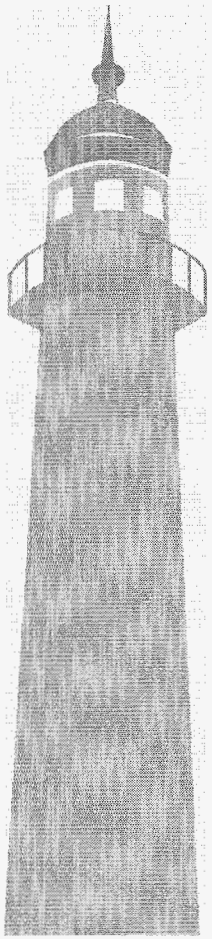
Jet Propulsion Laboratory
California Institute of Technology



DSpace at the Library, Archives and Records Section Status Report NASA STI ViTS

March 11, 2005

Barbara Amago
Jennifer Momjian

Jet Propulsion Laboratory
California Institute of Technology

DSpace Team

- Barbara Amago—Library Supervisor
- Teresa Bailey—Communication Plan
- Eric Li—Unix System Administrator
- Jennifer Momjian—Project Manager
- Nicole Parsons—Data validation and cleanup
- Joshua Phillips—Cataloging Manual, data validation and cleanup
- Iris Young—SQL programmer, workflow, screens

Jet Propulsion Laboratory
California Institute of Technology

Agenda

- Overview
- Why DSpace
- Process Improvement
- System Architecture
- Hardware and Software Configuration
- Data Conversion, Validation and Cleanup
- Data Entry Workflow Modification
- Testing—Functional, Performance Usability
- Documentation
- Customer Help
- Security
- Change Control
- Metrics and Reporting
- Communication Plan

Jet Propulsion Laboratory
California Institute of Technology

Background

- Library's Strategic Plan calls for making finding information "falling off a log easy."
- Old Server
- At capacity, even with additional disk drives
- Limited fields available for metadata creation
- Scripts for formatting descriptions into HTML inconsistent
- Metadata rules inconsistently applied, e.g., author name
- Quality of PDF files varied (some OCR'd, others not)
- Implemented OAI harvesting with NASA TRS, but other issues needed resolution

Jet Propulsion Laboratory
California Institute of Technology

Why DSpace

- Looked for better off-the-shelf solution
- Evaluated eprints and installed test version on Linux server
- Ordered, received, and installed new server
- By the time we were ready to look at new systems, DSpace was being used at a number of institutions
- Decided to test it

Jet Propulsion Laboratory
California Institute of Technology

Process Improvement

- Old:
  - Paper-based
  - Posted preprints to TRS “as cleared”
    - Author submitted paper to be reviewed
    - Cleared papers and forms sent to be filmed in 6-months batches
    - Papers then scanned and OCR’d
    - PDFs posted along with limited metadata
- Problems:
  - No easy process for replacing preprints with final versions
  - Early OCR process caused some data corruption
  - Paper-based
  - Uneven quality of PDFs and metadata

Jet Propulsion Laboratory
California Institute of Technology

Process Improvement (2)

- New:
  - Need to take advantage of electronic environment.
  - Authors submitting papers via email, but no easy way to capture and convert from native format.
  - Improve metadata – more fields, more consistency in data entry, easier to correct errors after posting

Jet Propulsion Laboratory
California Institute of Technology

Process Improvement (3)

- Process for Meeting Papers:
    1. Authors submit papers via email.
    2. Reviewers save in native format. Replace drafts with final copies when received.
    3. Library receives meeting papers in weekly batches, instead of semiannually
    4. Converts files directly from Word or Powerpoint to PDF
    5. After posting, files are sent on for filming
    6. Metadata “rules” applied and followed consistently

Jet Propulsion Laboratory
California Institute of Technology

Process Improvement (4)

- Process for Journal Articles:

  1. Journal articles downloaded directly from journals and posted

  2. Library processes publication charges for authors and captures new papers via that process, so they can be posted soon after journal published.

