International Conference on Digital Libraries (ICDL)
February 2004, New Delhi, India

Practical Considerations in Creating and Providing Access to Digital Repositories

Dr. Vinod Chachra, CEO
VTLS Inc.
Blacksburg, VA, USA
www.vtls.com
VTLS is an International Company
Corporate HQ in Blacksburg VA
VTLS is used 35 Countries
VTLS has offices in 7 Countries:
Brazil, India, France, Malaysia, Spain, Switzerland and USA
About VTLS (2 of 3)

VTLS has Three Products (many services)

Virtua -- ILS (Intelligent Library Solutions)

Vital – for Digital Libraries

Vtrax -- Radio Frequency Id (RFID)
About VTLS (3 of 3)

Examples of VTLS Users in India

University of Hyderabad
National Library of India (Calcutta)
CIIL, Mysore
Dera Community Library, Punjab
Indian School of Business, Hyderabad
Bhoj College of Engineering for Women
South Asia International Institute
Since ICADL 2000 - Bangalore

• VTLS has worked on more than 80 Digital Library Projects. Examples are:
  – Steven Spielberg Digital Library Project
  – University of Maine Music Project
  – Texas Tides Historical Project
  – New York Public Library Digitization Project
  – AMICO – Art Museum Image Consortium
  – NDLTD – Networked Digital Library for Theses and Dissertations

• Gained a lot of experience in what works and does not work
• Developed expertise on what to do and what to avoid
• Developed a new product called VITAL to incorporate these ideas
• Developed services to take advantage of the experience and the products
• This is what I hope to share with you today
Practical Considerations

• Part 1: Metadata

• Part 2: Digital Object Repository

• Part 3: Putting it together – VITAL
  • Workflows
  • Web Delivery System

• Part 4: Working with vendors
Disclaimer

• This presentation is based on information from a variety of sources; Not all sources have been identified in this “version” of the presentation.

• We will see later that “versioning” is an attribute of digital objects
Some observations:

There is more to metadata than meets the eye.

Marc21 is insufficient for digital collections

XML is in your future – prepare for it!
Metadata is very important
What is Metadata?

Simple Definition:
Information about information or “data about data”

Other Definitions:
“… the Internet-age term for structured data about data” - Joint NSF-EU Working Group on Metadata (1998)

“… structured data about data that imposes order on a disordered information universe” - Carl Lagoze (Cornell University)

“… machine understandable information about web resources or other things” - Tim Berners-Lee (World Wide Web Consortium)
Purpose of Metadata

• Helps add/extract meaning from an information object
  – *an information object is anything that can be addressed and manipulated by a human or a system as a discrete entity. The object may be comprised of a single item, or it may be an aggregate of many items.*

• Helps organize information
• Helps describe information
• Helps discover information
• Helps preserve information

“*imposing a veneer of regularity on the natural disorder of the artifacts we encounter*”
Types of Metadata

1. **Descriptive**
   - Characterizes the content itself

2. **Technical**
   - Records technical aspects and changes

3. **Administrative**
   - Helps track changes over time

4. **Rights Management**
   - Resolving rights of content

5. **Preservation**
   - Archiving of digital content

“All of these can be multiple views of the same information object”
1. Descriptive Metadata

- **Describes content**
  - Aids indexing
  - Aids discovery
  - Aids identification

- **Describes structure**
  - Aids display and navigation
  - Aids understanding of organization or structural divisions.
    - Example: chapters in a book
Descriptive Metadata Examples

- **MARC (MAchine-Readable Cataloging)**
  - Used to *describe* bibliographic information (books)
  - Foundation of library catalogs

- **EAD (Encoded Archival Description)**
  - Used to *describe* finding aids such as inventories, registers, indexes and other documents created by archives, museums, libraries and manuscript repositories

- **TEI (Text Encoding Initiative)**
  - Used to *describe* literary and linguistic texts created by libraries, museums, publishers and individual scholars

- **DC (Dublin Core)**
  - Used to *describe* a “broad range” of information objects

- **VRA Core (Visual Resources Association)**
  - Used to *describe* works of visual culture as well as the digital images that represent them
2. Technical Metadata

