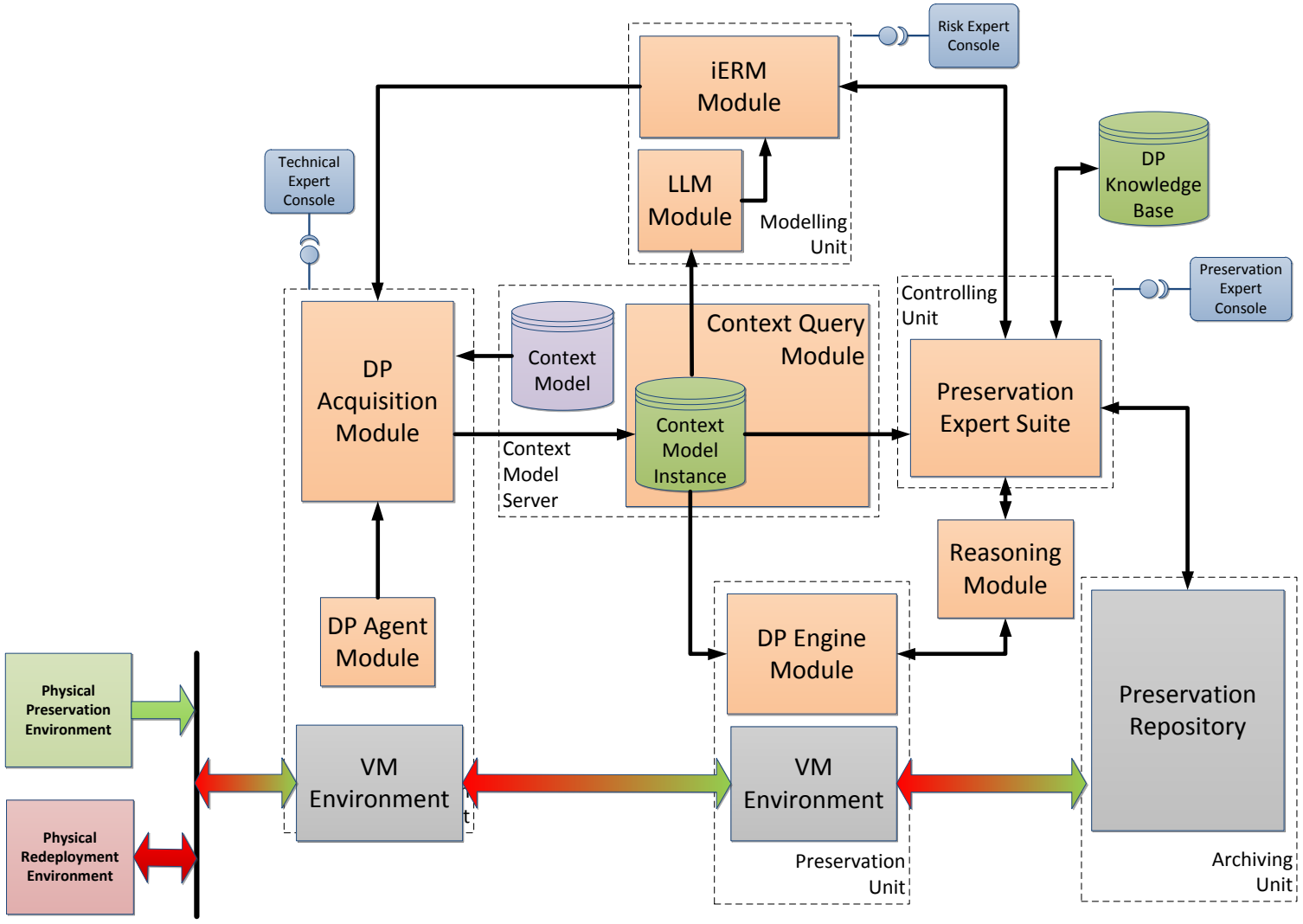




Context

Context Model + Dependency Model

Architecture Overview



The Context Model

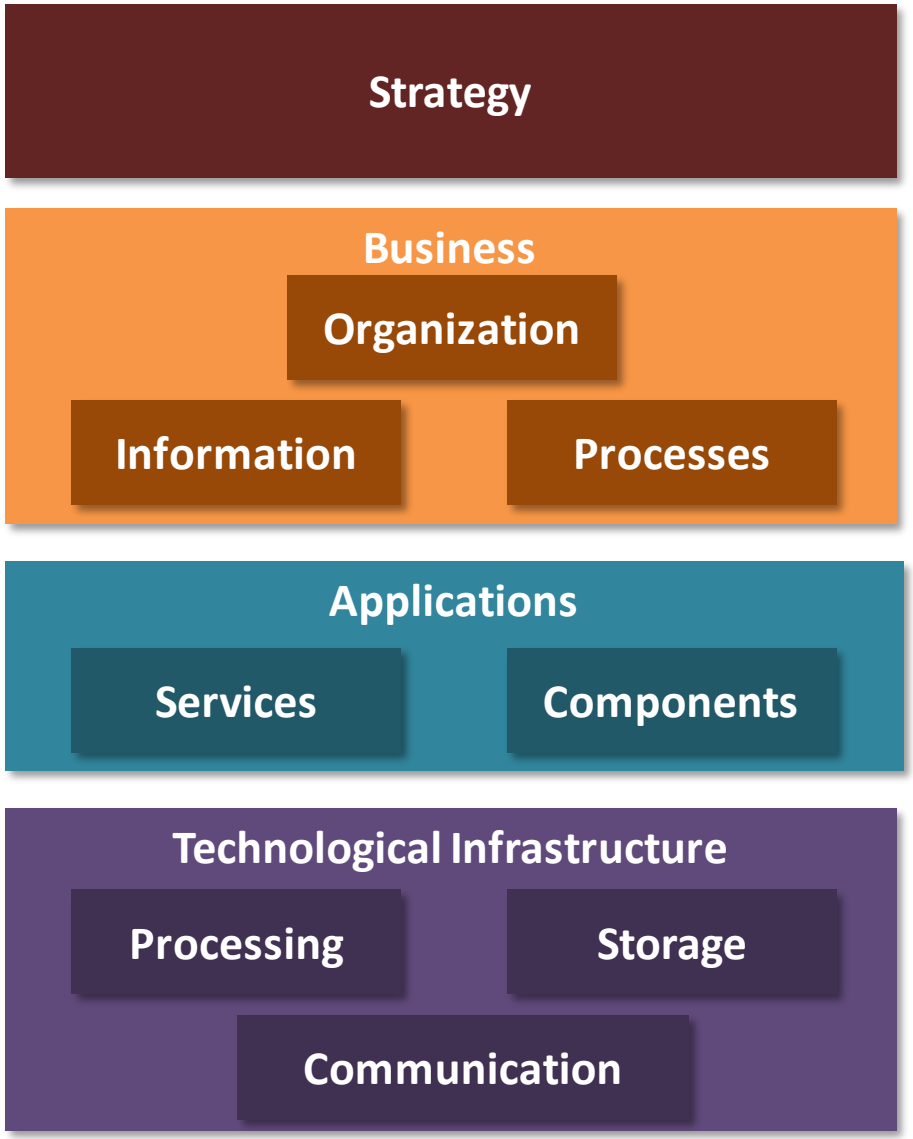
- The context model supports business process
  - preservation,
  - redeployment and
  - analysis.
- The context model defines the semantics of business process modelling in TIMBUS.
- The context model usage crosscuts the overall TIMBUS architecture.

Stakeholder Questions

Survey involving all TIMBUS partners with +100 answers.

