

Digital Repository Projects at the North Carolina State University Libraries

James Jackson Sanborn and Jim Tuttle The International Association of Aquatic and Marine Science

Libraries and Information Centers (IAMSLIC) is a global non-profit

North Carolina does NOT accept paper theses

NCSU - Electronic Theses and Dissertations

NCSU - Authors Databases

Technical Reports Print Collection

- Campus institutes and departments

- Massie fall-off in print distribution

Special Collections Resource Center

- Digitized texts, photos

- Campus newsletters

GIS Data

- Library

- Homegrown

Target Research collections first

- Technical reports

- ETDs

- Faculty Pubs, citations

Treat each collection as a project

Actively pursue common tech, but willing to try various approaches

Technical Reports

DSpace Application

Lightly Customized

Libray Harvested

- Local catalog/metadata database

Scripting ingest object creation
Batch ingest

Mix of ongoing submission by dept personnel and library capture

Providing URIs to copy in repository

http://www.diglib.org/pubs/news05_02/ncsunews5.htm

ETDs into dSpace

Partnership with grad school

Hybrid with DSpace and ETD-db

See <http://etd.vt.edu/background/>

Faculty Publications

Originally access/cold fusion

In Oracle/PHP

Re-modeled data

Added functionality

- Open URL

- Vita like citation display

- Full text or submission links

Full text stored in dspace

- Citation metadata and file exported by script

- DSpace identifier currently manually entered

<http://www.lib.ncsu.edu/searchcollection/>

NCGDAP: Geo-spatial content in NC

<http://www.lib.ncsu.edu/ncgdap/>

Preservation is the key to the project, not access

Repository agnostic ingest and export

Simple curation functions

- Periodic MD5 checksum validation

- Structured metadata index

Expected archival data exchange

Leverage os

Python to wrap

- Antivirus

- Compressed files

- At risk fbr

ESRI ArcGIS toolbar (cool)

Metadata: Communities and collections

- Search by community type for 100+ communities

Repository agnostic approach

Spokes for each transformation

Facilitates export from Dspace into other repositories

Generate DSpace QDC, METS; populate workflow database (!)

Extra-repository AIP mgmt

Workflow Management Database (WMD) populated as a spoke on the metadata/ingest hub

External tracking NOID, Handle, ISO keywords

Integrates with GIS lookup tool

Upcoming Repository Related Projects

- XTF search interface

- Inter-archive exchange

Digital Collections Repository

- Special collections research center

- Other non-faculty collections

Digital Repository

Scientific Data
Statistical resources

james_sanborn@ncsu.edu

Mark Hedges: Workflow for the ingest
AHDS: Art and Humanities Data Services

Descriptive metadata in atomic model

PREMIS

How does PREMIS differ from CIDOC

Cornell: If we built it will they come?

Active: Theses, dissertations

Rhododendron

Zucker shrub

Senior seminars

Downloads: Scanned classic textbook for classes, video biography, physics video

Growth patterns: Great plains, terraced gardens, rocky mountains

one time, punctuated, steady

Comparison to other institutions: 7 academic institutions, at least 6 months,
harvested collection and pop info from public web interfaces

Wide diversity of items, communities, collections, empty collections

Growth: Majority are plateau or staircase. Seems to point to how the repositories are being used.

Faculty interviews: 11 from sciences, social sciences, humanities

M=6, F=5, 4 assistant, 3 associate

Access to literature a non-issue for faculty 10:1

Using alternatives to institutional repositories (own websites). Liked control and ease of use

Research community has higher salience than institutional rep. Move from school to school

Normative culture of disciplines

Faculty did not know about dSpace (7/11)

Permanence	Redundancy
Timeless	Learning curve
Registration for discoveries	Copyright
Policy of funding	Quality association
	Plagiarism
	Reputation
	Publishing issues

Access is not an issue

I am hoping this is being done

Conflicting cultures:

Collision course

Don't want to lead

Conclusions:

Largely underpopulated and not being used

Faculty behavior guided by inertia

Most faculty have no motivation to use dSpace

Is this a case of build it and they won't come?

<http://hdl.handle.net/1813/5195>

Philip Davis: cornell.edu

Dept of communication

NYU Digital Library Program: Joseph Pawletko

Motivation, architecture, approach, background

OAIS Reference Model

Implementation details

Motivation: Digital Library program working on several grants. HIDVL, Afghanistan, NDIIP

Separate or PR?

Monolithic or distributed?

Built single repository with distributed architecture.

