Creating Virtual CD-ROM Collections

Kam Woods
Geoffrey Brown

Indiana University
Department of Computer Science
2008

"Over the past 20 years, more than 100,000 CD-ROMs containing vital social, economic, cultural, and scientific data have been produced. The methods used for the creation, distribution, and storage of these materials have introduced technical challenges related to preservation, cataloging, and access that are of immediate concern to any holding institution.

Using readily available open-source software, we present techniques to improve archival practices for CD-ROM collections, including bit-level image preservation, web-based access using format migration and remote execution in emulated environments, and improved metadata handling. Our techniques generalize to other collections and can be integrated with existing archival software."

The Project

- CD-ROM Access
  - Format identification
  - Web-based browsing of CD-ROM collections
  - Format migration
    - Emulation (accessing legacy executables in virtual environments)
- CD-ROM Preservation
  - Creating and validating “bit-faithful” copies
- Distributed Collections
  - Building reference images
  - Authentication

Indiana University Department of ComputerScience

Indiana University School of informatics

IU Home | IUB Home | Informatics home

Home
Search
List All

Chemical Emergency Preparedness and Prevention Office library on CD : a one-stop reference

Contains information on chemical accident prevention and emergency preparedness and response.

United States Environmental Protection Agency. Office of Solid Waste and Emergency Response. ; United States. Environmental Protection Agency. Chemical Emergency Preparedness and Prevention Office.
Washington, DC : U.S. Environmental Protection Agency, Office of Solid Waste and Emergency Response2000

System Details : 76.8 MB hard drive space; Adobe Acrobat Reader + Search version 3.0 (included on this CD-ROM); Microsoft Windows 95 or Windows 98; Adobe Acrobat for Macintosh.

Notes : "Issued April 2000."
"EPA 550-C00-001"--disc label.

Subject Headings : United States. Environmental Protection Agency. Chemical Emergency Preparedness and Prevention Office.
Chemicals -- Safety measures
Hazardous substances -- Safety measures

Record Info : (InU)CBB9138BB
(OCoLC)ocm44648346

Indiana University Catalog Information

EP 1.104: C 42/2

1. Download : Browse

Issues

- Preservation and access issues unique to CD-ROM collections
  - Distribution across libraries
  - Require physical access
  - Examples:
    - FDLP GPO collection: ~5000 images
    - Indiana University collections: 14,000+
    - OCLC WorldCat: 120,000+
  - Collections not easily searched
    - Incomplete metadata
    - Obsolete file formats
  - Subject to bit-rot
  - Large objects (up to 8GB)
  - Need to be mounted for access
  - Typical user requires only partial access

Our Strategy

- Semi-automatic construction of a virtual collection accessible from internet-enabled locations
- Format migration to web-friendly formats
- Use of VM technologies to augment or replace existing workstations
- Collective maintenance of images and metadata via modern network filesystem (AFS)
- Differentiate between public and restricted materials via ACLs, Kerberos domain

Image Access

- ISO images are good preservation targets, but can be problematic for access
  - Files stored in obsolete formats
  - Executables and links may depend on explicit mount points
  - Macintosh format images not fully supported under Linux
- Extensions
  - ISO standard has been modified via extensions (Joliet, Rock Ridge) to overcome metadata and naming issues associated with the original spec
  - Correct rendering of original file names becomes difficult
- Security
  - Mount events triggered by actions on ISOs made available over the Web is a security risk

Building The “Raw” Archive

- Relatively inexpensive, provides basic browsing capabilities
- Many technical and preservation issues
  - Limited search capability
  - Copyright/access control issues
  - Limited metadata
  - No way to validate bits
  - No universal ID for items in collection
  - IU collection incomplete

Image Creation

- CD-ROMs governed by well-defined (ISO 9660) standard
  - ...but not in the real world: conformance to standard varies significantly in existing archives
    - Additional problems with contextual dependencies embedded in data
- Bit-level preservation
  - Differentiating between meaningful and irrelevant extraction errors
  - Volume header issues

ISO Image Handling

- Multiple issues
  - Advertised volume size errors (typical - TAO)
  - Truncation during image creation
  - Verifying image identity through checksumming