Which business actors are required to *execute* business process P?	List of **Business Actors**
Which technological entities *support* business process P?	List of structural and behavioural **technological entities**
Which application components *support* business process P?	List of **application components**
Which legal requirements are *verified by* business process P?	List of **legal requirements**
Which are the licenses required to *execute* software application S?	List of **licenses**
…	…

Context Model Domains



Strategic Indicators, External Services, Contracts, Regulations, Licenses, Legal Requirements, Patents

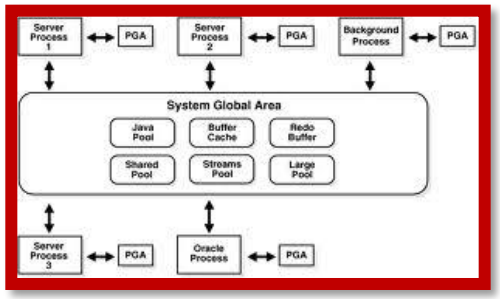
Organizational Structure, People, Business Processes, Operational Indicators

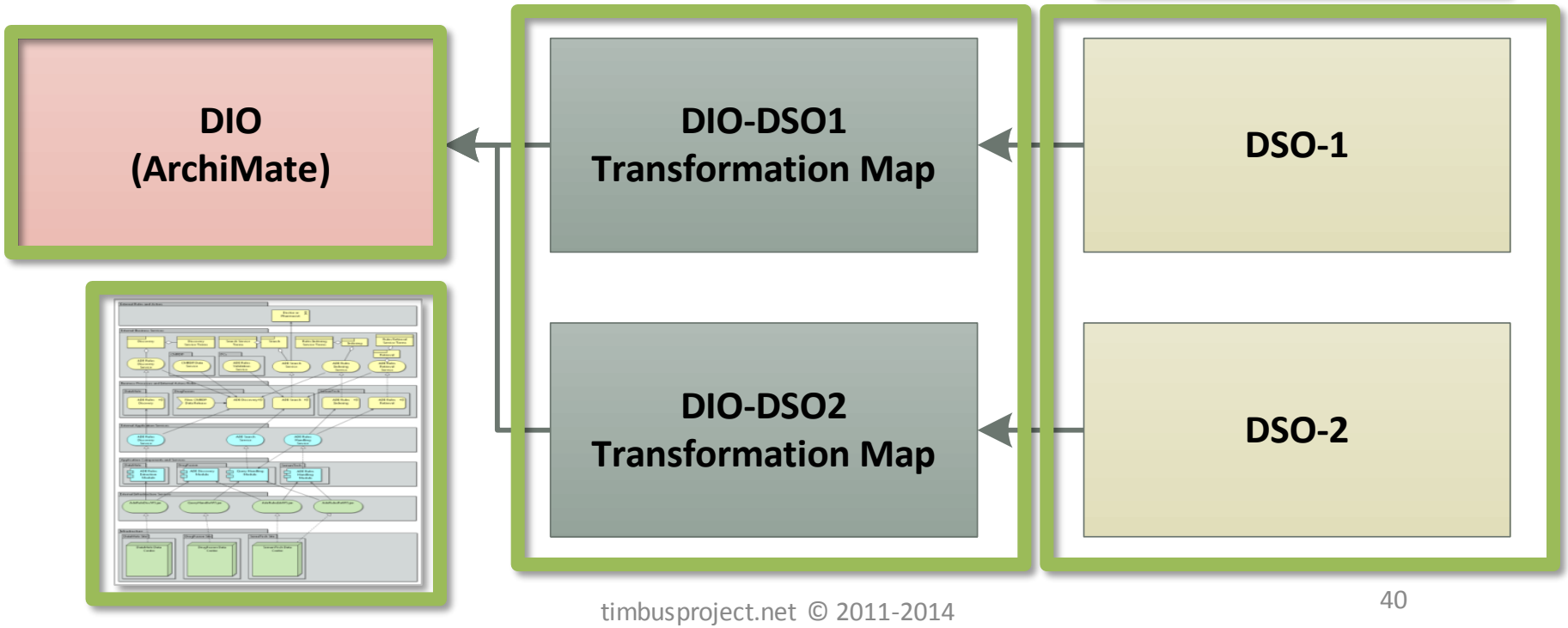
Applications, Services, Virtualization Applications

Deployed software applications and services, Hardware nodes, Communication nodes

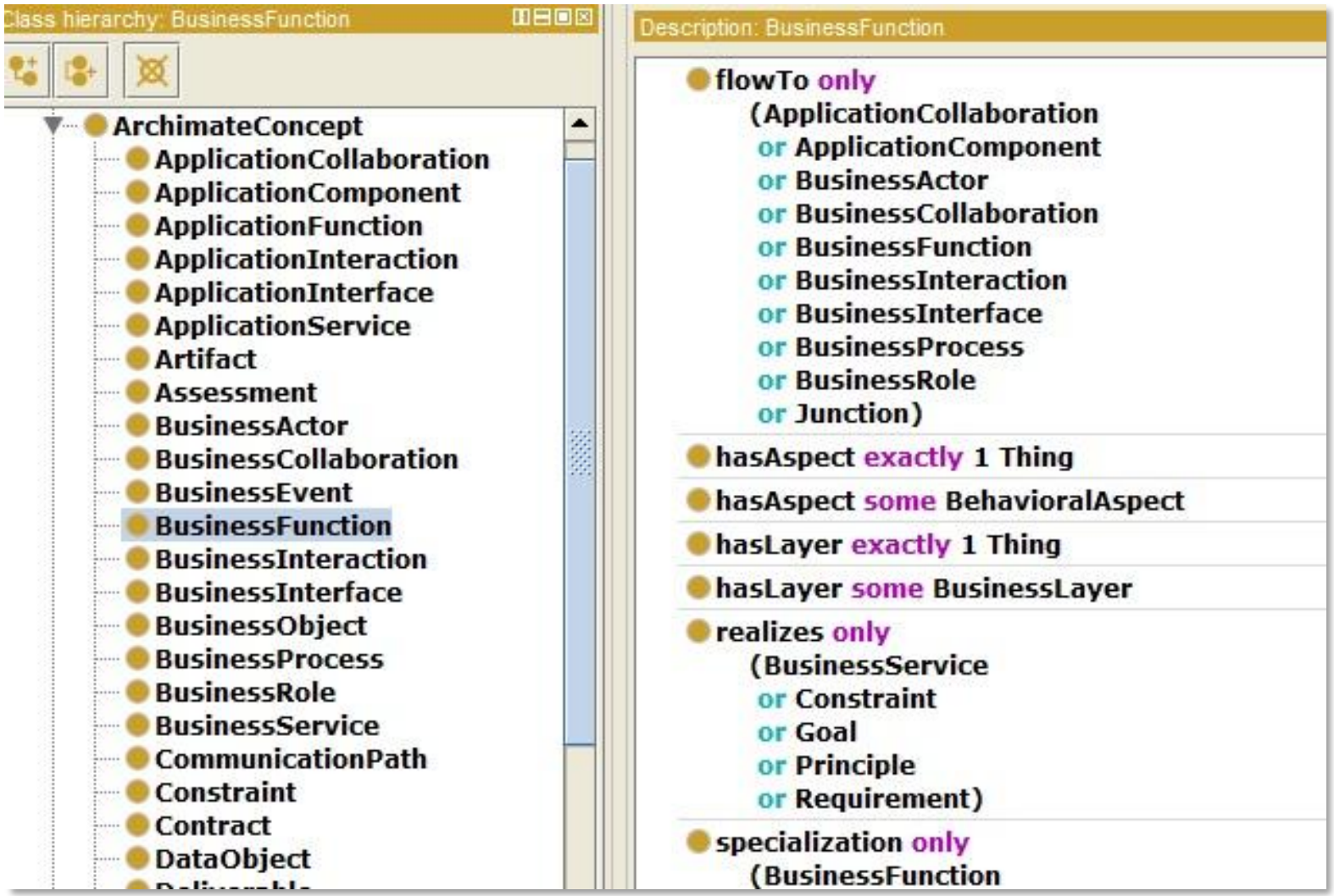
Architectural Concepts

- DIO: Domain-Independent Ontology
- DSO: Domain-Specific Ontology
- Ontology integration (transformation maps)
- Model transformation and extraction





