

## Contribution to JTC 1 Study Group on DCOMP

Title: On Designing an Interoperability Framework for Digital Preservation

Status: Contribution

Author: Rainer Schmidt (AIT)

Date: 10-08-2010

### 1. Introduction

Digital preservation has been recognized as an important challenge across many different areas of society. The problem is caused by various reasons like the fast growth of humanly produced digital materials [1], enormous amounts of scientific data generated by digital instruments [2], as well as computer hard and software that is subject to periodic changes.

Digital preservation aims to ensure long-term access to existing and future digital holdings. This includes preservation of the physical bit-streams (e.g. through redundant and distributed storage using data grids<sup>1</sup>) as well as preservation of the logic that is required to understand and interpret the digital materials (e.g. the Planets Suite<sup>2</sup>, the Representation Information Repository<sup>3</sup>).

### 2. Open Archival Information System

Research in this area has been significantly driven by the archival community. A prominent outcome is the establishment of the Open Archival Information System (OAIS) reference model (ISO 14721:2003). OAIS provides a conceptual model and vocabulary for archival systems identifying commonly required system components, information packages, as well as their interaction. The model however does not provide a technical specification against which a system could be validated.

Some OAIS Concepts that might be interesting for creating a Digital Preservation Interoperability Framework (DPIF):

- The specification defines following six functions/interfaces: Ingest, Archival Storage, Data Management, Administration, Preservation Planning, Access.
- It further addresses the migration of information and names different types of possible migrations: Refreshment, Repackaging, Transformation, Replication,
- It identifies different levels of interaction between OAIS instances: Independent, Cooperating, Federated, Shared resources.

---

<sup>1</sup> <http://www.sdsc.edu/srb/index.php>

<sup>2</sup> <http://sourceforge.net/projects/planets-suite/>

<sup>3</sup> <http://registry.dcc.ac.uk:8080/RegistryWeb/Registry/>

### **3. Applying Preservation Actions**

Logical digital preservation pertains to the continuous maintenance of content and its associated metadata in order to ensure long-term interpretability of digital information. Preservation actions are data management operations that support the technical long-term viability of digital materials, in particular regarding changing of formats and platforms. This includes the specification and establishment of required technical conditions (hardware/software), the prevention of format obsolescence (through migration and emulation), as well as the preservation of significant characteristics (like for example color, size, look-and-feel).

The requirement for applying preservation actions to archived content is implicitly covered by the OAIS component Preservation Planning. The model however does not detail how a preservation plan (a sequence of preservation actions) is applied to an archival system. Possible approaches might be: (1) to access and process the content outside the repository, followed by data re-ingest and/or metadata enrichment; (2) to integrate a process execution engine into the repository system that directly manipulates the content. Possible implications are that the former approach might not scale well for larger amounts of data while the latter approach might raise concerns regarding the trustworthiness of a repository system. Another key issue is the provision of quality assurance for automated preservation strategies.

In order to support preservation actions, the minimal requirements for a preservation archive are support for re-ingest/versioning and support for associated preservation metadata. This is however not facilitated by most commercial products on the market hindering the application of preservation strategies within archival institutions. A proof-of-concept implementation that integrates this functionality with the Fedora open-source repository system has been implemented in the context of the Planets Interoperability Framework (IF)<sup>4</sup>.

We see a clear need for standardization regarding the application of preservation actions with OAIS environments as well as for guidelines that detail how this can be supported by existing repository software.

### **4. Interoperability**

Depending on the user community, repository systems have been designed for the archival of a broad range of different types of digital information. As a matter of fact, the various implementations often differ significantly regarding interfaces, data modeling, and representation, even if their core functionality might be very similar.

---

<sup>4</sup> <http://www.ijdc.net/index.php/ijdc/article/viewFile/157/220>

This has caused lacking interoperability between different repositories as well as between the components that make up a repository system.

#### **4.1 Repository Infrastructures:**

The need for interconnecting existing repositories – both with similar as well as complementary types of content (e.g. linking publications with scientific data) – has been widely recognized. Consequently, a range of infrastructure programs and projects have been established that aim at interlinking a variety of information sources.

##### *Some Initiatives:*

- UK JISC Information Environment Architecture<sup>5</sup> and Data Management Infrastructure Programme,
- EC infrastructure projects like Driver, Sherpa, TEL
- Scientific data infrastructures: GENESI-DR (distributed earth science data), Tardis (federated X-Ray Image repository), I2S2 (Integration in Structural Sciences)

Interoperability regarding information exchange between repository systems and/or their client applications will require standard information exchange protocols and formats. Many existing open repository systems support access based on OAI-PMH<sup>6</sup> metadata-harvesting. However, the potential for automatic record exchange based on this protocol is very limited. Open-source systems (Fedora, DSpace) typically provide public APIs (e.g. web, file, web service-based) for ingest or access. However, the representation, syntactical, and operational characteristics differ between the various systems.