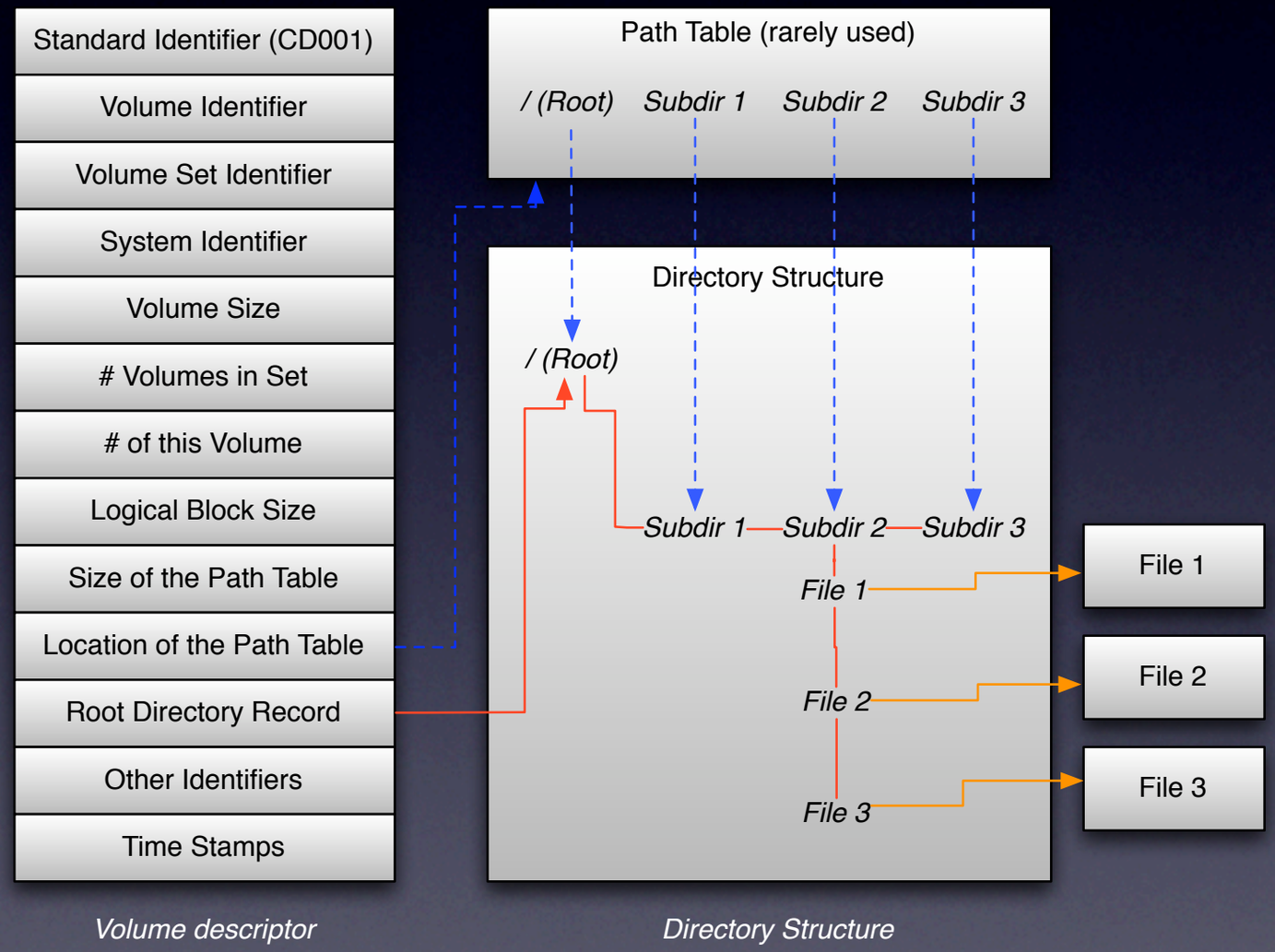