- **Describes Technical characteristics**
  - File size and format
  - Resolution and Colorspace
  - Character set
  - Software used to create
  - Software required for viewing
  - Example: MIX (Technical Metadata for Digital Still Images) - [http://www.loc.gov/standards/mix/](http://www.loc.gov/standards/mix/)
  - Example: MPEG-7 (Combines descriptive and technical) - [http://www.dlib.org/dlib/september99/hunter/09hunter.html](http://www.dlib.org/dlib/september99/hunter/09hunter.html)
3. Administrative Metadata

• **Describes Context**
  • Who, what, where, why, when information object was created
  • Who, what, where, why, when information object was modified
  • Sometimes includes “rights” information
  • Example: A-Core (Admin Core - Administrative Container Metadata) -
    [http://dublincore.org/groups/admin/](http://dublincore.org/groups/admin/)
4. Rights Management Metadata

• Indicates ownership
  • Supports copyright law
  • Supports use licenses (subscriptions, etc.)
  • Example: RoMEO (Rights Metadata for Open Archiving) -
    http://www.lboro.ac.uk/departments/ls/disresearch/romeo/index.html
  • Example: XrML (eXtensible rights Markup Language) – http://www.xrml.org
5. Preservation Metadata

- Preserves information integrity
  - Quickly changing technology environment
  - Ensures “bit stream” does not become obsolete
  - Example: OAIS (Open Archival Information System) - http://www.rlg.org/longterm/oais.html
Metadata Issues

• Standardization
• Cross-domain, cross-disciplinary
• Cross walking
• Representations
  – Conceptual representation
  – Organizational representation
  – Temporal representation
  – Spatial representation
Metadata Issue: Cross walking

- Mapping of one metadata format to another
- Metadata is often created with a particular community in mind but must be shared across communities
- **Cross walking is essential to the creation of consistent finding aids**
- Example: MARC21 → DC → EAD
  - 245 $a Hello World
  - *GOES TO* <dc.title>Hello World</dc.title>
  - *GOES TO* <archdesc><did><unittitle>Hello World</unittitle></did></archdesc>
Metadata Issue: Cross walking
Cautions in cross walking

Cross walking

• Requires in-depth knowledge and expertise
• Possible loss of semantics (or meaning)
  • Example: Creator in one metadata standard could be <author> and in another <artist>
• Element to element mapping losses
  • Example: One to many, Many to One
  • Example: Mandatory, Optional, etc. (Cardinality)
• … And Much More. Please refer to:
  • http://www.niso.org/press/whitepapers/crswalk.html
Metadata Issue: Organizational Representation

- Describes the “organizational” structure of an information resource

- Example:
  - EAD Metadata Standard
    - Collection
    - Box
    - Folder
    - Item

Metadata Issue: Temporal Representation

• Describes the manner in which the properties of an information object are transformed over time (from creation through continuing evolution)

• Traditional “descriptive” models focus on a stable or static information object; Digital objects evolve and change over time.

• Important in museums where time is crucial (e.g., it’s discovery, classification, exhibit history)

• Example:

  • ABC Ontology and Model
Digital Objects Evolve Over Time

Internet-Age Example

Digital Collections
Metadata for digital content has additional requirements

- Digital objects evolve, therefore need version control
- Single digital object can have many metadata data-streams
  - In the same metadata format
  - In different metadata formats
- Single metadata data-stream can point to many digital objects
- Preservation information for digital objects is more extensive

**Conclusion** - At present no single metadata format is sufficient for all objects; therefore system needs flexibility and versatility to handle all formats and even ones not yet invented.
How are XML and Metadata related?
What is XML?

eXtensible Markup Language

- metalanguage – a language for describing other languages
- Extensible? – Not a fixed format type. It can be used to design any number of document types (e.g. Metadata Document types)
- Markup? – The tags or field labels that identify an element
  - Example: <dc.title>Hello World</dc.title>
XML Components

• DTD (Document Type Definition)
  – Used to define the “allowable” structure of a particular XML document (tags, content of tags, etc.)
  – Allows for the creation of “Application Profiles”

• XSD (XML Schema Definition)
  – Successor of DTDs
  – Written in XML
  – Allows richer XML definitions and is itself extensible

