Digital Repository on Cloud Infrastructure
Issues & Challenges

Dr. Mukul K Sinha

Expert Software Consultants Ltd., New Delhi
mukulks@yahoo.co.in

Digital Repositories

Threats

- **Malicious Attacks / Unintended Mistakes**
  – *Elaborate Security Regime*
- **Accidents & Natural Disaster**
  - *Disaster Recovery Sites*

- **++ Technological Obsolescence** **- Since 90s**
  - *of <u>Media</u>, <u>H/w</u>, <u>S/w</u> & <u>Middleware</u> Platforms ??*

:: *<u>Unaware of Threat</u> of Technological Obsolescence*

Technological Obsolescence

- **Degradation & Obsolescence of Digital Media**
- **Proprietary Format**

  Withdrawal of Format, or Company Closure!
- **Obsolescence of H/w, S/w & Mid/w Platforms**

  Perpetual in nature

→ ***Total & Irrecoverable Loss of Digital Records***

→ ***Digital Vulnerability***

Effects of Digital Vulnerability

Digital Black-hole:

Un-aware **of Irrecoverability of Digital Records**

- **Loss of Asset / Information /Records / History**
- **Loss of Credibility**

- **Financial Loss**
- **Non-Fulfillment of Statutory Requirements**

→ → ***Need of Long Term Digital Preservation***

Long Term Digital Preservation

Social Needs /Business Needs / Statutory Needs

Digital Preservation

- **Long Term Eligibility**
- **Continued Accessibility, and Understandibility**

Requires

- **Additional Administrative Infrastructure**
- **Additional Technical Infrastructure**
- **Additional Financial Investment**

Long Term Digital Preservation

Bit Preservation *Media Degradation / Obsolescence*

- **Data Refreshing** *Media Degradation*
- **Data Migration** *Media Obsolescence*

Logical Preservation *Format /Platform Obsolescence*

- **Data Transformation** *New Format / Platform*

  → **Bit Mutation**

  Need to Assure Readability / Understandability

Open Archival Information System (OAIS)

OAIS Reference Model:

Blue Book 2002/ Pink 2009/ Magenta Book 2012

- **Archive Information Package (AIP)**
  (Content. Info. + Preservation Description Info.)

- **Content Info.** *(Content Data Object + Rep. Info.)*
- **Preservation Description Info.** *Logical Preservation*
  (Ref. Id., Provenance, Context - AIPS, Fixity)

Dig. Preserv. → *Provenance Tracking* + *Fixity Checking*

Long Term Dig. Preservation: Approaches

- **Community Based Digital Repositories**
  - *Common Domain, Common / Shared Funding*

- **Trusted Digital Repository (TDR)** (OAIS Compliant)
  - *Specialized Service Organization*
  - *Service to Customers against a Fee*
  - *Deep Infrastructure (Distributed / Multi-Site)*
  - *Coupled with other Trusted Digital Repositories*
    (Trusted Inheritor of Repository – Closure!)

→ **Dig. Preserv. Infrastructure: *Coop. Network of TDRs***

Audit & Certification of TDR
Criteria & Checklist

- **Objective Criteria**

  Documentation / Transparency / Adequacy / Measurability

- **Certification Process Checklist**

  Organizational Infrastructure / Digital Object Management / Technologies, Technical Infrastructure, Security

  - Initial Certification / Periodic Survell. / Audit Re-Certification

- ***Certification of O*rganizational Archival Program**

  Archival Process /Archival Data /Archival Staff

Trusted Digital Repository: Roles

- **Core Business: Digital Object Management**
  - *Archivists / Information Scientists*
- **Backend Service: Data Center Management**
  - *Computer Scientists*

?? Can Backend Service <u>Out-Sourced</u> <u>Economically</u>

- Cloud Service Provider

Cloud Service Providers: Offerings

- **On Demand Service** - ***Through APIs***
- **Network Access** - ***Through Internet***
- **Shared Pool of Resources**
  - ***Storage / Compute / Appl.*** **Software**

- **On-the-fly Rapid Provisioning –** ***Scale up/down***
- **Measured Service** - ***Pay as you go***
- **No Up-front Commitment**