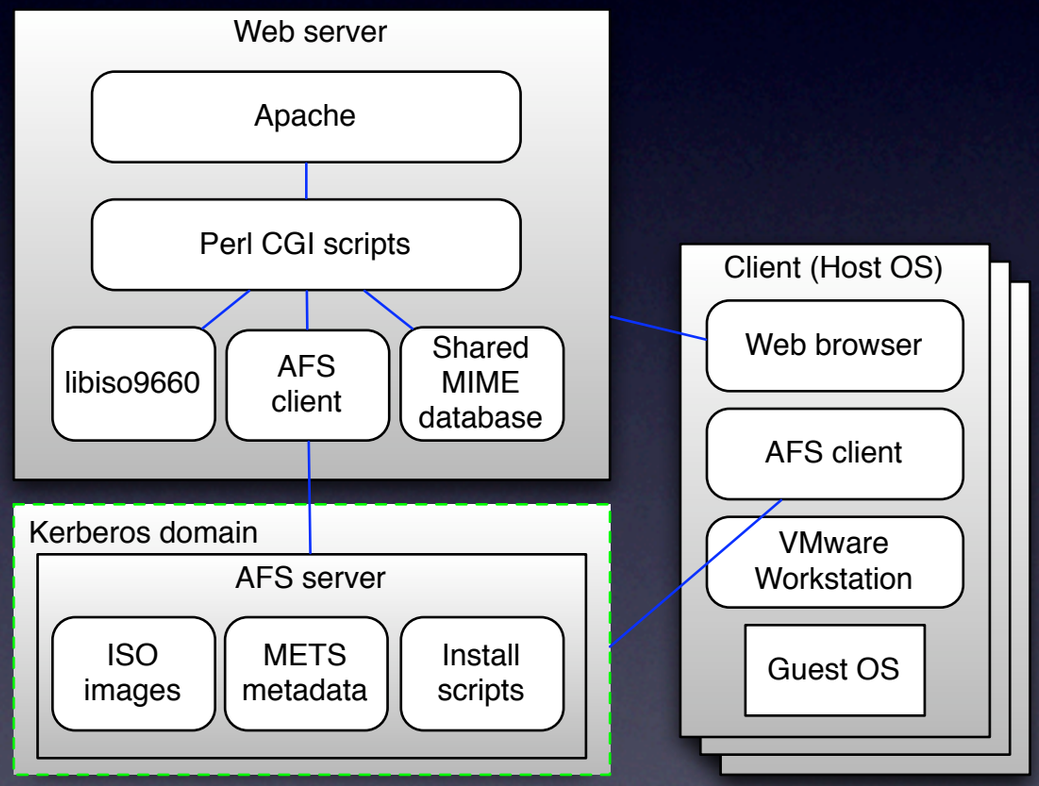


