



# Towards a Global Digital Format Registry

---

David Seaman  
Executive Director  
Digital Library Federation

**DPC Forum: *Digital Preservation – the global context.***  
**Wednesday 23rd June 2004, the British Library Conference Centre**



# Digital Library Federation

<http://www.diglib.org/>

---

- ❑ Thirty-three members – major academic and national libraries, including The British Library; four allies (CNI; RLG; OCLC; LANL)
- ❑ Created in 1995 by directors of US research libraries; fills a need not simply met by larger library organizations: focus exclusively on DL needs and strategies for large libraries
- ❑ Be nimble, agile, collaborative
- ❑ Practical and strategic areas of activity



# DLF Work -- background

---

## ***USER SERVICES***

- ❑ *Dimensions and use of the scholarly information environment* [www.diglib.org/pubs/scholinfo](http://www.diglib.org/pubs/scholinfo)
- ❑ IMS – learning technologies and courseware integration
- ❑ Distributed single collection of our own material

## ***METADATA STANDARDS***

- ❑ Open Archives Initiative support
- ❑ METS (Metadata Transmission Standard)



# DLF Work -- background

---

## ***RESOURCE MANAGEMENT***

- ❑ XML format for license content
- ❑ Registry of Digital Masters

## ***PRODUCTION***

- ❑ Production standards and benchmarks

## ***PRESERVATION***

- ❑ Journals preservation –  
[www.diglib.org/preserve/ejp.htm](http://www.diglib.org/preserve/ejp.htm)
- ❑ Registry of Digital File Formats



# Why Do We Need a Registry?

---

“In order for a document to be readable in the future, two conditions must be met. First, the bits that constitute the document must be readable from the medium and transferable to a computer memory. **Then, software must be available to interpret the data.**”

*A Project on Preservation of Digital Data.* Raymond A. Lorie. <http://www.rlg.org/preserv/diginews/diginews5-3.html#feature2>



# Why Do We Need a Registry?

---

- ❑ Once you have retrieved your document, you need to know something authoritative about its format in order to load, convert, or emulate it. Where do you find that information?
- ❑ Format is central to the workings of a DL, preservation program, or institutional repository -- repository functions are performed on a format-specific basis, for example.



# Why do we need a *Global* Registry

---

- ❑ We are beginning to build local format registries and this is clumsy, duplicative, and miserable.
- ❑ A shared global registry is a core infrastructure component of a distributed program for preservation and *data viability*
- ❑ We have a great will to solve core issues collaboratively
- ❑ It provides a means for *more projects* to have *more sophisticated information* about *more formats*



# What's Wrong with MIME Types?

---

- ❑ **Insufficient depth of detail and insufficient granularity**
  - Both tiled RGB TIFF with LZW and striped bi-tonal TIFF with Group 4 → image/tiff
  - All of PDF 1.0 – 1.4, PDF/X-1 – 3, and PDF/A → application/pdf





# Background

---

- During summer 2002 the Harvard LDI and MIT DSpace teams met to discuss shared concerns.
- DLF-sponsored invitational meetings, growing out of discussions at the Fall 2002 DLF Forum
- DLF committee 2003-2004
  - Collected use cases
  - Working groups on data and governance models
  - “Strawman” registry set up at the University of Pennsylvania
  - Funding talks underway to build this out



# Ad-Hoc Committee

---

- |                                                           |                                                      |
|-----------------------------------------------------------|------------------------------------------------------|
| <input type="checkbox"/> Bibliothèque nationale de France | <input type="checkbox"/> NARA                        |
| <input type="checkbox"/> British Library                  | <input type="checkbox"/> National Archives of Canada |
| <input type="checkbox"/> California Digital Library       | <input type="checkbox"/> New York University         |
| <input type="checkbox"/> Digital Library Federation       | <input type="checkbox"/> NIST                        |
| <input type="checkbox"/> Harvard University               | <input type="checkbox"/> OCLC                        |
| <input type="checkbox"/> IETF                             | <input type="checkbox"/> Public Records Office, UK   |
| <input type="checkbox"/> JISC                             | <input type="checkbox"/> RLG                         |
| <input type="checkbox"/> JSTOR                            | <input type="checkbox"/> Stanford University         |
| <input type="checkbox"/> Library of Congress              | <input type="checkbox"/> University of Pennsylvania  |
| <input type="checkbox"/> MIT                              |                                                      |



# Global Digital Format Registry

---

The registry will maintain persistent, unambiguous bindings between public *identifiers* for digital formats and *representation information* for those formats.