- **Multi-Site Distributed System –** ***Site of your choice***

Storage As A Service (SaaS): Amazon

- **Simple Storage Service (S3)**
  - Reliability – 99.99% (Four 9s)
- **Elastic Block Storage (EBS)**
  - Filing System can be mounted
  - Accessible by Virtual Machine (IaaS)

- **Transfer Speed** ***- S3: as local disk, slower with VM***
  - EBS: as local disk for VM

- **Charging** ***- Volume * Time***
  + Data Transfer – From / to SaaS

Cloud Storage Service for Dig. Repository

- **Data Integrity** - ***Checksum / User run Checksum***
- **Data Confidentiality** – ***Data Encryption***
- **Data Availability** - ***Four 9s / Eleven 9s***

- **Data Portability** - ***?? Interoperability***
  - Cloud Broker Service
- **Data Audit** - ***?? Full Audit – Data Transfer Cost!***

- **Operational Cost - ?? LOCKSS vs. Amazon - High**

Specificity of Digital Repository

- **Data Loss Unacceptable - *Multi-site Replication***
- **Data Access Infrequent – *Lower Availability OK***
  No Hot Spot
- **Handling Large Number of Heterogeneous Systems**
  - ***Large Volume***
  - ***Slow Growth over Time***
  - ***Multiple Platforms / Technology***
  - ***Multiple Vendors / Formats***

Functional Requirements for Logical Preservation

- **Data Migration across Platforms**
  - *Technological Obsolescence*
- **Maintaining Provenance**
  - *Change in PDIs*
- **Fixity Computation**
  – *Change in AIPs*

→ Compute Intensive
→ → Preservation Aware Storage System

Preservation Data Store: PDS

PDS: Infrastructure Component of CASPAR
with Preservation Aware Object Store (OSD)

- **Encapsulate Raw Data with large Metadata**
  - Provenance, Rep Info., Fixity
- **Creates AIP and associated AIPs**
- **Transforms AIP into Physical Storage Object**

- ***Preservation Related Functions within Storage***
  - Media Migration, Provenance Updates, Fixity Computation

Preservation Data Store on Cloud

PDS Cloud - ***Offers Preservation Service to Customer***

- ***Avails <u>Storage Service from Cloud Service Providers</u>***

Decomposed OSD into

- Preservation Engine + Multi-Cloud Service

- **Preservation Engine**

  - Preservation Functions to AIPs

- **Multi Cloud Service**

  - Heterogeneous Cloud Storage

  - Amazon, Rackspace, Open Stack

→ **<u>Cloud Broker Service</u>!**

Preservation Vendors as Cloud Broker

Preservation Vendors

- ***Offers (limited) Preservation Service to Customer***
- ***Avails <u>Storage Service from Cloud Service Providers</u>***

- **PDS Cloud**
- **DuraCloud** ***- Storage – Amazon, Rackspace, SDSC Cloud***
  - ***Preservation – Replica on Different Cloud (Synchronized)***
  - ***Data Migration across Cloud Vendors***
  - ***Data Integrity – Checksum Verification***
  - ***Health Checkup – Checksum Recalculation of all contents at pre-specified time – report on dashboard***

?? Organizational Closure / Failure

Preservation Vendors with Cloud Storage

Preservation Vendors

- ***Offers (limited) Preservation Service to Customer***
- ***Offers <u>Multi-site Cloud Storage Service</u>***
- ***<u>Niche market</u> of Long Term Digital Preservation***

- **Preservica**

  Preservation Functions:

  - ***Multi-site Replica (cross-checking & recovery)***
  - ***User can store with metadata and data tags***
  - ***User Access to Data as well as Metadata***
  - ***Tools for Data Migration across various Formats***
  - ***….***

?? Organizational Failure ? Closure

Cloud: Digital Preservation as a Service

Near Future Scenario

- **Number of Trusted Digital Repositories to Grow**
- **Repository Audit**
  - ***Compute Intensive & Data Transfer Intensive – Very Costly***
- **Long Term Digital Preservation Market to Grow**

→ Preservation / Audit Functionalities by Storage Service

→ → Digital Preservation as a Service (DPaaS) by Cloud Service Providers!!

Thank You