Image Distribution

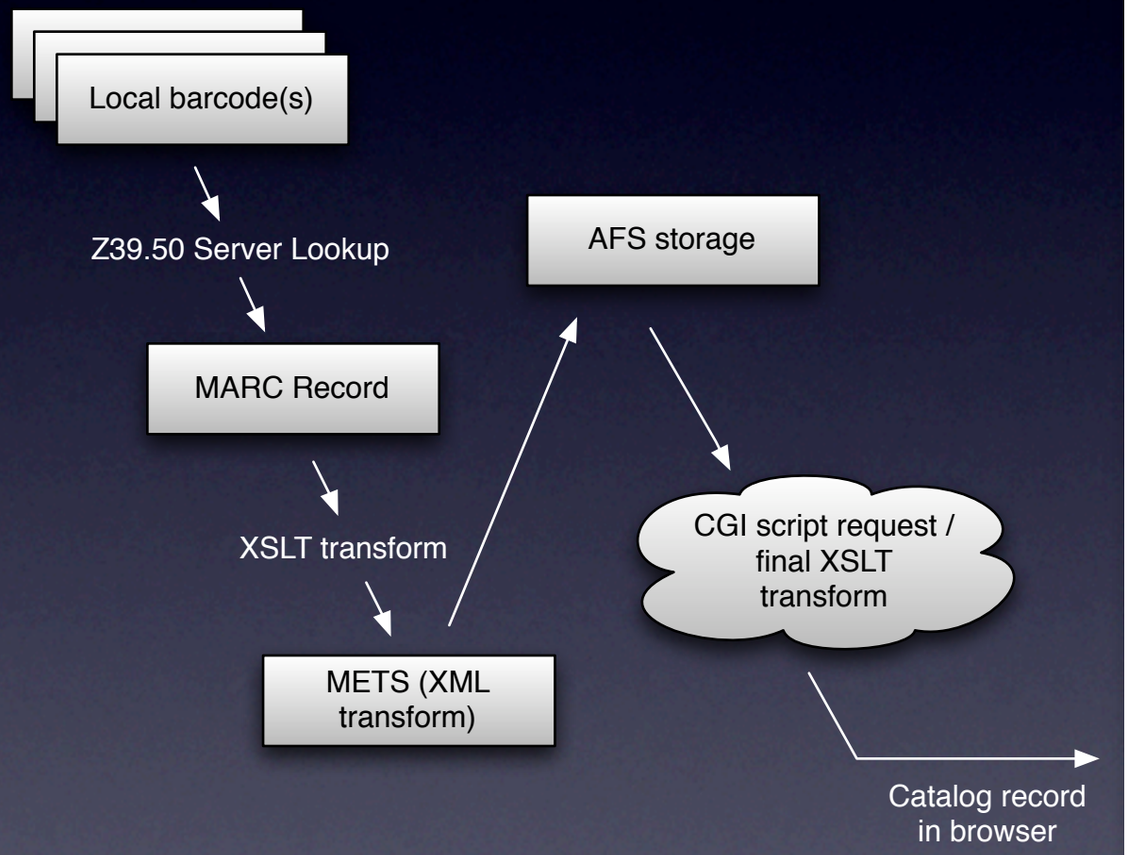
- Downloading complete ISO images is inefficient for most access scenarios
  - AFS provides access via web, remote mounts, retains download facility
- Why AFS?
  - Global namespace for distributed participation
  - Transparent storage migration
  - Storage mirroring
  - Multidomain authentication + ACLs

