MacKenzie Smith
Associate Director for Technology
MIT Libraries
Institutional Repositories

- Institution-based
- Scholarly material in digital formats
- Cumulative and perpetual
- Open and interoperable
The DSpace Repository

- Institutional Repository for MIT faculty’s digital research materials
- MIT Libraries - Hewlett Packard Research Labs collaborative development project
- Open Source system
- Federated system
- Preservation archive
DSpace

- **Captures**
  - Digital research material in various formats
  - Directly from creators (e.g. faculty)

- **Describes**
  - Descriptive, technical, rights metadata

- **Distributes**
  - Via WWW, with necessary access control

- **Preserves**
DSpace Offerings

- Large-scale, stable, managed long-term storage
- Support for range of digital formats
- Easy-to-use submission process
- Persistent network identifiers
- Access control
- Search and delivery interface
- Digital preservation services
Why Libraries

Expertise
- Large-scale collection management
  - Assessment/collection policies
  - preservation
- Metadata
- Solid business practices

Commitment
- Long time frames
- Mission scope
Possible Content

- Preprints, articles
- Technical Reports
- Working Papers
- Conference Papers
- E-theses
- Datasets
  - statistical, geospatial, matlab, etc.
- Images
  - visual, scientific, etc.
- Audio files
- Video files
- Learning Objects
- Reformatted digital library collections
Information Model

- Communities
- Collections (in communities)
  - Distinct groupings of like items
- Items (in collections)
  - Logical content objects
  - Receive persistent identifier
- Bitstreams (in items)
  - Individual files
  - Receive preservation treatment
Information Model

Versioning

- Item “versions” can be
  - All instances of a work in different formats
    - E.g. the XML, PDF, and PostScript versions
  - All editions of a work over time
    - Official changes (e.g. addenda or new release)
    - Periodic snapshots (e.g. web sites)
- Metadata lists all available versions of items
Communities

- Departments, Labs, Research Centers, Programs, Schools, etc.
- Localized policy decisions
  - Who can contribute, access material
  - Submission workflow
    - Submitters, approvers, reviewers, editors
  - Collections definition, management
- Communities supply metadata
MIT Early Adopters

- Sloan School of Management
- Dept. of Ocean Engineering
- Center for Technology, Policy and Industrial Development (CTPID)
- Lab for Information and Decision Systems (LIDS)

- MIT Press – out-of-print books
MIT Collaborators

- Academic Computing
  - Open Knowledge Initiative
    - Learning Object repository
  - OpenCourseWare
    - Course content, including 3rd party
- Information Systems (computer center)
- W3C
  - Research collaboration
Library Collections

- Libraries another DSpace community
- Reformatted print collections
  - Scanned theses, tech reports, etc.
    - Especially to complete a series
  - Digital image collections, etc.
- 3rd party archiving (e-journals, etc.)
Standards-based

- Modular architecture, well-defined APIs
- 100% open source
  - Programmed in java
  - RDBMS and SQL for metadata
- CNRI “handles” for persistent identifiers
- X.509 certificate-based access control
- OpenURL linking
- OAI-PMH for exposing metadata
OAIS-compliant

- METS AIPs in bitstore
- Designated Community are scholars, researchers
- Knowledge Base
  - Interdisciplinary content
  - Digital archaeology
Technology Stack

- Apache, Tomcat, OpenSSL/mod_ssl
- Java 1.3, JSP 1.2, Servlet 2.3
- PostgreSQL 7, JDBC (rdbms)
- CNRI Handle System 5 (persistent ids)
- Lucene 1.2 (index/search)
- Jena (RDF History system)
- JUnit (testing), Log4j (logging)
- HP/UX, Linux, Solaris, etc.
Metadata

- Qualified Dublin Core
  - based on Library Application Profile
- Crosswalk from MARC
  - based on Library of Congress crosswalk
- Minimally effective preservation metadata
- METS-encoded OAIS AIP in bitstore
- Support for collection/community-specific schemas in development (SIMILE)
System Comparison

- Extends preprint archive model
  - e.g. eprints.org, arXiv.org
  - Accepts non-document material
  - Community paradigm
  - Preservation commitment

- Differs from Digital Library model
  - e.g. FEDORA, Greenstone
  - Faculty-produced content
  - Distributed responsibility
    - Selection, policies, cataloging, submission, etc.
  - No guarantee of future renderability
Challenges

- Faculty Acceptance
  - Valuing and trusting an institutional archive
  - Myriad disciplines with different cultures
  - Copyright/IP policies
- Learning and Sharing
  - Designing DSpace system to federate
  - Tracking digital library research, reporting out
- Sustainability
  - Institutional, financial
- Digital Preservation
Faculty Acceptance

- Variety of content
  - Preprints and publications
  - Digital research material
  - Educational material

- Respect for discipline differences
  - Access control, review process, etc.