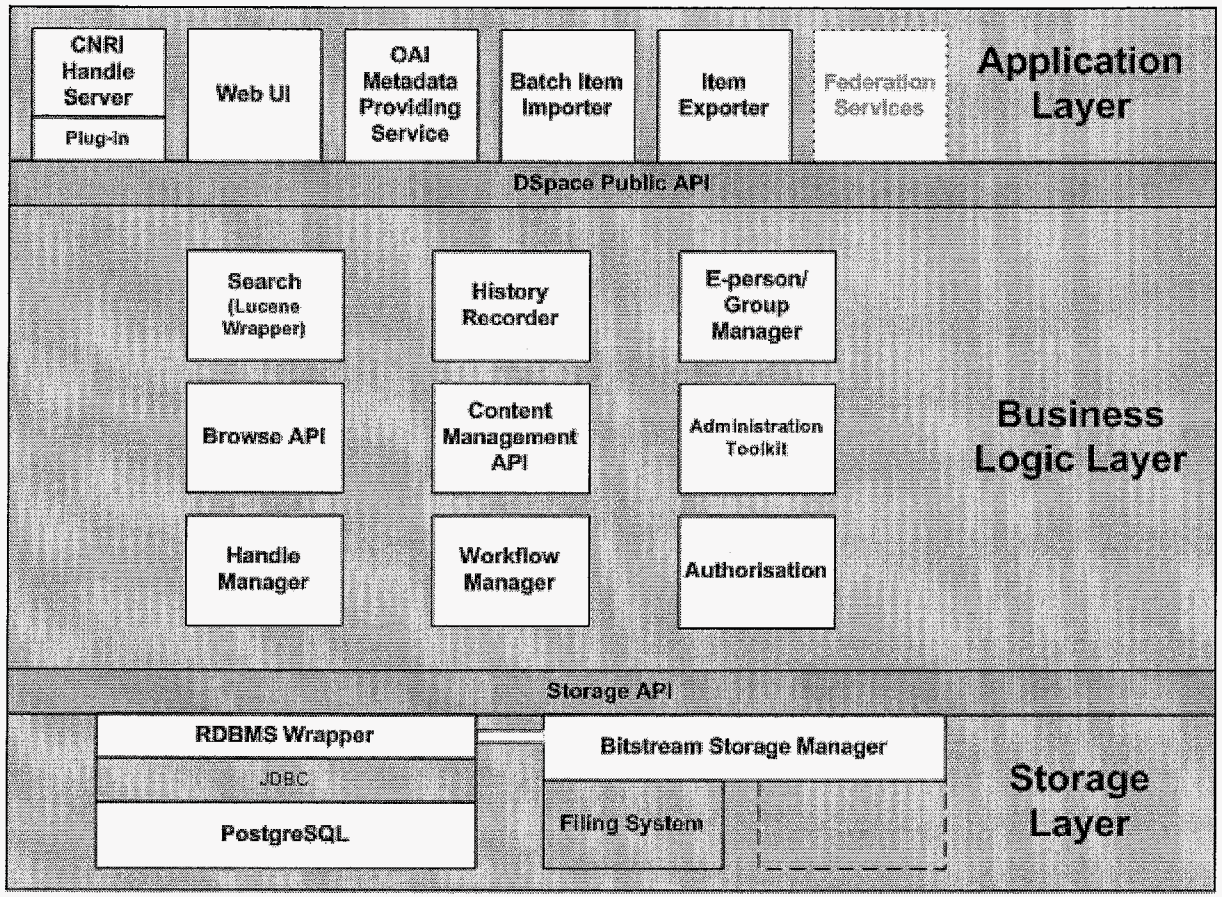
Jet Propulsion Laboratory
California Institute of Technology

DSpace Improvements

- DSpace will allow additional fields, application of AACR2 rules
- Improved accessibility for authors, e.g., posting papers to be given at meetings
- Persistent identifiers for posted papers
- Future possibilities – more of a community concept.
- Authors and projects able to establish their own communities