Standardized data exchange mechanisms might greatly improve the interoperability of preservation systems. For DPIF, it will be important to investigate interfaces (and protocols) that are specifically designed for automated data exchange between repositories. These may have to support different acquisition and delivery methods like *harvest*, *search*, *subscribe*, or *notify*.

A minimal, extensible specification that can be easily supported by commercial vendors would also be desirable. The Planets IF, for example, utilizes the concept of *Digital Objects* (a minimal data abstraction that encapsulates a repository record) as well as *Digital Object Managers* (providing standardized access mechanisms to repository interfaces) as wrappers in order to support the integration of a range of existing repository systems.

#### **4.2 Repository Components:**

---

<sup>5</sup> <http://www.ukoln.ac.uk/distributed-systems/jisc-ie/arch/>

<sup>6</sup> <http://www.openarchives.org/OAI/openarchivesprotocol.html>

Interoperability may also pertain to the internal components of an archival system in order to support a modularity and exchangeability. It would be desirable to specify the interfaces provided and/or required by the main components a preservation system consists of. The Storage Networking Industry Association (SNIA), for example, develops such standard interfaces for establishing interoperability with different types of storage components.

### 4.3 Identification

In a distributed system, it is of crucial importance to define the naming system that is used to identify and locate concepts like services, data, or events. This is typically implemented based on a naming schema as well as registries, and metadata catalogues. Examples are the Domain Name System (DNS)<sup>7</sup>, Data Management and Naming Conventions in Grid Computing<sup>8</sup>, the ICAT<sup>9</sup> catalogue for facility science data, the PLANETS Technical Registry.

Below, is a short summary how identification is ought to work using the Planets IF.

#### *Description of the Work*

Describes the work that a Digital Object is a manifestation of (not the physical representation). Bibliographical metadata can be expressed using the Dublin Core schema. The ID of a work should be a Uniform Resource Name (URN), e.g. a UUID.

#### *Description of Representation*

Is done based on a global/permanent identifier (e.g. DOI), which can be resolved into a locator, represented by a repository identifier. The Identifier should be expressed as a Uniform Resource Locator (URL).

#### *Description of Concepts*

Planets-specific concepts like events, services, characteristics are described using a metadata registry (e.g. the PRONOM<sup>10</sup> registry for file formats). The Identifiers are expressed as system-specific URIs (like *planets://repository/event/ingest/* for an event or *info:pronom/fmt/122* for a specific file format).

## 5 Recommendation

Many approaches and infrastructures for federating distributed information sources exist. A focus on the technology and user community that are addressed by DPIF might be required, e.g. OAIS-type digital libraries/archives.

---

<sup>7</sup> <http://www.ietf.org/rfc/rfc1034.txt?number=1034>

<sup>8</sup> e.g. [http://wiki.egee-see.org/index.php/SG\\_Data\\_Management\\_File\\_Names\\_and\\_LFC](http://wiki.egee-see.org/index.php/SG_Data_Management_File_Names_and_LFC)

<sup>9</sup> <http://code.google.com/p/icatproject/>

<sup>10</sup> <http://www.nationalarchives.gov.uk/PRONOM/Default.aspx>

Many established schemas for describing different types of data exist, e.g. for web, scientific, technical, financial, administrative, scholarly, historical data. Depending on the domain, the required descriptive, technical, and also preservation metadata might also vary greatly. Thus, a specification of the internal data structures of an archive will be very difficult to accomplish and might not be desirable.

DPIF should establish a layer that resides atop content dependent formats and descriptions. There is a clear need for technical specifications that define interaction methods with/among repository systems, and ways to exchange and map between diverse digital object representations, independently from concrete and content dependent description schemas. It will be important to describe the required interfaces, information containers, and transformation protocols, rather than prescribing specific metadata schemas. The design should consider existing approaches for supporting interoperability between repository systems (like OAI-ORE or the Fedora Content Model Architecture<sup>11</sup>).

## 6. References

[1] John F. Gantz, Christopher Chute, Alex Manfrediz, Stephen Minton, David Reinsel, Wolfgang Schlichting, Anna Toncheva, The Diverse and Exploding Digital Universe - An Updated Forecast of Worldwide Information Growth Through 2011, An IDC White Paper - sponsored by EMC, March 2008. Available at: <http://www.emc.com/collateral/analyst-reports/diverse-exploding-digital-universe.pdf>

[2] T. Hey, A. Trefethen, The Data Deluge: An e-Science Perspective, in Grid Computing - Making the Global Infrastructure a Reality, 2003. Available at: [http://eprints.ecs.soton.ac.uk/7648/1/The\\_Data\\_Deluge.pdf](http://eprints.ecs.soton.ac.uk/7648/1/The_Data_Deluge.pdf)

[3] Avoiding Technological Quicksand: Finding a Viable Technical Foundation for Digital Preservation, Jeff Rothenberg, 1999. Available at: <http://www.clir.org/pubs/reports/rothenberg/contents.html>

---

<sup>11</sup> <http://www.fedora-commons.org/documentation/3.0b1/userdocs/digitalobjects/cmda.html>