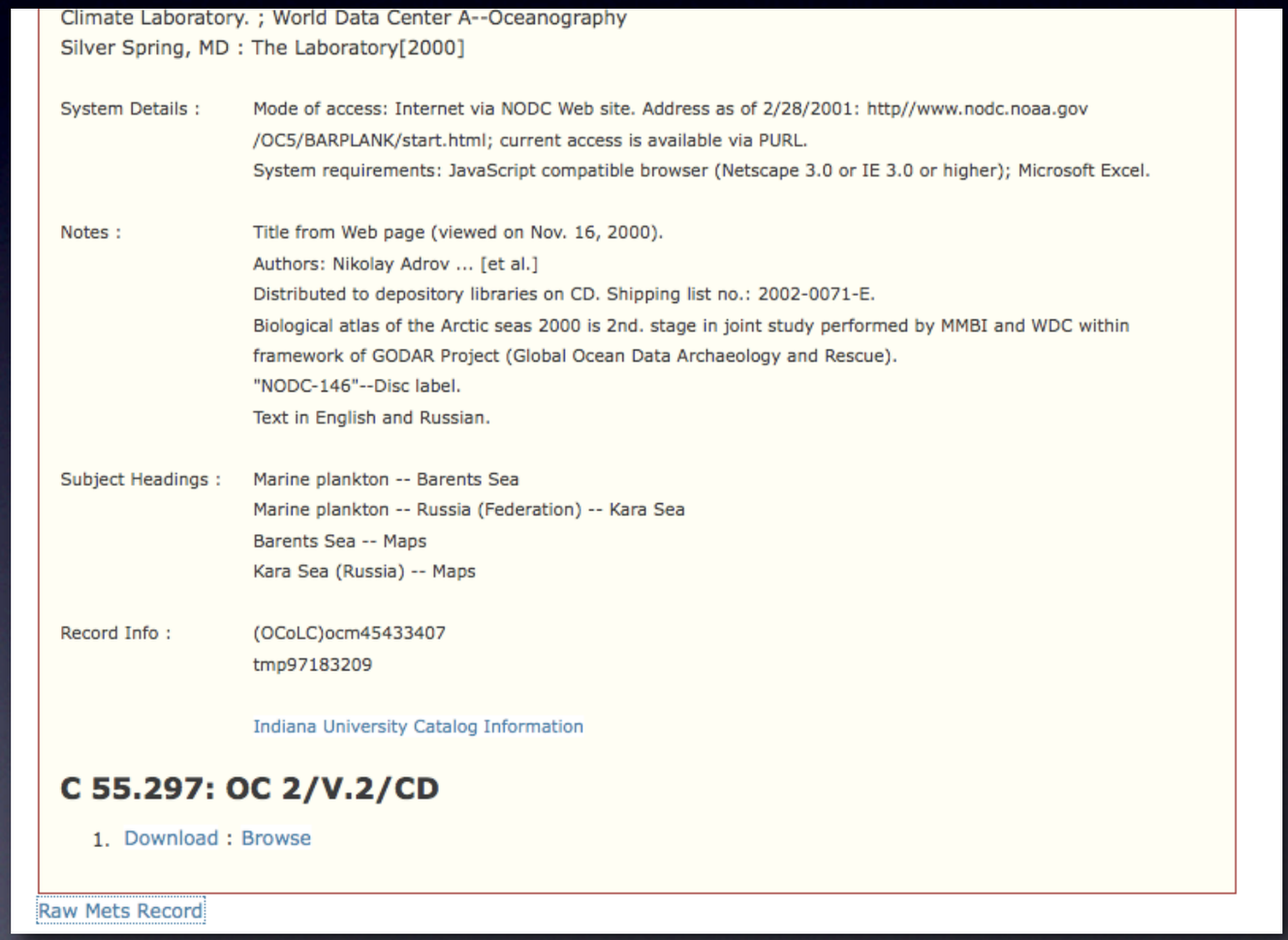
Service Overview

- Thin CGI script implemented in Perl (~600 lines)
- Additional Perl scripts transform MARC to METS, generate Swish-E tags for records
- File browsing information extracted directly from ISO images using libiso9660



Metadata Handling





Climate Laboratory. ; World Data Center A--Oceanography
Silver Spring, MD : The Laboratory[2000]

System Details : Mode of access: Internet via NODC Web site. Address as of 2/28/2001: http//www.nodc.noaa.gov /OC5/BARPLANK/start.html; current access is available via PURL.
System requirements: JavaScript compatible browser (Netscape 3.0 or IE 3.0 or higher); Microsoft Excel.

Notes : Title from Web page (viewed on Nov. 16, 2000).
Authors: Nikolay Adrov ... [et al.]
Distributed to depository libraries on CD. Shipping list no.: 2002-0071-E.
Biological atlas of the Arctic seas 2000 is 2nd. stage in joint study performed by MMBI and WDC within framework of GODAR Project (Global Ocean Data Archaeology and Rescue).
"NODC-146"--Disc label.
Text in English and Russian.

Subject Headings : Marine plankton -- Barents Sea
Marine plankton -- Russia (Federation) -- Kara Sea
Barents Sea -- Maps
Kara Sea (Russia) -- Maps