• XSL (eXtensible Stylesheet Language)
  – Used for defining presentation of XML document
Example of XML Syntax (for EAD)

```
<!DOCTYPE ead SYSTEM "ead.dtd">
<?xml-stylesheet type="text/xsl" href="sfm.xsl"?>
<ead relatedencoding="MARC21">
  <eadheader langencoding ="UTF-8" findaidstatus ="edited-full-draft"
    audience ="external">
  </eadheader>
  <archdesc level ="collection" type ="inventory">
    <did>
      <repository encodinganalog ="852">
        <corpname>Stephen F. Austin State University</corpname>
        <subarea>Stone Fort Museum</subarea>
      </repository>
      <physdesc encodinganalog ="300">
        <extent>93 cm x 51 cm</extent>
      </physdesc>
      <unittitle encodinganalog ="245">Chair, rocker</unittitle>
      <physloc encodinganalog ="852">On-site Storage</physloc>
    </did>
    <acqinfo encodinganalog ="541">
      <p>Donation/Gift  <date>4/12/1985</date></p>
    </acqinfo>
    <scopecontent encodinganalog ="520">
      <p>SCOPE AND CONTENT NOTE</p>
      <p>"It has been in J.J. Pitt's family for approximately 102 years. It was previously owned by Nettie P. Pitt's mother, who gave it to Nettie after her marriage to J.J. Pitt in 1984."
      </p>
    </scopecontent>
  </archdesc>
</ead>
```
Need XML Editing Tools

Microsoft Office InfoPath™ 2003
The Microsoft Office information gathering and management program

xmlspy™

COREL™
XMetaL®
Making XML Content Creation Easy
Challenge for everyone

• *Developers* – Create software to manage XML and XML document workflows

• *Librarians* – Learn the emerging metadata standards to help customers describe their information objects correctly (a new age of cataloging? Heard of FRBR?)

• *Technical support* – Learn the tools that help create/manage XML data as well as XSL stylesheets that display XML content
Part 2: Digital Object Repository

Some observations:

Digital object repositories need special handling; standard file systems are insufficient.

Knowledge and experience in this area is limited; requiring caution and co-operation.

Risks can be reduced by using open source software.
Shortcomings of existing products for managing digital libraries

- Narrow focus on specific media formats (e.g. image databases, document management)
- Fail to effectively address interrelationships among digital entities
- Fail to address interoperability
- Fail to provide facilities for managing programs and tools that deliver digital content.
- Not extensible; do not enable easy integration of new tools and services

Tim Sigmon (Director, Advanced Technology Group UVA)
What is a Digital Object Repository?

- Stores and maintains digital objects
- Provides external interface for Digital Objects
  - Creation
  - Modification
  - Access
- Enforces access policies
- Provides for content type disseminations
- Offers preservation facilities
What is Fedora™?

Flexible Extensible Digital Object Repository Architecture
History of Fedora™

• 1997-Present
  – DARPA and NSF-funded research project at Cornell
    (Conceptual framework developed by Sandra Payette and Carl Lagoze)
  – Reference implementation developed at Cornell

• 1999-2001
  – University of Virginia digital library prototype (Thornton Staples and Ross Wayland)

• 2002-Present
  – Andrew W. Mellon Foundation granted Virginia and Cornell $1 million to develop a production-quality Fedora system
  – Fedora 1.0 released in May 2003 as Open Source under the Mozilla public license.
Fedora™ Digital Object Architecture

Globally unique persistent id

Public view: access methods for obtaining “disseminations” of digital object content

Internal view: metadata necessary to manage the object

Protected view: content that makes up the “basis” of the object

The Mellon Fedora Project (Page 4)
Digital Object with multiple datastreams

Datastreams

Digital Object

DC

EAD

Admin Metadata

Datastreams

EAD
Example Disseminators

- Persistent ID (PID)
  - Disseminators
    - Default
    - Simple Image
  - System Metadata
  - Datastreams

Methods:
- Get Profile
- List Items
- Get Item
- List Methods
- Get DC Record
- Get Thumbnail
- Get Medium
- Get High
- Get VeryHigh
Fedora™ Repository

- **Client Application**
- **Web Browser**
- **Batch Program**
- **Server Application**