- Institutional support
  - Broad advocacy
  - Mission relevance
Business Plan

- One year, Mellon funded project
- Written by business consultants
- Built cost models for running DSpace
- Developed revenue options
  - Core services (free)
  - Premium services (for-fee)
Digital Preservation

Philosophy

- Lots of digital material *is already lost*
- Most digital material is *at risk*
- Better to capture it, guarantee bit preservation than to lose it completely
- Capture maximum information to support functional preservation
- Analyze Cost/benefit tradeoffs
MIT's commitment levels

- **Known/supported**
  - TIFF, SGML/XML, AIFF, PDF

- **Known/unsupported**
  - Microsoft Word, PowerPoint (common)
  - Lotus 1-2-3, Visicalc, WordPerfect (less common)

- **Unknown/unsupported**
  - One-of-a-kind software program
Digital Preservation

- **Supported** = migration and/or emulation
  - Migration for texts, images, audio, etc.
  - Emulation for software, multimedia?
- **Unsupported**
  - Bit preservation at minimum
  - Batch migration where possible
    - Commercial conversion services
Digital Format Registry

- DLF-sponsored initiative
- Centrally hosted, network accessible
- More granular than MIME types
  - E.g. TIFF 5.0 rather than TIFF
  - *Not* TIFF 5.0 600dpi color
- Captures format documentation
  - E.g. specifications, possibly authoring/rendering tools, samples, etc.
  - *Print* and digital
DSpace Federation

Goals

- Drive DSpace development
  - open source development model
- Build critical mass of content
  - support useful interoperation
- Leverage distributed expertise
  - metadata
  - digital preservation
Federation Benefits

- Socio-political
  - Shared direction, leadership, priorities, goals, resources
  - Standards development
    - Putting weight behind “best practices”
    - e.g. W3C, NISO, IETF, ARL/DLF standards
    - Drive commercial developments
Federation Benefits

Technical

- Virtual collections
  - Networked Digital Library of Theses and Dissertations
- New publishing ventures
  - “Overlay” e-journal located at multiple institutions
- Cross-institution/cross-collection search
  - NCSTRL project
  - OAI indexes in any subject
- Distributed services
  - Leverage industry services supporting preservation, etc.
Federation

- Cambridge University Library in the UK
  - 3-year project, focussed on digital preservation
- Mellon project in the US
  - Columbia University, Ohio State University, Universities of Rochester, Toronto, and Washington
- Cornell University
  - Dean of faculty, collaborating with the library
Research Areas

**SIMILE**
- Funded through HP/MIT Alliance
- Flexible metadata infrastructure
- W3C’s Semantic Web activity
- David Karger’s research on personalized information management
- Library provides expertise, test-bed, real-world applications
Research Areas

- Digital preservation
  - Digital files (e.g. audio, video, image, text)
  - Web sites (e.g. W3C)
  - Software programs

- Personal Archiving strategies
  - DSpace on your laptop
  - Proactive collaboration with content creators
Research Areas

- **New Publishing Models**
  - Open Archives Initiative
    - Collection sharing, aggregation
  - Content delivery/publishing infrastructure
  - E-communities (e.g. MIT Press CogNet)
  - E-journals authored and edited by faculty

- **Open, standardized Rights Management**
  - Micropayment support
  - Fine-grained access control (individual subscribers)
Schedule

- Early Adopters (beta release)
  - Now through summer 2002

- MIT public release (DSpace 1.0)
  - October 3, 2002

- Begin federating
  - Starting Fall 2002

- Begin research program
  - Starting Fall 2002

- Open Source to the world
  - November 4, 2002