Record Info : (OCoLC)ocm45433407
tmp97183209

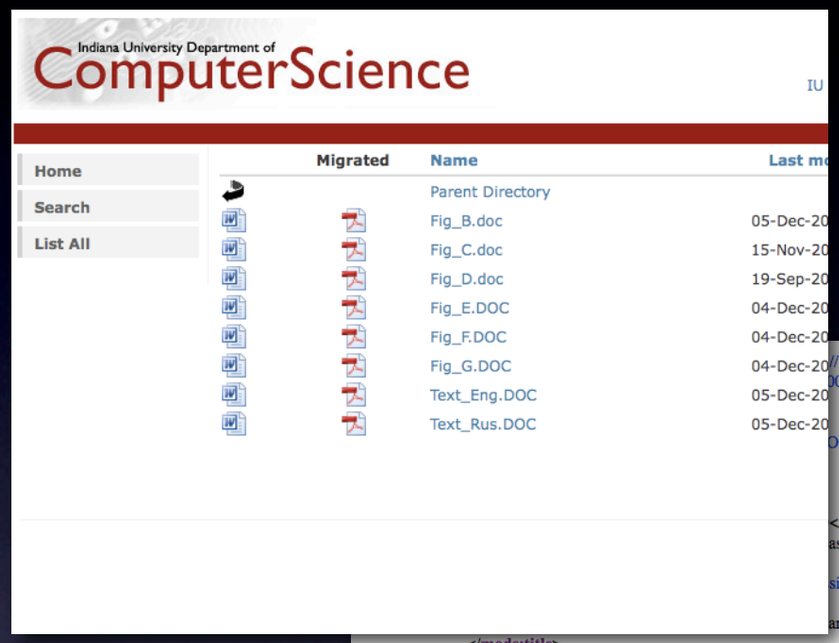
Indiana University Catalog Information

C 55.297: OC 2/V.2/CD

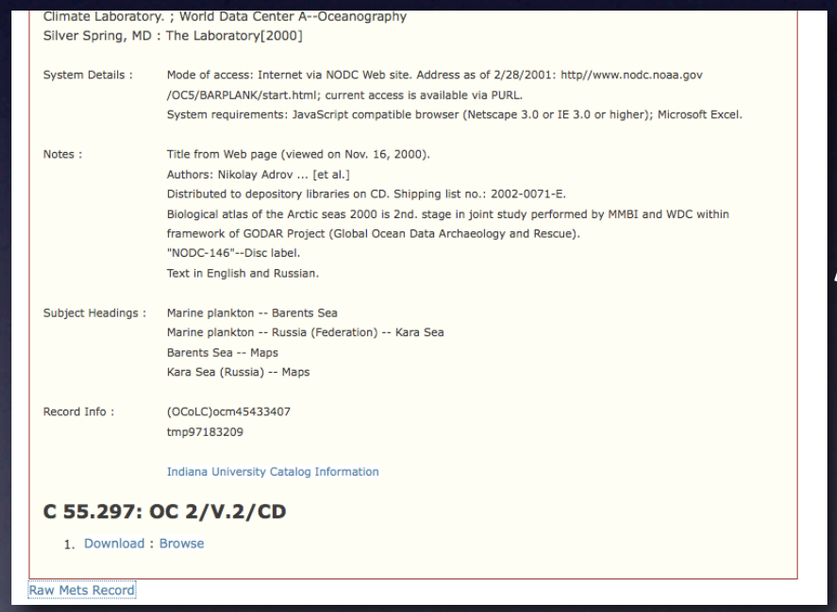
1. Download : Browse

Raw Mets Record

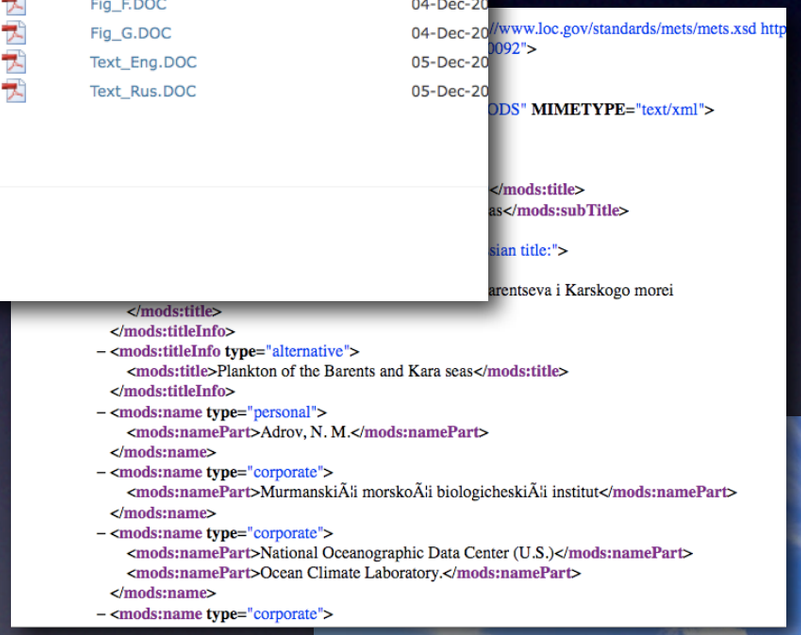
Browsing



File browsing



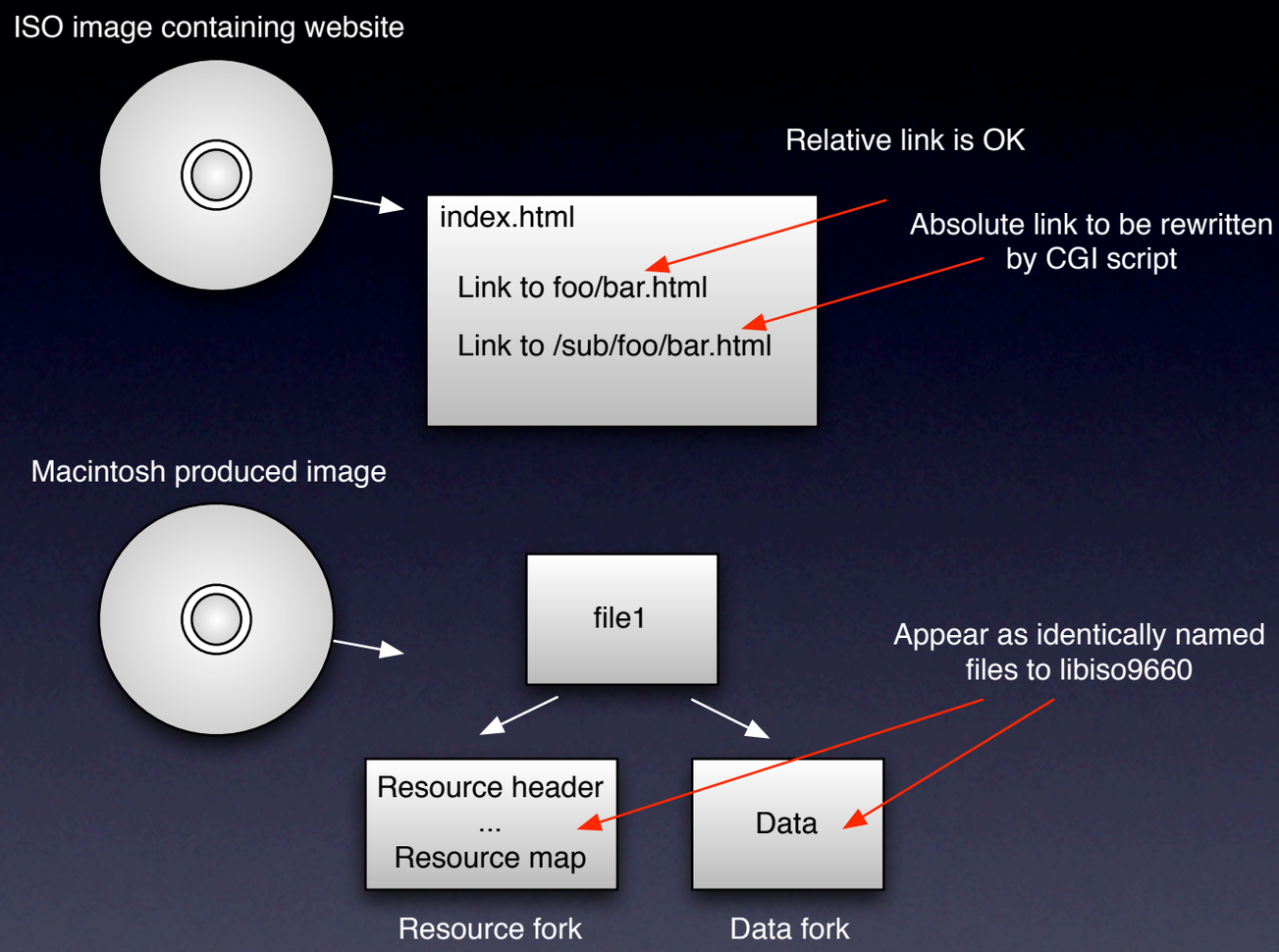
Catalog record



Raw METS

Virtualization

Browsing



- Archived websites frequently require on-the-fly rewriting of absolute links for browsing access
- Resource forks from Macintosh file systems can confuse libraries expecting uniquely named items in file hierarchy

Migration for Access

- Identifying good candidates for migration is difficult
- Focus on migration for access
- Open-source tools, scripted tests, heuristics
  - Shared MIME-info Database
    - Fast, integrated seamlessly with UNIX environment, easily modifiable database with strong community support
  - OpenOffice in “headless” mode and Gnumeric provide majority of translation filters required
  - Migration instances threaded, monitored for failure based on expected job duration