ArchiMate DIO

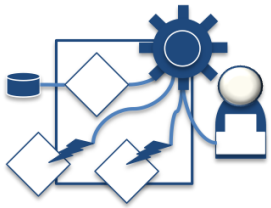


Implementations

TIMBUS Tools
(Demo)

Risk Management Cycle

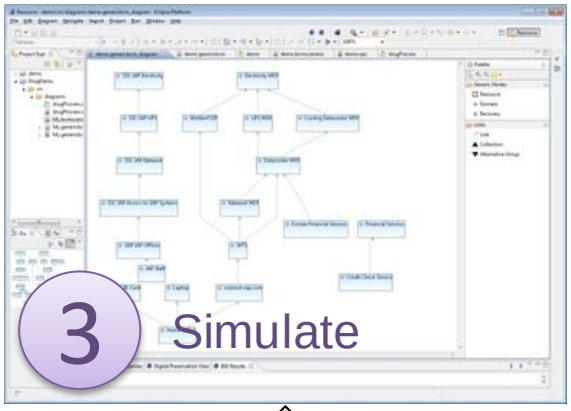
BPMN



Context Model Instance

Load

1



3 Simulate

Risk Assessment Web-Service

4

iERM

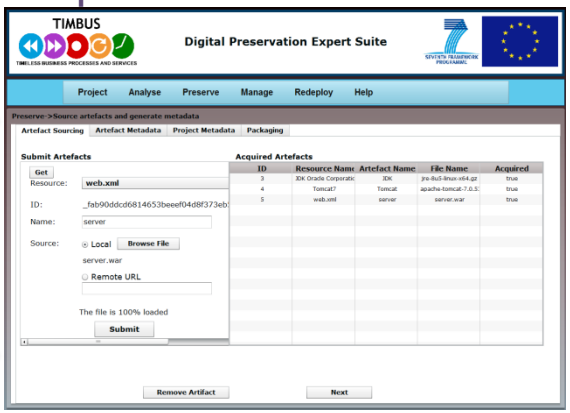