**Fedora™ Repository**

**Management Subsystem**
- Object Mgmt
- Component Mgmt
- Object Validation
- PID Generation

**Security Subsystem**
- Policy Mgmt
- Policy Enforcement
- Users/Groups
- Policies

**Access Subsystem**
- Object Reflection
- Object Dissemination

**Storage Subsystem**
- Datastreams
- External Content Retriever
- Content
- XML Files
- Relational DB

**External Content Source**

**Web Service Exposure Layer**

**SOAP**

**HTTP**

**FTP**

**Component Mgmt**

**Object Validation**

**PID Generation**

**Policy Enforcement**

**Users/Groups**

**HTTP**

**SOAP**

**HTTP**

**Web Service**

**Remote Service**

**Local Service**
Fedora Advantage

• Extensible digital object model
• Repository exposed by Web services APIs
  – Management (Creation, Deletion, Maintenance, Validation)
  – Access (Search, Disseminations)
• Scalable, persistent storage for content and metadata
• Content can be local and/or remote
• Content versioning
• Open source solution
Fedora™ Statistics

- Total downloads (since May 2003): 1427
- Average downloads per day: 9
- Number of Countries: 32
- Types of organizations:
  - Universities: Libraries, IT, Departments
  - Software and Technology Companies
  - Defense/Military
  - Banks
  - National libraries and archives
  - Publishers
  - Research Laboratories
  - Scholarly societies
More Info?

http://www.fedora.info
Some observations:

It is easy to digitize and manage a few images; scalable solutions are more difficult to create.

Quality has to planned for before the project starts; it cannot be introduced afterwards.

Productivity is essential for ultimate success.
VTLS Imaging Technology for Advanced Learning
VITAL - Introduction

• Digital Asset Management System - based on the Fedora – **Open-Source** Digital Object Repository Architecture

• Software for creating, storing, managing, cataloging, indexing, searching & retrieving your digital collections

• Backed by VTLS software and service solutions designed to meet your needs
VITAL / Fedora Relationship

Application Developers

Public Access

Staff

Access Control

Web Services

Web Interface

Vital Manager

Open Source Fedora™ Repository

Other Digital Collections

OAI

Dublin Core/EAD/TEI/XML

Digital Content
How does VITAL work with Fedora™?

- **VITAL has workflow tools** that simplify the creating, storing, managing, cataloging, indexing, searching & retrieving of digital objects
- **VITAL uses Web Service Interfaces (API’s)**
- **Management Service (API-M)**
  - Ingest – XML-encoded object submission
  - Create – interactive object creation via API request
  - Maintain – interactive object modification via API requests
  - Validate – application of integrity rules to objects
  - Identify – generate unique object identifiers
  - Secure – authentication and access control
  - Preserve – automatic content versioning and audit trail
  - Export – XML-encoded object formats
How does VITAL work with Fedora™?

- **Access Services (API-A)**
  - Search – search repository for objects
  - Object Reflection – what disseminations can the object provide?
  - Object Dissemination – request a view of the object’s content
Reasons to base VITAL on Fedora™

[1 of 2]

1. The “Flexible” and “Extensible” aspects.
   **Flexible** – The development can be in any programming language because the API is based on Web services.
   **Extensible** – We can write workflow tools that facilitate different types of workflows.

2. XML Submission and Storage – Digital objects are stored as XML-encoded files that conform to an extension of the METS schema.
3. Focuses on Object Repository model and not how the repository will be used.
4. Native OAI-PMH support.
5. Open Source – VTLS is spearheading the open source movement in the library industry.
6. There is a community of users developing other add-ons to Fedora that will add value and services for all Fedora adopters.
7. Low cost procurement for customers
It’s VITAL to manage the resources of today’s digital libraries.
Four Components of VITAL

1. Fedora™ Repository
   UNIX/LINUX, SUN and Windows

2. VITAL Manager
   Based on Windows 2000 and XP
   Has XML Cataloging Utility
   Uses EAD, DC & MARC XML Templates
   Has a digital object loader

3. VITAL Web Portal
   UNIX/LINUX and Windows

4. Oracle Database (Optional)
2. VITAL Manager

2.1 Advanced Collection Management Functions

2.2 XML/METS Metadata Storage, Linking, Retrieval & Export

2.3 XML (Dublin Core) Editing & Indexing

2.4 Uses Fedora™ Digital Object Search Tool

2.5 Easy Image Management & Import with some Automatic Metadata Creation and Linking
2.1 VITAL Collection Management

- Supports multiple Fedora™ repositories
- Collections can be dispersed across locations
- Repositories can contain diverse digital object types
- VITAL facilitates easy loading and searching of repositories
2.2 VITAL XML Metadata Storage