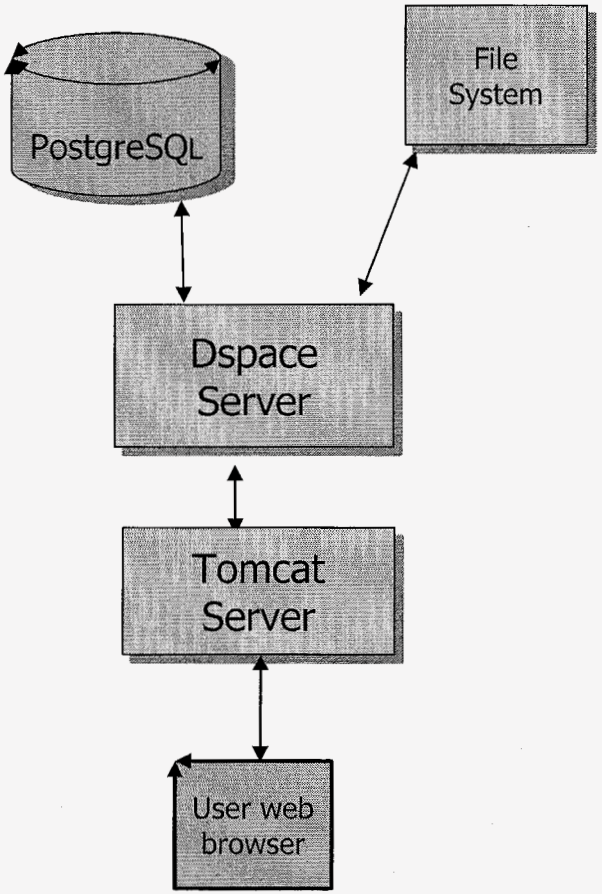
Jet Propulsion Laboratory
California Institute of Technology

System Architecture



Jet Propulsion Laboratory
California Institute of Technology

Dspace System Overview



TomCat Server: Servlet and JSP container
Dspace Server: open source digital repository software
PostgreSQL: database server (store metadata, authorization, etc)
File System: store bitstream(electronic documents)

Jet Propulsion Laboratory
California Institute of Technology

Hardware & Software Configurations

- Hardware
  - SunFire V240
  - 2 x UltraSPARC-IIIi 1.2Ghz CPU
  - 2 GB memory
  - 2 x Ultra160SCSI 73GB disks hot swappable
  - Redundant power with separate power cords
  - Smart UPS 1400

- System Software
  - Solaris 9
- Tomcat Server 5.0.28
  - Dspace 1.2
- PostgreSQL 7.4.3
  - User Software
  - JPL supported browsers

Jet Propulsion Laboratory
California Institute of Technology

Data Conversion

- Data mapping from the JPL Technical Report Server to Dublin Core on DSpace.
- Larger set of Dublin Core to be implemented in workflow for new records.

TRS RFR	Other Value	Dublin Core Element	Dublin Core Qualifier	Display in Simple item Display
%T		title		Yes
%A		contributor	author	Yes
%J		Identifier	citation	yes
%C		Identifier	citation	yes
%D		date	issued	Yes
%K		subject		Yes
%U	[Testonly]	identifier	uri	Yes
%X		description	Abstract	Yes
	[CL #]	identifier	Other	Yes
	Applicatipn/pdf	Format	Mimetype	no

Jet Propulsion Laboratory
California Institute of Technology

Data Validation

- Scripts were written to compare the number of records and pdf's
- All records were converted but pdf's were missing
  - Discovered that some pdf's were not successfully uploaded onto the old TRS when the disk space was full.
  - Recovered the pdf's from hard drive.
- Compared 25 records from each year
  - Discovered that data from 1995 was corrupt.
  - Requires manual cleanup
- Listed data by date, author, and title to reveal anomalies.
  - Discovered that variation in data entry caused problems in browse indexes.
  - Results reported for data cleanup.

Jet Propulsion Laboratory
California Institute of Technology

Data Cleanup

- Dates
  - Some dates could be reformatted with a script.
  - Other dates had to be manually corrected.
  - Some do not exist.
- Authors
  - Names were corrected to improve the author browse.
  - Some corrections could be made by script.
  - Name source was the paper in the pdf.
  - No authority file exists.
- 1995 data
  - Big mess.
  - The data exists along with many other characters
  - Corrections must be done record by record.