Annotate

2

Business Impact Data

5

Request



6

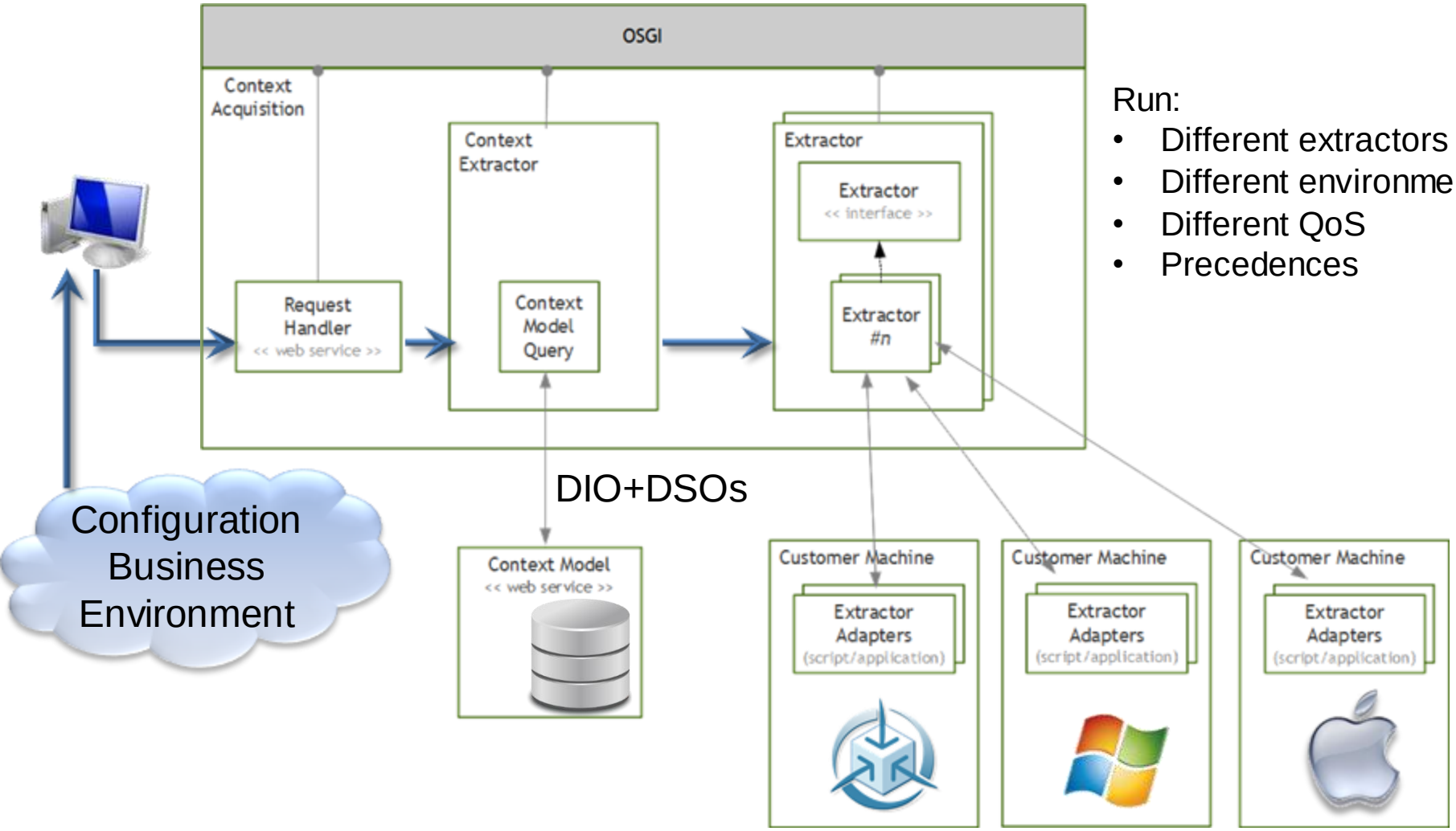
Request

DPES

Preservation

Preservation recommendations

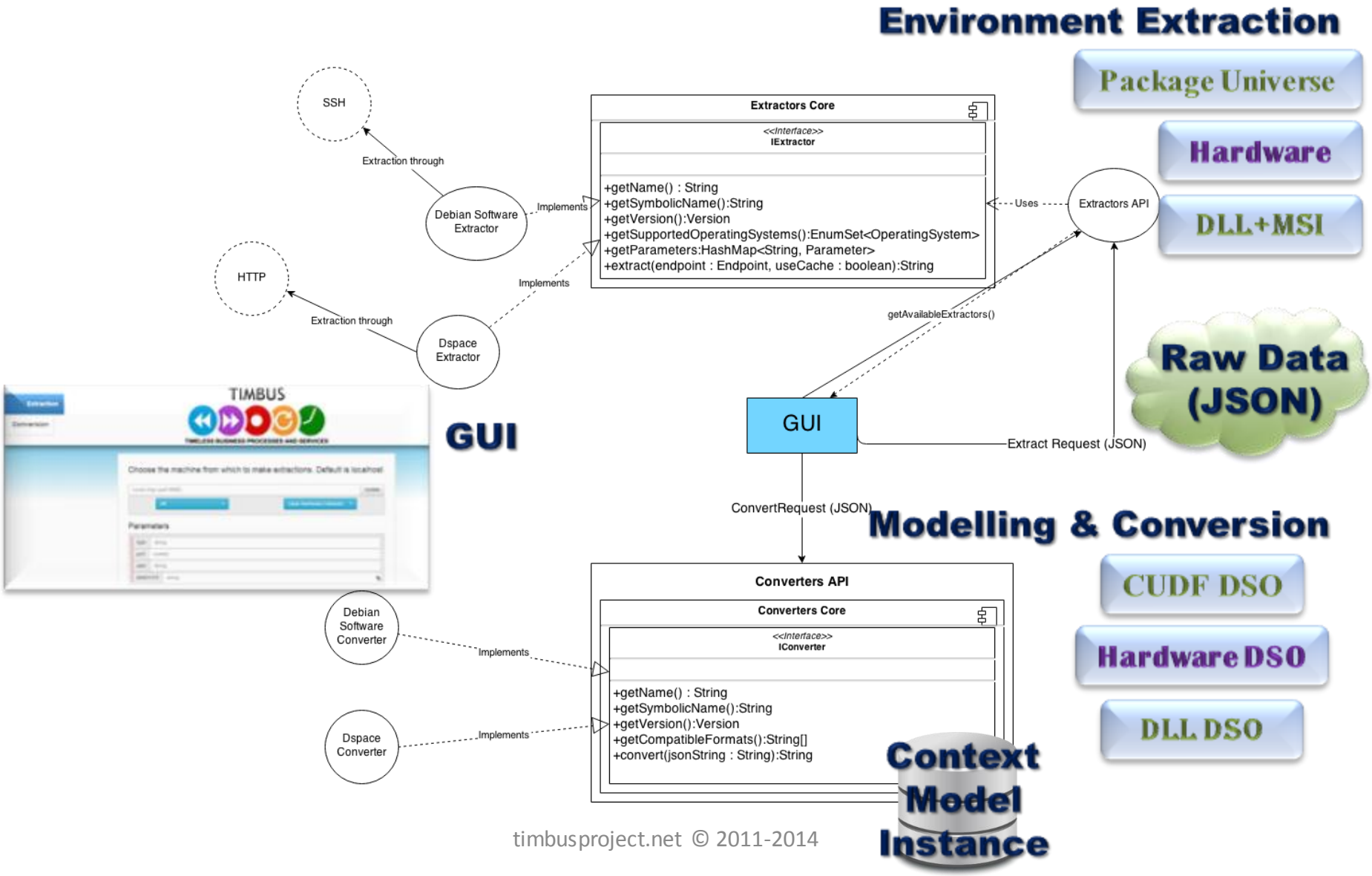
Context Acquisition Framework



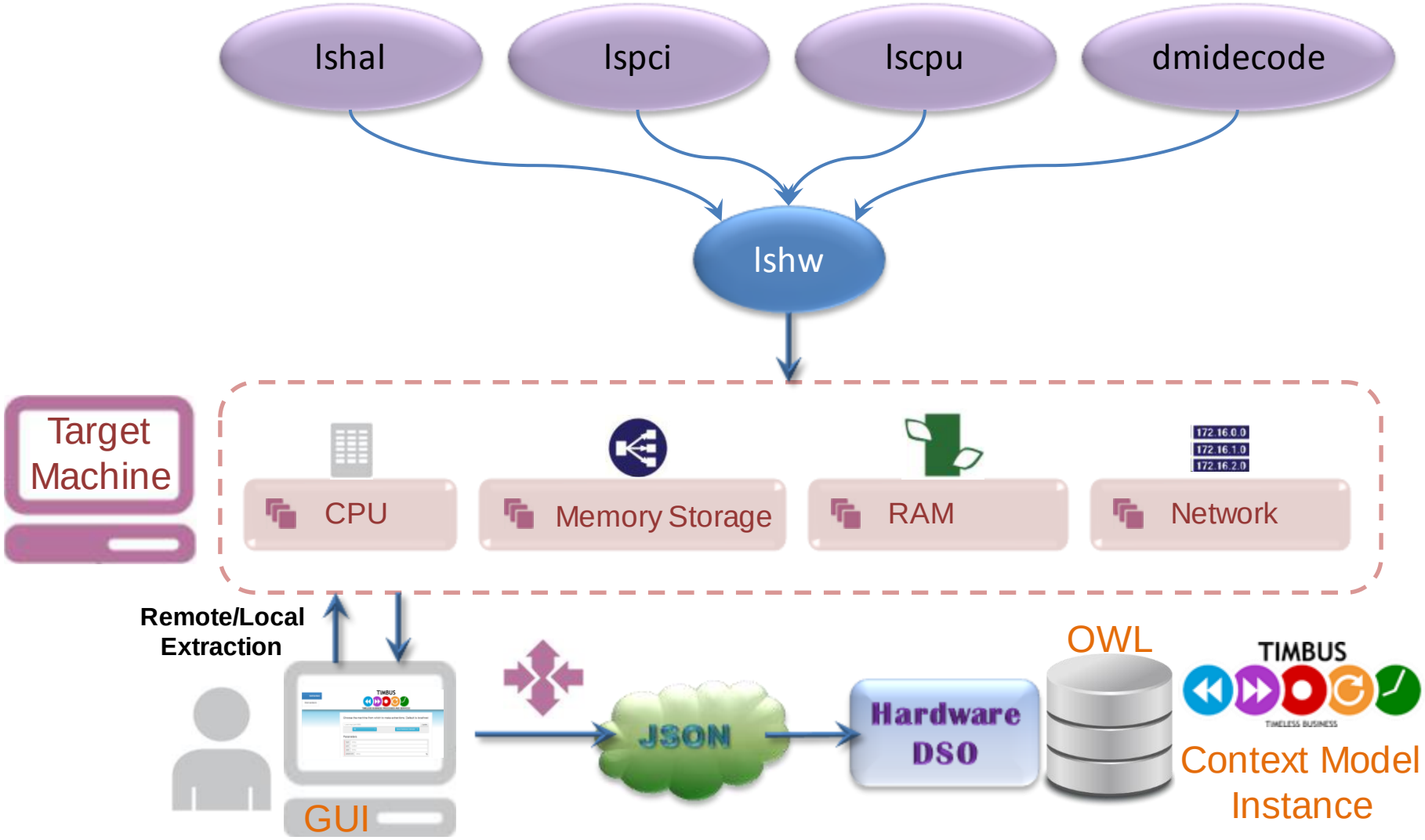
Run:

- Different extractors
- Different environments
- Different QoS
- Precedences

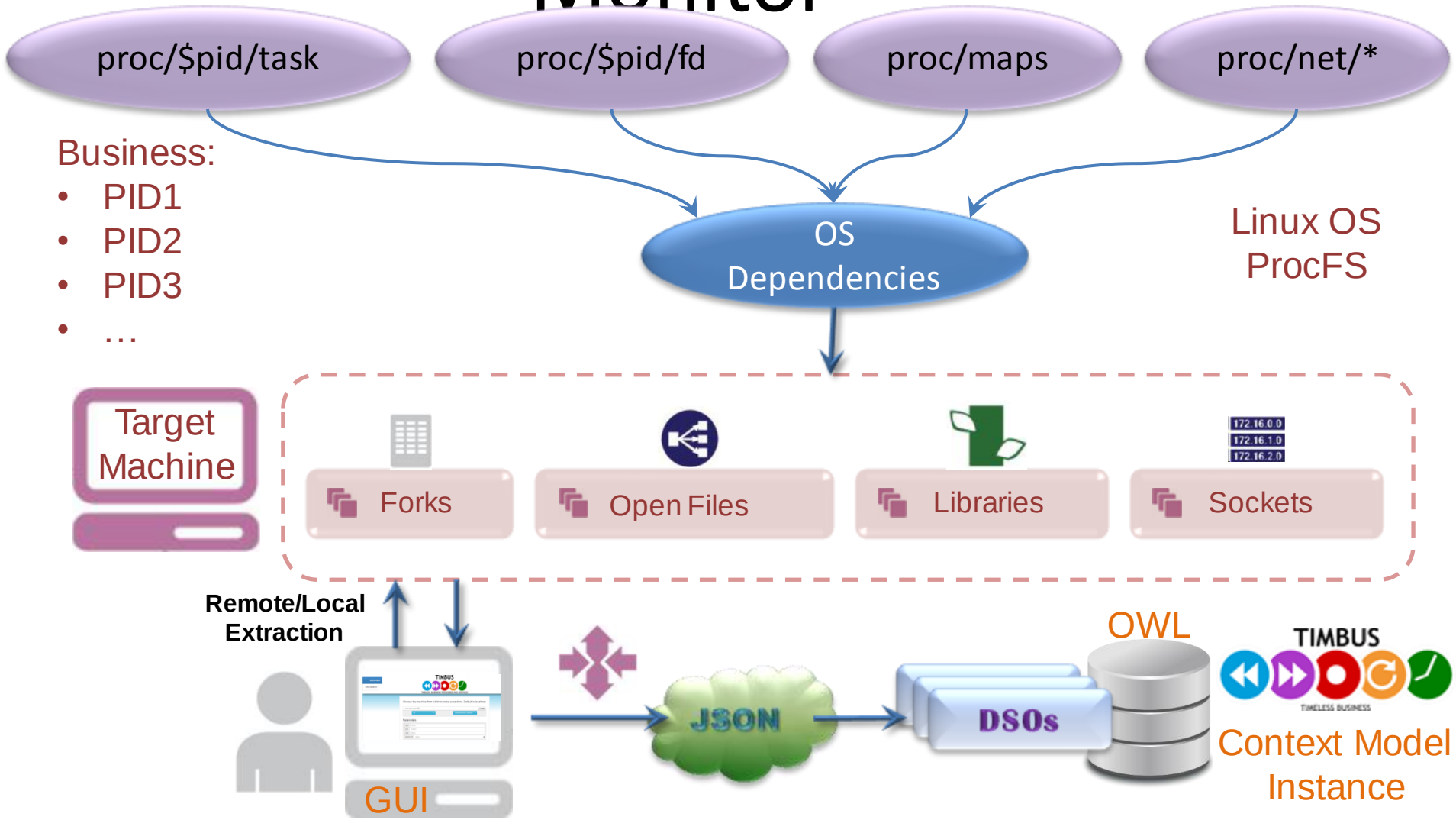
Context Acquisition



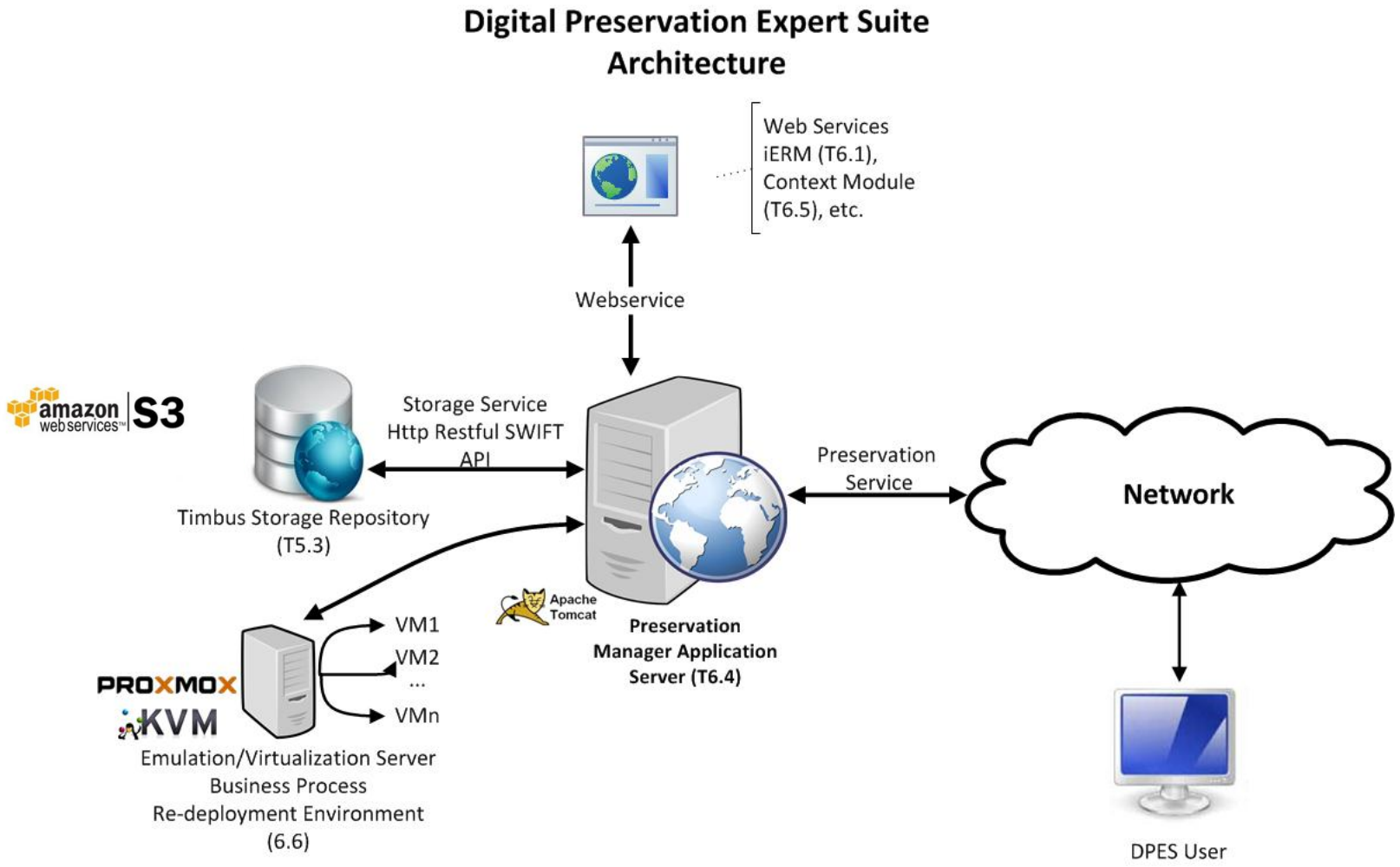
Extractor – Linux HW



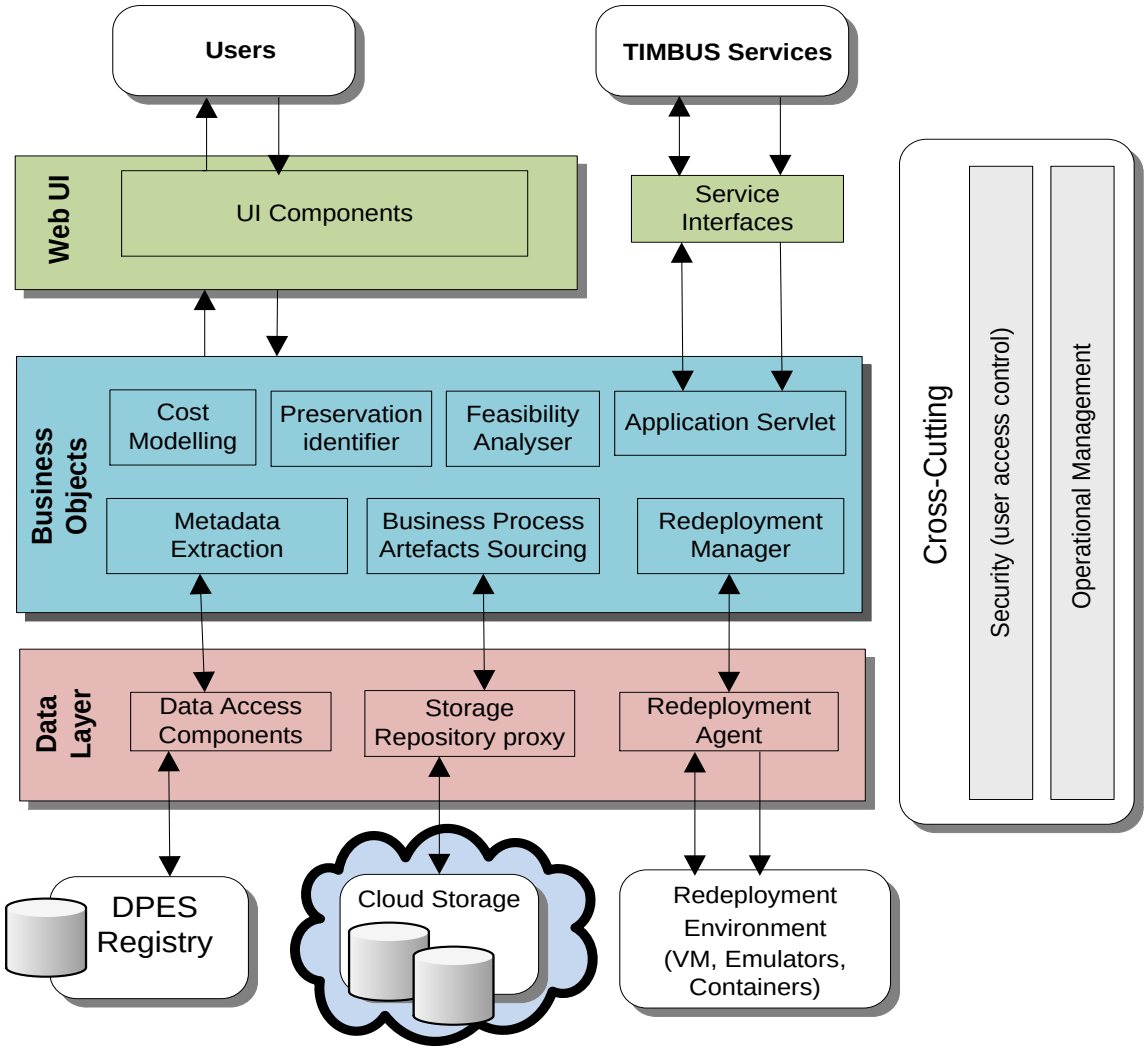
Extractor – OS Resource Monitor

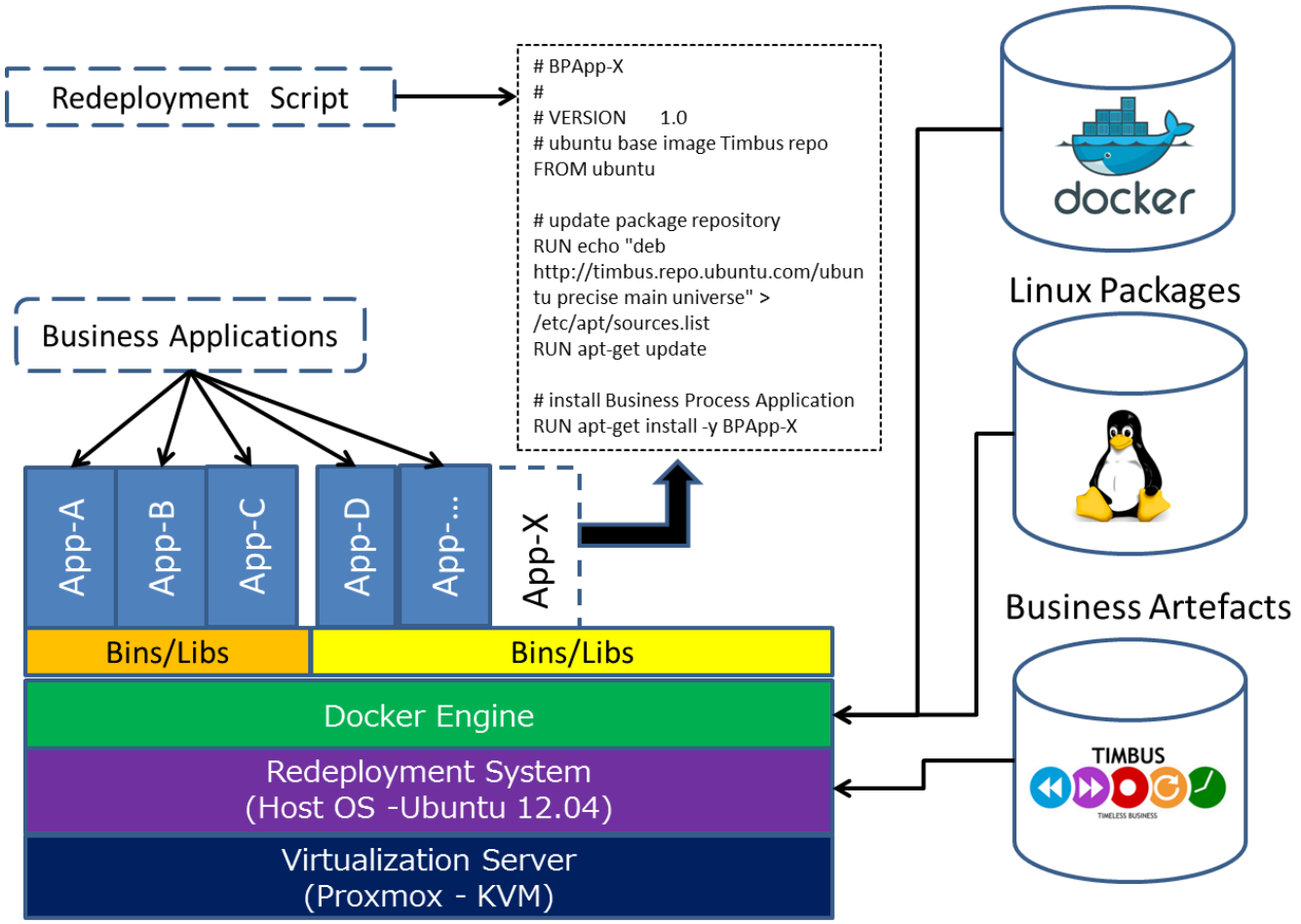


DPES - Implementation

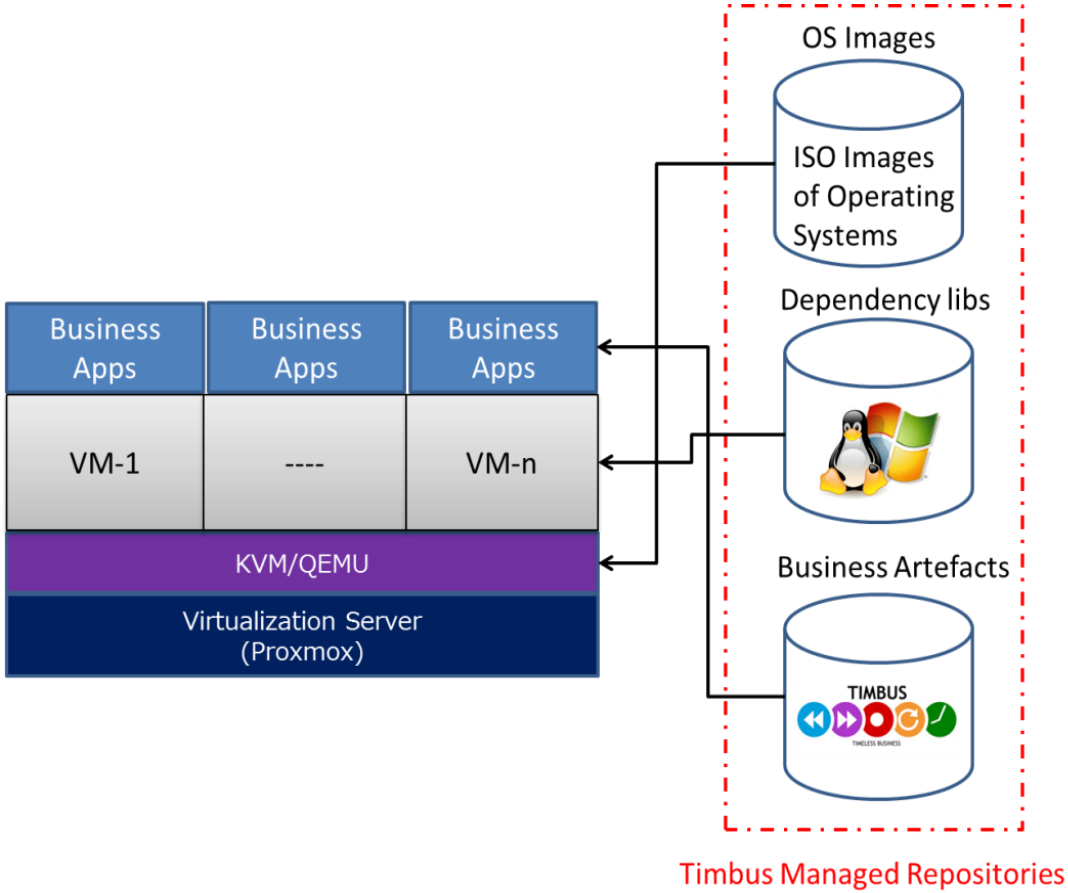


DPES - Internals