- Standards based
  - XML/METS Schema
  - Dublin Core
  - EAD
  - MARC
  - AMICO XML Format
  - Additional formats can be added quickly

- Metadata may be exported in XML for use in other applications
2.3 VITAL XML Editing

- Cataloging/editing with XMLSpy Software
  - Templates provided for Dublin Core, EAD, and MARC
- Additional 3rd Party XML tools may be used
  - XMetal from Corel
  - Microsoft InfoPath
- Collection metadata may be imported from VTLS MetaCat
2.5 VITAL Image Management

- Easy import of digital objects and images
  - Watched Folder
  - VITAL Import Tool
- Digital object versioning
  - Changes made to the digital objects are recorded in the repository
- VITAL automatically creates technical metadata for the digital object by recognizing the imported files mime type
2.5 VITAL Image Management

- Integrates with any TWAIN source scanning software or imaging application
- Images can be immediately verified prior to load - through the VITAL Manager preview window
- Tools to facilitate the digitization of all materials including, rare objects and historical documents
VITAL Manager Client Details
VITAL Manager Client Details

- Search the repository to locate digital objects and their associated image, text and metadata
- Launch the software of a TWAIN compliant scanner or digital camera directly from VITAL and load the digitized images in one step
- Import one or many image, text, sound and other digital files into the repository and have the basic metadata created dynamically based on mime type
- Configure a “watched” folder from your favorite application to automatically move files into the repository
The VITAL Manager Client allows for easy navigation and searching of your digital object repository.
A search of the repository produces the digital object reference and its associated datastreams - reflected here with a local digital image file and the Dublin Core metadata describing this object.
VITAL Manager Client Details

Datastreams can be edited using linked applications – Metadata datastreams such as Dublin Core are modified by integrated XML editors such as XMLSpy or XMetal.
VITAL facilitates adding additional scanned images to the repository by providing an easy to use interface compliant with any TWAIN scanning or digital camera device.
VITAL features an Import/Ingest tool for loading digital images, text, metadata, etc., from your local or networked file system into the repository — individual or multiple files may be added to the repository using this workflow.
3. VITAL Web Portal
3. VITAL Web Portal

- Z39.50 compliant – compatible with any integrated library system

- Sophisticated display for Encoded Archival Description (EAD), Dublin Core and MARC

- Includes the VTLS Hi-Res Image Navigator – uses Wavelet compression for incredibly detailed viewing of your images
  - Supports MrSID and JPEG2000 encoded image files

- Instant access to digital content anytime, anywhere, to anyone with a web browser
3. VITAL Web Portal

The VITAL Access Portal is a Z39.50 compliant, web-based software interface for searching and retrieving digital objects from the repository.

Copyright © 2004 VTLS Inc. All rights reserved.

For specific details regarding the purchase or evaluation of the VTLS VITAL solution, contact askvtls@vtls.com.
3. VITAL Web Portal

The VITAL Access Portal has a completely configurable interface – institutions can create their own look and feel for the front-end and provide a variety of search options including pre-defined searches to assist their users in locating groups of digital objects in the repository.
3. VITAL Web Portal

The results screen presents a list of digital objects that satisfy the search term(s) – clicking on the hyperlinks to the left will bring up the digital object summary screen.
3. VITAL Web Portal

Text documents in Word, PDF and DjVu may be launched into the browser by clicking on the "Content" datastream icon.

Dublin Core and Digital File (DjVu) Datastreams
Tides in Early Texas History –
Stephen F. Austin University
3. VITAL Web Portal

Institutions have considerable flexibility in the way they present their collections – the examples here show two different approaches to presenting EAD (Encoded Archival Description) metadata objects.

Clicking on the thumbnail image from this screen will launch the VITAL Hi-Res Image Navigator – a tool which provides for detailed examination of these wavelet compressed image files.
3. VITAL Web Portal

MrSID and JPEG2000 wavelet compressed images can be stored in the repository and displayed to the user via the integrated VITAL Hi-Res Image Navigator.
The AMICO Library™

The AMICO Library™ in partnership with VTLs Inc.