Emulation for Access

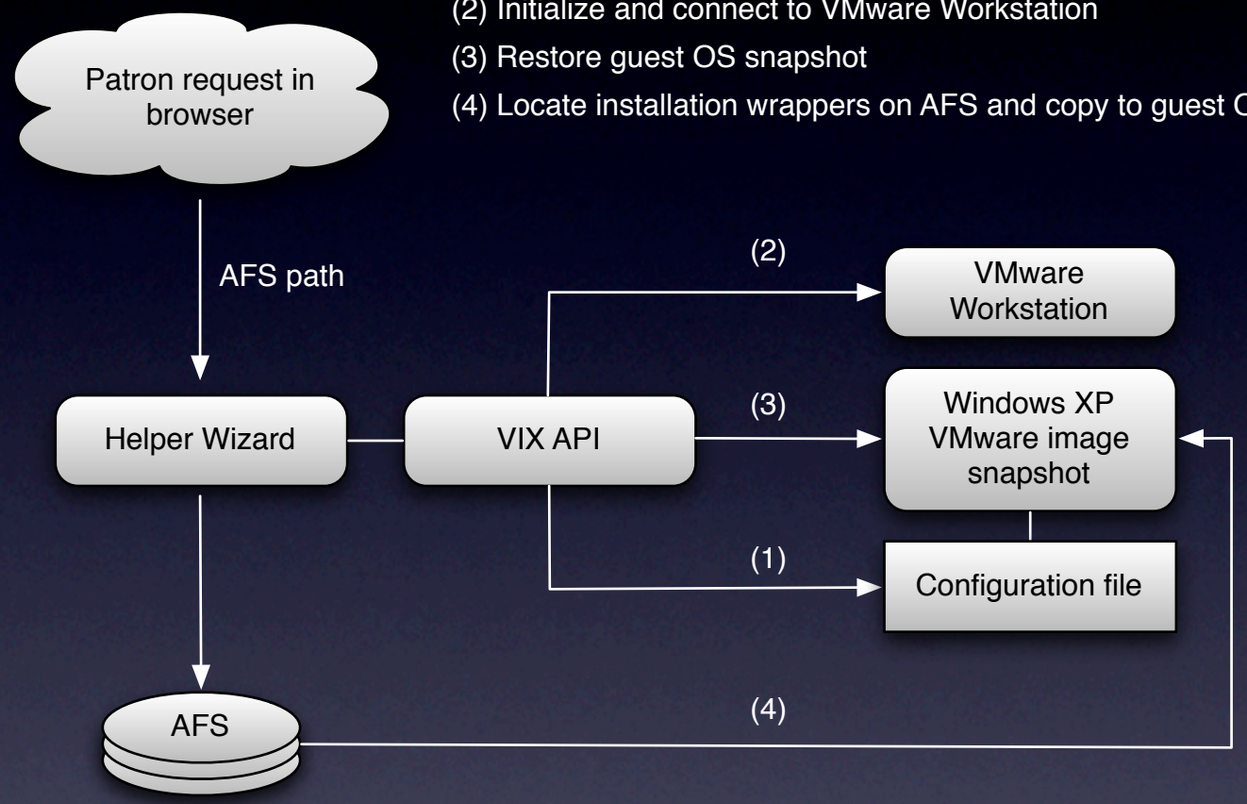
- GPO specifications for library reference workstations are inadequate
- Focus on current technology incrementally introduces additional access issues
- "Libraries should also consider keeping [existing] equipment in order to access electronic products that cannot be read with newer hardware and software." (http://www.fdlp.gov/computers/rsissues.html)

Emulation Strategy

- A pre-configured emulator is provided on local workstations
- Emulator is customized
  - ISO is mounted on a virtual hardware device
  - ISO specific installations (if available) copied and executed
  - Shared file directories created for patron use
- Links to ISO on AFS storage and web accessible file system provided alongside item metadata
- Emulator executes as preferred application for ISO images under patron control

Emulation Overview

(1) Rewrite configuration file with new AFS path
(2) Initialize and connect to VMware Workstation
(3) Restore guest OS snapshot
(4) Locate installation wrappers on AFS and copy to guest OS



- Guest OS is restored to snapshot after every session
- Remote mounts from AFS
  - Only transfer information from ISO as requested
  - Transparent to guest OS - continues to see ISO as physical mount

Preparing the Environment

- Analyze software requirements of individual ISOs
  - May depend on specific hardware conditions (presence of D: drive)
  - May require specific versions of legacy commercial software
  - May pollute environment during install process
- Build software images (OS, supporting applications and readers)
- Build and test customization scripts

Emulation Testbed

- VMware Workstation 6.04 for Windows
  - VMware VIX API used to automate configuration, startup, reversion to snapshots on exit
  - Windows XP SP3 image prepped with Office 2007
- 66 ISO images prepped with customized installation scripts
  - Installation scripts compiled to executables and stored in AFS alongside ISOs and metadata

Status

- Over 4000 CD-ROMs from IU collection available online
- Test implementation available
  - http://www.cs.indiana.edu/svp/
- ~1M files migrated and retained on AFS
- Public access to most materials

Future Work

- Formalize image creation and metadata transformation for use in SIPs/AIPs
- Shared pool of software images and licenses
- Further enable participating institutions to share expertise in supporting various document collections
- Collaboration to maintain copies of shared resources
- Improve access by limiting specialized software required on local workstations

Acknowledgments

- Lou Malcomb (IU Head GIMSS)
- Julianne Bobay (IU Head SLIS Library)
- Stewart Howard, Mitchell Lutz, Valkyrie Savage, Amanda Farag