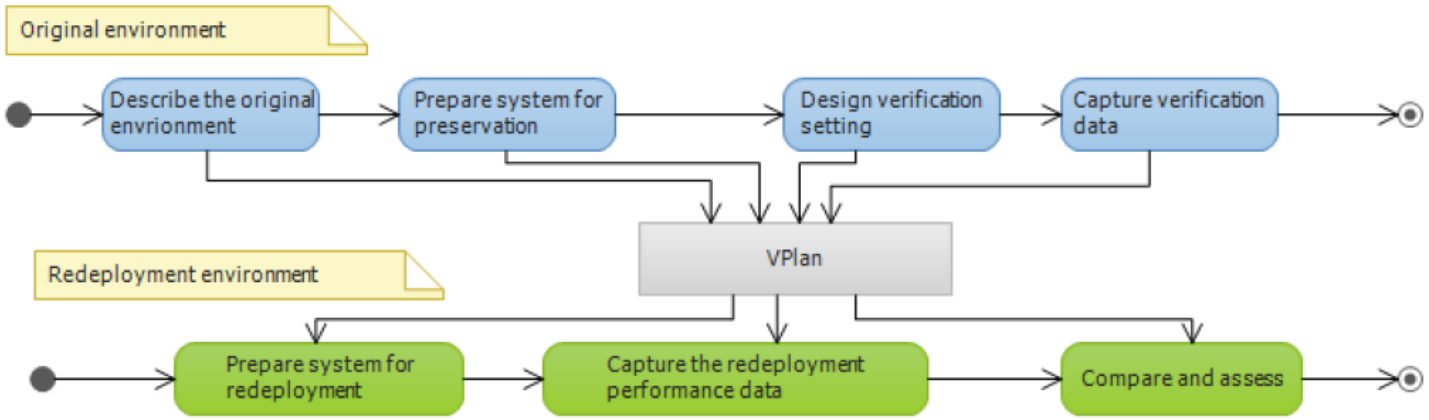

Redeployment Script
BPApp-X
#
VERSION 1.0
ubuntu base image Timbus repo
FROM ubuntu
update package repository
RUN echo "deb
http://timbus.repo.ubuntu.com/ubun
tu precise main universe" >
/etc/apt/sources.list
RUN apt-get update
install Business Process Application
RUN apt-get install -y BPApp-X
Business Applications
App-A
App-B
App-C
App-D
App-...
App-X
Bins/Libs
Bins/Libs
Docker Engine