Scan the Collection
Enter your search terms:
Select a search type:
Artist
Search

Search by Keyword

Featured Item

AMICO Library™ is a licensed digital educational resource available under subscription to universities, public libraries, elementary and secondary schools, and museums.

Types of Works:
- over 10,000 paintings
- over 5,000 sculptures
- over 12,000 drawings and watercolors
- over 14,000 prints
- over 25,000 photographs
- over 1,400 textiles
- over 1,400 costumes and jewelry
- over 6,500 works of decorative art
- over 600 books and manuscripts
3. VITAL Web Portal

The AMICO Library in VITAL
Implementation Options

• The Fedora™ package
  – Fedora™ open source software (free)
  – VTLS installation, training, and support
Implementation Options

• The Full VITAL package
  – Fedora™ open source software (free)
  – VTLS software and hardware extensions, with features and workflows
  – VTLS installation, training, support, integration and documentation
Implementation Options

• VITAL Hosted Solution
  – VTLS provides ASP services for your digital collections

• VTLS Professional Digital Imaging Services
  – Imaging services and project consulting can be combined with any of the above packages to provide a solution tailored to your needs
Just the Beginning

- **Fedora™ 2.0 planned enhancements**
  - Digital Rights Management
  - Authentication (Shibboleth)
  - [http://www.fedora.info](http://www.fedora.info)

- **VITAL Directions**
  - Advanced collection management features, indexing, searching and statistical reporting
  - VTLS Knowledge Portal for Fedora™ resources and collaboration
  - Additional features for multimedia collections
  - Integration of VTLS V-Commerce Solutions
Part 4: Working with a vendor

As a general rule (there are exceptions) in-house projects cost more, take longer, and often remain incomplete.
Reasons to work with a vendor

• **Space:** No need to convert for scanning activities
• **Equipment:** It is the vendor’s responsibility to stay ahead of the technology curve
• **Staff:** Imaging activities required specialized training and personnel
• **Downtime:** Not your concern, the onus is on us
• **Economies of Scale:** Larger projects can be done faster, more efficient and at less cost to you
• **Security:** Protection of a contracted price
• **Experience:** Technical & Creative
Practical considerations in working with a vendor

- Communicate your project goals to the vendor
- Consider a small pilot project to work out the kinks and learn the process
- Identify points of contact in each organization and work through issues as they come up
- Spend the time up front making sure that the promised deliverables meet your expectations
- Do your part to keep the project moving and expect the same of the vendor
- Even with a vendor there are some things you must do
Working with a Vendor
What You Must Do [1]

This outline is an abbreviated version of Library of Congress – steps in a digitization project.

• Select a collection
  – Analyze Collection
  – Determine scope of digitization (entire or subset?)
  – Assess the physical condition
  – Assess restrictions and copyright
  – Determine finding aid

• Plan digitization strategy
  – Develop method for collection preparation
  – Develop preservation treatment plan
  – Determine formats (capture, archiving and presentation)
  – Determine physical size
  – Determine scheme for file name assignment
Working with a Vendor
What You Must Do [2]

- Estimate resource requirements
  - Disk space
  - Number of scanners
  - Number of people and working days

- Develop restriction plan [copyright plan]
  - Review copyright restrictions
  - Seek required permissions
  - Add notices to all restricted items
Working with a Vendor
What You Must Do [3]

• Prepare Documents and Scan
  – Prepare targets
  – Prepare scanning instructions for collection
  – Scan collection
  – Process scanned images
  – Review images for quality
  – Coordinate rework
• Create Database
  – Archive images in repository
  – Develop Finding aid
    • Modify existing finding aid
    • Create new finding aid
Working with a Vendor
What You Must Do [4]

- Create Links
  - Use productivity tools
- Test and refine
  - Review for accuracy and completeness
  - Test links
  - Make any necessary changes
- Release Collection
Outsourcing Responsibilities of the Site

- Selecting the materials to be digitized
- Determining the purpose of digitization and the nature of the desired final output
- Establishing the quality targets
- Verifying the quality of completed work
- Enjoying the extra attention you and your institution will receive from the fantastic collections you make available in digital format
Conclusion

Go as far as you can see for
You will be able to see farther when you get there!

_action is better than inaction and reaction_