A format is a fixed, byte-serialized encoding of an information model, e.g. PDF; TIFF.

The registry is an enabling technology underlying many digital repository operations and preservation activities



# Potential Use Cases

---

- **Identification**

- “I have a digital object; what format is it?”

- **Validation**

- “I have an object purportedly of format  $F$ ; is it?”

- **Transformation**

- “I have an object of format  $F$ , but need  $G$ ; how can I produce it?”



# Potential Use Cases

---

## □ **Characterization**

- “I have an object of format  $F$ ; what are its significant properties?”

## □ **Risk assessment**

- “I have an object of format  $F$ ; is at risk of obsolescence?”

## □ **Delivery**

- “I have an object of format  $F$ ; how can I render it?”



# Informative, not Evaluative

---

The format properties stored in the registry should be factual, not judgmental.

- ❑ Legal liability
- ❑ May discourage deposit of proprietary information
- ❑ Investigate ways to include (by reference?) third party evaluations/recommendations
  - Insofar as this doesn't hamper primary goal



# Data Model Informed by Prior Work

---

- ❑ **ISO 14721, Open archival information system -- Reference model**
  - CCSDS OAIS reference model
  - Representation information
    - ❑ Interpret, or provide “additional meaning” to Data Object
    - ❑ Structure and semantic information
- ❑ **PRONOM**
  - Public Records Office, UK
  - “information about file formats and the application software needed to open them”
  - Format, vendor, product



# Data Model Informed by Prior Work

---

## □ **Diffuse**

- EC's Information Society Technologies programme
- “reference and guidance information on available and emerging standards and specifications”
- Business Guides
  - “application of standards and specifications in specific areas”

## □ **OCLC/RLG Preservation Metadata Framework**

- “information necessary to render/display, understand, and interpret the Content Data Object”
- Based on CEDARS, NEDLIB NLA, OAIS, and OCLC metadata





# Data Model Informed by Prior Work

---

- JISC File Format Representation and Rendering Project
  - Assessment of formats and rendering software
  - Representation system to track formats and their rendering software
- NIST National Software Reference Library
  - File profiles for the NSRL Reference Data Set
    - Vendor, product, operating system
  - Used for forensic identification



# Core Registry Services

---

## □ Management Services

### ■ Approval

- Level of review, level of public disclosure

### ■ Maintenance

- Add, update, delete format entries

### ■ Notification

- Notify registry clients of new/updated format or trigger events (e.g. obsolescence, new transformation service, etc.)

### ■ Introspection

- Determine local policies (scope, coverage, implemented services, etc.) of a given registry to identify appropriate registry to use



# Core Registry Services

---

## □ Access Services

### ■ Description

- Representation information returned on request for single format

### ■ Export

- Entire registry or selected subset sent to external repository
- Export is critical both for some services, for preservation, and for engendering trust



# Supported Services

---

## □ Representation Services

### ■ Identification services

- Determine format of a specific digital object by comparing its attributes to the attribute profiles retrieved from the registry

### ■ Validation services

- Verify format of a specific digital object (DO) by comparing its attributes to the attribute profile retrieved from the registry for that format.



# Supported Services

---

## ❑ Brokerage Services

### ■ Rendering service

- ❑ Identify current rendering conditions for supplied digital object

### ■ Transformation service

- ❑ Convert digital object from current (source) format to target format

### ■ Metadata Extraction services

- ❑ Registry returns information supporting automated extraction of attribute metadata from a digital object of a specific format



# Registry Operation

---

A global registry is valuable when it is *trustworthy* and *sustainable*.

- ❑ Trust is necessary to encourage deposit of proprietary information
- ❑ Sustainability is necessary to justify expense
  - As for all preservation activities, how do we generate income today, for services not needed until tomorrow?
  - In this case, we see short-term services as well.



## Next (3 years' duration)

---

- ❑ Secure funding.
- ❑ Conduct a formal study of the options for hosting, governance, staffing, and financing the registry.
- ❑ Develop working prototype of the registry, including critical mass of format entries. Test interoperability with a set of distributed digital preservation systems.
- ❑ Move prototype into a production service.