Redeployment System
(Host OS -Ubuntu 12.04)
Virtualization Server
(Proxmox - KVM)
docker
Linux Packages
Business Artefacts
TIMBUS
Timbus Managed Repositories

DPES – Redeployment Model-2



Verification and Validation

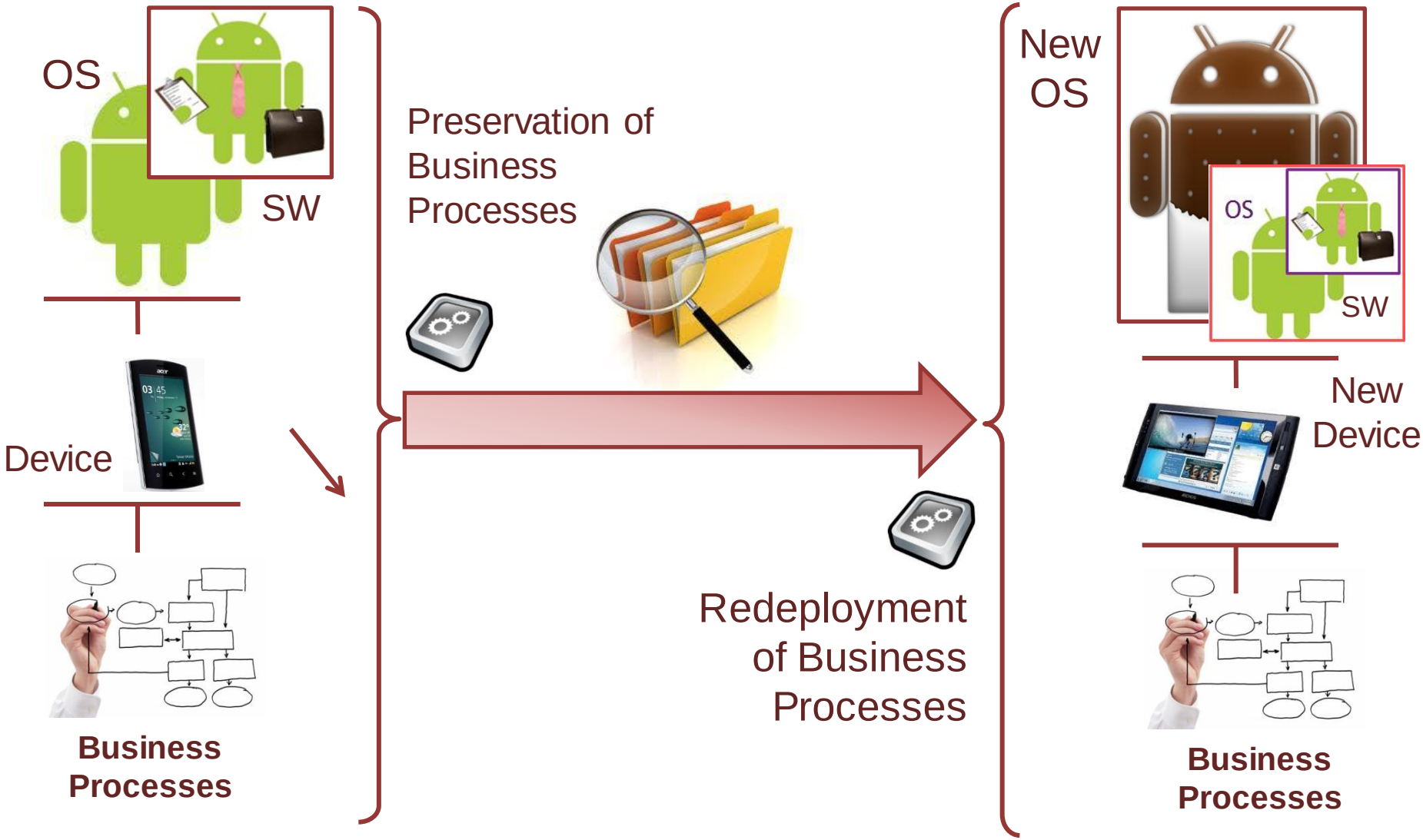
- Define Significance properties and metrics
- Describe methodologies to measure
- Acquire values and Preserve
- Redeploy, capture new values and Compare



