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 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** (i.e., Cloud is for storage and **compute**)

7/22/2014
DP Futures

‘Digital Universe’ Nears A Zettabyte
May 4th, 2010 : Rich Miller

Figure 1: The Digital Universe 2009 – 2020
Growing by a Factor of 44

2020
35 ZB*

2009
0.8 ZB*

*Zettabyte = 1 million gigabytes

Source: IDC Digital Universe Study, sponsored by EMC, May 2010

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.

... it’s not going to be about obsolescence so much as workflow and capacity

7/22/2014
Virtualisation and Digital Preservation
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/](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?)
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

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

• Virtualization
  o Virtualisation puts a layer between physical hardware and controls access to that machine.
  o 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.
  o 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.

<table>
<thead>
<tr>
<th>Applications</th>
<th>Applications</th>
</tr>
</thead>
<tbody>
<tr>
<td>Operating system</td>
<td>Operating system</td>
</tr>
<tr>
<td>Emulator</td>
<td>Emulator</td>
</tr>
<tr>
<td>Operating system</td>
<td></td>
</tr>
</tbody>
</table>

• Emulation
  o 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

TYPE 1 native (bare metal)

Hosted

TYPE 2 (hosted)
Emulation strategies for LDP

1. SW Emulation for HW

<table>
<thead>
<tr>
<th>Original situation</th>
<th>Emulated situation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original digital document</td>
<td>Original digital document</td>
</tr>
<tr>
<td>Original rendering software</td>
<td>Original rendering software</td>
</tr>
<tr>
<td>Original operating system</td>
<td>Emulator</td>
</tr>
<tr>
<td>Original hardware</td>
<td>Future hardware</td>
</tr>
</tbody>
</table>

2. Stacked Emulation

Source: Jeffrey van der Hoeven

3. Migrated Emulation over time

Source: (Jeffrey van der Hoeven)
Examples

- Hypervisors
  - KVM
  - Xen
  - Windows Server
  - VMware
  - Hyper-V
- Emulators
  - QEMU
  - 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 BUusiness processes and Services.
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
Topics

- Motivation
- Objectives
- TIMBUS Process
- Architecture
- Context Model
- Owl Ontologies
- Tools
- Demo
# Comparison of Cloud Providers

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

**TIMBUS targets this**

---

22 July 2014  timbusproject.net © 2013  29
TIMBUS Why

PRIMARY MOTIVATIONS

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

Context Model + Dependency Model

Overview

Risk assessment
Monitoring & review
Establish the context
Risk treatment

Context model

Reasoning

Technical Framework for ESCROW

Context and dependencies

Process Flow

Risk process

Software ESCROW

TIMBUS Process Framework

Risk assessment
Establish the context
Risk treatment
Monitoring & review

Overview

Establish the context
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.

<table>
<thead>
<tr>
<th>Question</th>
<th>List of</th>
</tr>
</thead>
<tbody>
<tr>
<td>Which business actors are required to execute business process P?</td>
<td>List of Business Actors</td>
</tr>
<tr>
<td>Which technological entities support business process P?</td>
<td>List of structural and behavioural technological entities</td>
</tr>
<tr>
<td>Which application components support business process P?</td>
<td>List of application components</td>
</tr>
<tr>
<td>Which legal requirements are verified by business process P?</td>
<td>List of legal requirements</td>
</tr>
<tr>
<td>Which are the licenses required to execute software application S?</td>
<td>List of licenses</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Context Model Domains

- **Strategy**
  - Strategic Indicators, External Services, Contracts, Regulations, Licenses, Legal Requirements, Patents

- **Business**
  - Organizational Structure, People, Business Processes, Operational Indicators

- **Applications**
  - Applications, Services, Virtualization Applications

- **Technological Infrastructure**
  - Deployed software applications and services, Hardware nodes, Communication nodes

- **Organization**
  - Information
  - Processes

- **Components**
  - Services

- **Processes**
  - Technologies
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

1. Load
2. Annotate
3. Simulate
4. Risk Assessment Web-Service
5. Preservation recommendations
6. Preservation

BPMN
Context Model Instance
iERM
Annotate
Business Impact Data
DPES
Request
Request

timbusproject.net © 2011-2014
Context Acquisition Framework

Run:
- Different extractors
- Different environments
- Different QoS
- Precedences

Configuration Business Environment

DIO+DSOs
Extractor – Linux HW

Ishal  Ispci  Iscpu  dmidecode

Ishw

Target Machine

CPU  Memory Storage  RAM  Network

Remote/Local Extraction

GUI

OWL  Context Model Instance

Hardware DSO  JSON

timbusproject.net © 2011-2014
Extractor – OS Resource Monitor

Business:
- PID1
- PID2
- PID3
- ...

Target Machine
- forks
- open files
- libraries
- sockets

Remote/Local Extraction

GUI

JSON

DSOs

OWL

Context Model Instance

Linux OS ProcFS

proc/$pid/task
proc/$pid/fd
proc/maps
proc/net/*
DPES - Implementation

Digital Preservation Expert Suite Architecture

- Web Services
  - iERM (T6.1), Context Module (T6.5), etc.
- Web service
- Storage Service
  - HTTP Restful SWIFT API
- Timbus Storage Repository (T5.3)
- Preservation Manager Application Server (T6.4)
- Network
  - Preservation Service
  - VM1, VM2, ..., VMn
  - Emulation/Virtualization Server
  - Business Process Re-deployment Environment (6.6)

DPES User

timbusproject.net © 2011-2014
DPES - Internals

Users

TIMBUS Services

UI Components

Service Interfaces

Web UI

Cost Modelling
Preservation identifier
Feasibility Analyser
Application Servlet

Metadata Extraction
Business Process Artefacts Sourcing
Redeployment Manager

Business Objects

Data Access Components
Storage Repository proxy
Redeployment Agent

Data Layer

DPES Registry

Cloud Storage

Redeployment Environment (VM, Emulators, Containers)

Cross-Cutting

Security (user access control)

Operational Management

Cloud

Feasibility Analyser
Metadata Extraction
Redeployment Agent

Data Access Components

TIMBUS project.net © 2011-2014
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/ubuntu 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-...

Bins/Libs  Bins/Libs

Docker Engine

Redeployment System
(Host OS -Ubuntu 12.04)

Virtualization Server
(Proxmox - KVM)

Timbus Managed Repositories

Linux Packages

Business Artefacts
DPES – Redeployment Model-2

![Diagram showing the Redeployment Model-2]

- **Business Apps**
  - VM-1
  - ----
  - VM-n

- **KVM/QEMU**

- **Virtualization Server (Proxmox)**

- **OS Images**
  - ISO Images of Operating Systems
  - Dependency libs

- **Business Artefacts**

- **Timbus Managed Repositories**
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?
Preservation of Business Processes

Redeployment of Business Processes

New OS

New Device
Context Components and Proxys

Non-digital

Non-digital

Context Component

Digital

Preserved Context Component

Context Descriptive Metadata (Proxy)
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

Align DP activities with organizational strategy and objectives.

Rationalize DP from a Risk Management perspective.
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 $\Rightarrow$ equivalent to reproduction
    • Might be ok for private purposes, but not for business
  – Lossy $\Rightarrow$ infringement of exclusive right of alteration

• Preservation of software
  – Might require porting, decompiling, ...
  – Infringement on right of adaption

• DP requires special-purpose license agreements