Each component provides subset functionality

Project independent, reusable

Why components: Technology is changing rapidly. Flexibility, decomposable, improving software development return, decreases wasted effort

Producer: SIP to OAIS (archival information package), Discovery phase from consumer, sends to consumer results, then orders through DIP

Producer | Consumer

SIP - Ingest - Extracts metadata, produces AIP - data mgmt / archival storage - access - consumer. Descriptive info for discovery

Common services

Preservation planning and admin: human elements

DSpace: Leveraging existing capabilities + plug-ins, + scripts

Dspace Item Importer

Dspace API

Maintenance API

OAI/PMH

SRW

SRU

Ingest

Input Directory - CRON job looks for SIP, input monitor manages SIP, AIP generator preflights, AIP staging. QA phase, Ingest staging, Archive ingest, database entries, cross check

Large file ingest: PTV must ingest large files, 800GB

SRB features in dspace. Can ingest files of 130GB (<https://libnet.ucsd.edu/nara/>)

AIP: DMD, .foo, TMD foo, mets file (versioning), copy of file without version #
To update, a new METS file is built (points to new and existing)

Maintenance: Scope: enable authorized users to manage. Rollback. Leveraging existing components

Dspace Maintenance API: JAVA interface approach
Invoke functions through GUI

Common Services: Authentication, Authorization, Persistent Identifier Mgmt

Authentication: Centralized
Authorization: Centralized

Next talk

Plug-In Mgr:

Preservation Services Architecture: Rutgers

Preservation Monitoring

Event Messaging

Content model

Object Integrity: Fedora API, create and compare checksum

Event Messaging service:

An event is: action with an object, agent, rights entity (PREMIS)

Event outcome: a situation or state that follows an even and is a result of the event.

Generic framework: Messages represent

Atomic events: ingest, delete, modify, fixityCheck

Event messages - producers and consumers

Preservation Services

Fedora Commons

Core development team
Topaz project PLOS
Community-committers

501c3 charity
New proposal to Moore foundation
2010 sustainability timeline

Must build user base
Community leadership model
Develop income generating membership model
Evaluate

Evolve to steady state organization

Status: Phase 2
Complex objects
Networks of objects
Re-use
Process-orientation
Collaboration

Fedora as web application
Easy Installation

NSDL Presentation

TS Elliot - we have the experience but missed the meaning

Billions and billions of resources

Access was the thing - search

Why NSDL if there is Google

Metadata thing

- Not scalable, difficult, labor intensive

Is NSDL a library? CS folks. Battle between

'metadata is dead'

Metadata did not reflect the intellectual world

- The energy is not captured by this system
- Context is key

Fedora helps with collaboration, contextualization

Teaching Box

Concepts, Lessons, Activities, Resources, Notes
Personalized Collections

Teaching Box Model

Teaching Box API

- XML-RPC Servlet
- Integration of external DL services
-

Web 2.0

- UI Integration of search and browse
- Access to varied digital collections
- Streamlined delivery of multiple query results
- Web Service API achieved through
-
-
- huda.khan@colorado.edu
-
-

Repository Deposit Service Description

No easy way to deposit content

Encouraging deposit is the most difficult cultural issue

Technology needs to support

- Easy
- Multiple
- Auto
- Not closed or proprietary

1. Desktop authoring application
2. IMS compliant learning object to National Repository
3. Deposit in multiple repositories
4. Transfer between intermediate hosts
5. Repositories shared improved metadata (both ways)
6. Laboratory auto deposit

An analysis of Digital Repository Scenarios, Use Cases and Workflows

27 projects

- Metadata standards
- Metadata quality
- How do we achieve interoperability
-

Methodology

- Provided training
- Learn how to write UML and workflows, use cases
- Alistair Cockburn's work
 - Writing effective scenarios
 - Used trainers that had implemented a repository system using this approach
 - Intrallect
 - LADIE (JISC project)
 - Good and bad experience
-

Other Approaches

Extracting Internal Functions and external services

- CRUD is a given - Create, Retrieve, Update, Delete
- Versioning - filenames should include the date
 - Versions are valuable for extracting information
 - Enough metadata in versioning to describe
 - Version labeling should be easy
 - Permanence of a file location
- Open Access Versions closest to published version
- One click publish
- Functionality is too complex
- Open Access means?
- One .zip file download
- Copyright is key
-

Ecology

E-framework

- SOA approach

- Deposit or 'add' feature

Conclusions

-

SPARC: A Question of Access

Enabling the political interoperability

Expand dissemination

Reduce financial pressures on libraries

Leverage the networked digital environment

Immediate, free, online availability of research results

- Vision of scholarly communication in the networked digital environment
 - User toll barriers eliminated
 - Potential usage maximized
 - Value of research is fully realized
- Access model, not a business model

Less than 70% of the peer reviewed research is affordable at the wealthiest institutions

Barriers to sharing should be removed

Remix of data provides researchers new opportunities to do good things

Greater Access Benefits

- PLOS
- Publishing

Broad dissemination is essential and core to the mission of universities

Access is important to the tax payers

- 82% say the data should be accessible

Public access is a market issue

- Governments and taxpayers fund most academic research
- This is not sustainable

In order to remain competitive, research should be publicly available asap

A call for public access

- A distinction between public and open access

Common Elements in Public Access

- Copy of final manuscript
- Stable DR
- Free, public with embargo period (6-12 mos)

Dissemination is essential component of research

Expedite, expand and strengthen the national ability to leverage collective investments

Provide new avenues for use of federally funded research to stimulate new discoveries and innovations

Increase funders' ability to track results of research in which they have invested

Option of deposit in open repository or Open Access Journal

Explicit recognition of inclusion of deposit of data

Multiple repositories vs single central

- Portable PubMed Central
- Cross-Agency collaborations
- Public-private partnerships

Provisions to provide funding

US Policies

- NIH Public Access Policy
- Federal Research Public Access Act
- Senate Innovation Bill (S.2802)
 - Section 104

People FOR the policies

- Libraries
- Researchers

AGAINST

Publishers

The Eprints Application Profile: a FRBR

Provides application approach using FRBR entity model.