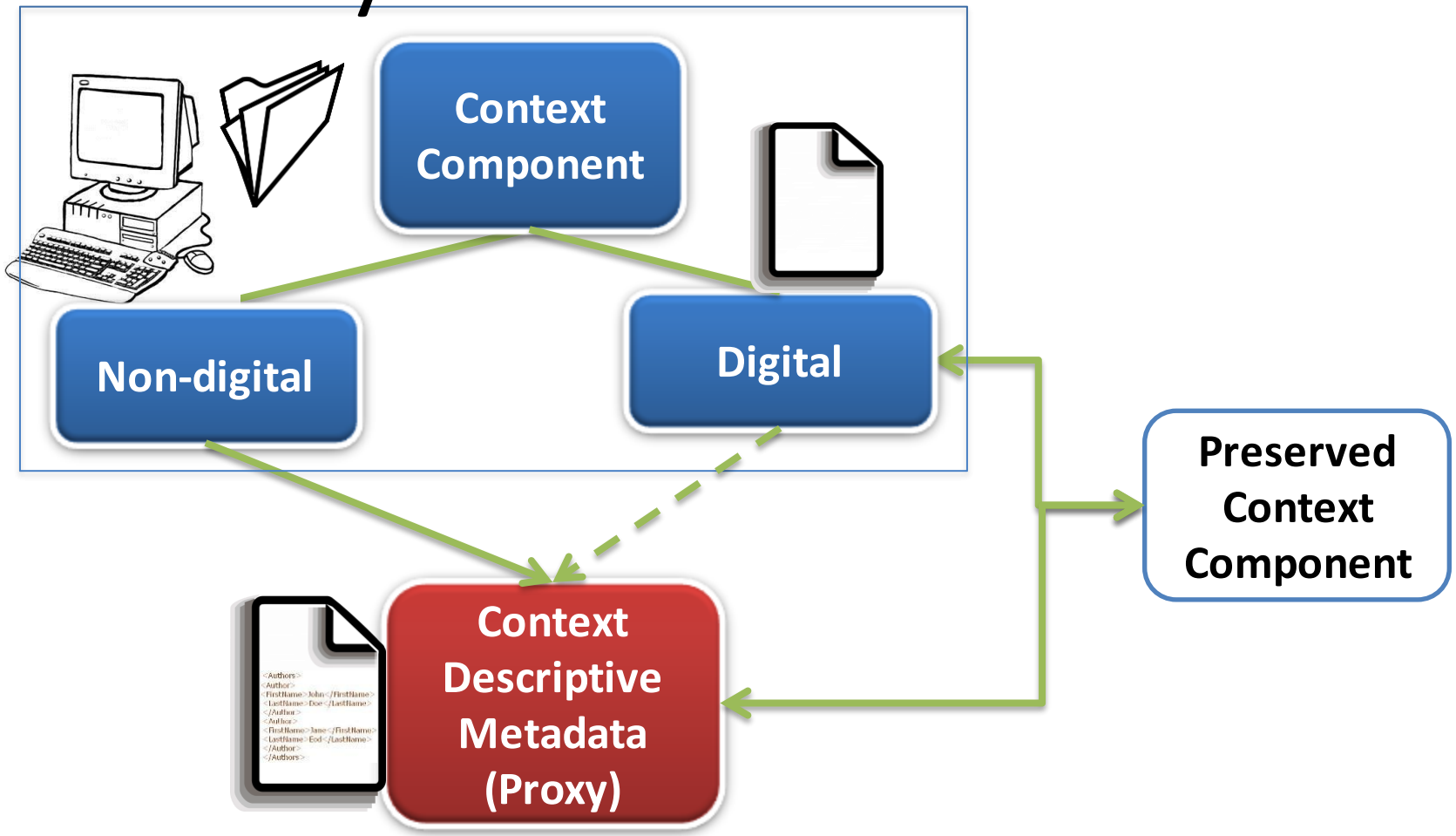
Questions

Any Questions ?

Vision



Context Components and Proxys



Business Process Contexts Model

- Business Process Context is based on a Business Process Context Model
  - a formal meta-model
  - can be instantiated
  - captures the relevant aspects of a business process and supporting software/technology
  - enables business process redeployment

Enterprise Risk Management

- Aim: Prevention and control mechanisms to address risks attached to specific activities and/or assets
- Risk = undesirable outcome posing a threat to the achievement of objectives, e.g.
  - Financial risks: e.g. credit, market risks
  - Operational risks: e.g. IT risks

- ERM = enterprise-wide approach to Risk Management
  - Looks at risks holistically
  - TIMBUS focuses on business processes “as a whole” and the risks affecting their operability

Align DP and Risk Management

Conceiving Digital Preservation as a risk mitigation action helps to

Legalities
Life Cycle

- Intellectual copyright
  - Information Society Directive (“works“)
  - Computer Program Directive
  - Database Directive
- Data protection laws

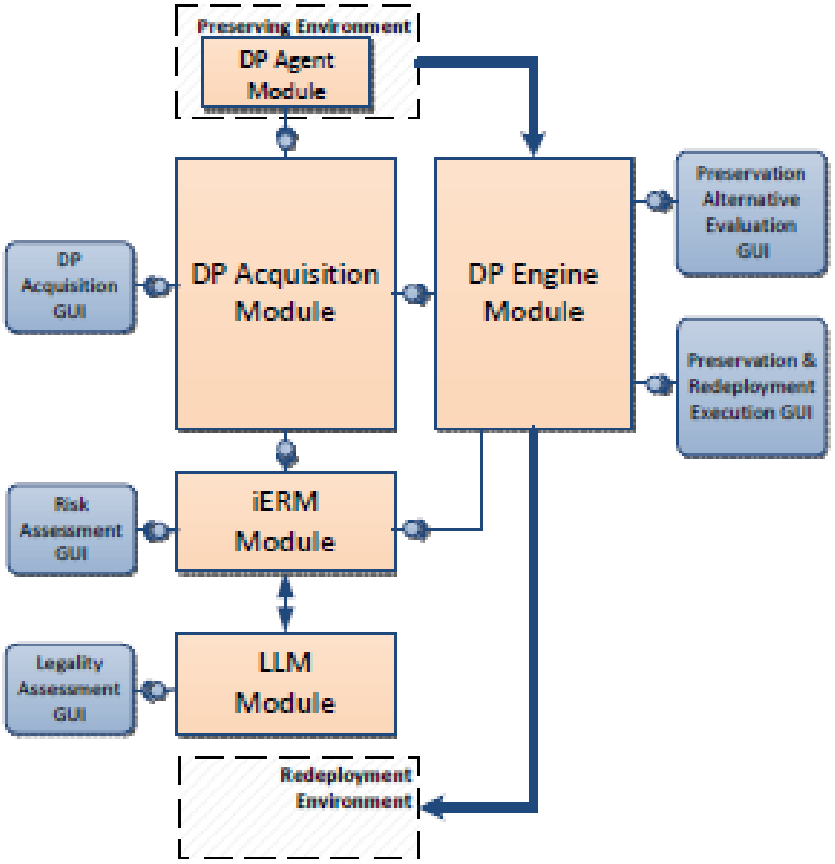
- Impact on
  - Backups, data carrier renewal (reproduction)
  - Migration, decompiling, … (alteration)
  - Retention (e.g. personalised data)

Legalities
Life Cycle

- Data format conversion
  - Lossless ➔ equivalent to reproduction
    - Might be ok for private purposes, but not for business
  - Lossy ➔ infringment of exclusive right of alteration

- Preservation of software
  - Might require porting, decompiling, …
  - Infrignement on right of adaption

- DP requires special-purpose license agreements


Preserving Environment
DP Agent Module
DP Acquisition GUI
DP Acquisition Module
DP Engine Module
Preservation Alternative Evaluation GUI
Preservation & Redeployment Execution GUI
Risk Assessment GUI
iERM Module
Legality Assessment GUI
LLM Module
Redeployment Environment