Dublin Core Abstract Model (DCAM)

- Set of relationships and attributes will allow for complex expressions
- Little mandatory, prescriptive statements minimized

OAI-PMH and community acceptance

- Dumb down
 - Still need simple DC
 - Simple DC about each Copy useful for getting to full-text, eg Google search crawl
- XML schema
 - Pete Johnston, EduServ
 - XML format (Eprints-DC-XML)
 - Create, expose and share DC xml

Community acceptance plan

- Deployment by developers
- Deployment by repositories, services
- Dissemination

JISC is funding other items such as images, spatial, etc...

Thoughts: CIDOC? Seems to be bandage around DC

Scholarly Communication Keynote 2

Tony Hey - Microsoft

e-Science is about global collaboration in key areas of science, and the next generation of infrastructure that will enable it - John Taylor

New Science Paradigm

- Experimental Science
- Theoretical Science
- Computational Science
- Now - e-Science or data-centric science

e-Science is shorthand for a set of technologies to support collaborative networked science

HPC and Information Management are key technologies to support this e-Science revolution

Example: Project NEPTUNE

- Programmable sensors and remote instruments
- Sensor networks

Data Workbench

- Raise level of abstraction to make it easy to combine data
- Combining data from disparate sources
- Effort needs to be at data centers or distributed, with work results coming back to data workbench
- Don't forget legacy data - it isn't going away and is mature

Research

- Reputation/Influence
- Live documents
- Searching and Visualization

Two examples of e-Science

- Astronomy
 - Jim Gray (Microsoft) and Alex Szalay

- Virtual Observatory
- Federation of 20 observatories
- Results
 - It worked and is used
 - Spatial extensions in SQL 2005
 - Good example of Data Grid
 - Good example of Web Services
- Chemistry
 - Comb-e-Chem Project
 - Digital capture of information
 - Scientists write everything down
 - Digital lab book with tablets

Data Life Cycle

- From Acquisition to Preservation
 - Data Acquisition
 - Ingest
 - Metadata
 - Annotation
 - Provenance
 - Storage
 - Cleansing
 - Mining
 - Curation
 - Preservation
- Case Study: CombeChem
 - Curation of laboratory experimental data as part of the overall data lifecycle
 - Jeremy Frey
 - End to end linking of data and information
 - Collect data with regard to how it could be eventually used
 - Metadata lifecycle
 - Human annotation is always important
 - Observations
 - Where is the data
 - Validation
 - Increasing value of data

- Need to bring all the necessary info together
- Science Blogs?
 - Lab notebook as blog
 - Encourages facilitation
 - Need a data repository behind the blog
 - Useful Chemistry blog
 - A record of an experiment that failed
 - Publishable?
 - Useful?
 - Publication as live documents
 - Can go back to the data, simulations, results
- Virtual organizations
 - Research
 - Research Group
 - Archives
 - International databases
- Data Publishing
 - Databases are replacing publications as a medium of communication
 - Data integration
 - Annotation
 - Provenance
 - Exporting/publishing in agreed formats
 - Security
- OECD

Scholarly Communication

- Open Access to Data and Publications
 - Libraries cannot afford to pay for journals
 - Publicly funded research should be available to all
- Mandates for Open Access
 - US Proposal - Cornyn-Lieberman bill
 - Supported by most top US research universities
 - EU Proposal
 - UK, France, Germany initiatives
- NSF Atkins
- 3 Prophets

- Paul Ginsparg
 - Electronic version of preprints
- David Lipman
 - Portable PMC deployed in UK (PubMedCentral)
- Stevan Harnad
 - OAI compliant institutional repositories
 - JISC funded TARDISProject links full text open access and links to publisher sites

NLM example: Entrez-GenBank

- Sequence data deposited with Genbank

PubMedCentral International

Univ of Illinois - OAI-PMH

- Over 1400 repositories

e-Science Mashups

- Repositories of the future will contain all sorts of complex objects
- We need a more powerful interoperability
- Augmenting interoperability
- New forms of peer review

Web 2.0 + Library = Library 2.0

- Reinvention of the libraries

Digital information lasts forever or five years, whichever comes first

- Optical
- Digital tape
- Magnetic disk

EU Planets Project

- 14M E, 4 year, FP6 project to preserve access to CH data
- Microsoft Open XML OOXML
 - Open, royalty free file format
 - Covenant not to sue by MS
 - Open XML translator project

Technical computing at Microsoft

- Advanced computing
- High productivity computing

- Radical computing

Summary

- Microsoft wishes to work with universities and library communities
 - Interoperable, new technologies
 - Develop interoperable repositories
 - New business models
-