Jet Propulsion Laboratory
California Institute of Technology

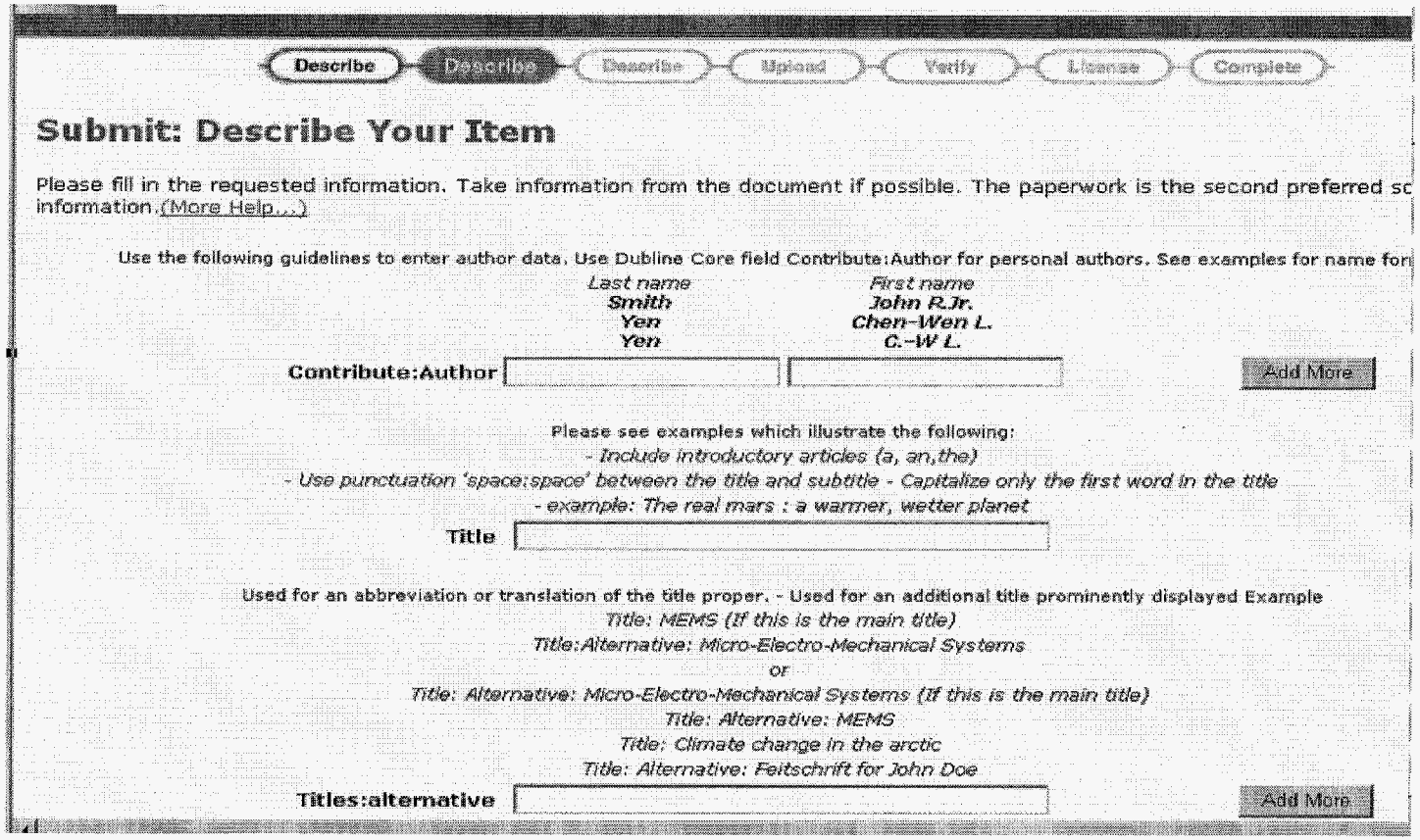
Data Entry Workflow Modification

- Requirements
  - Larger set of Dublic Core--Libraries Working Group Application Profile.
  - Element and qualifier in the label.
  - Enhanced screen instructions.
- Method
  - Set up a Linux box as a development platform.
  - Implemented DSpace 1.2.2 beta 1 which uses XML import form.
  - Modified the screen text, labels and submission forms.
  - Avoided complicated modification to the actual workflow program.
- Other Modification
  - Consolidated collections which were by year into one collection.

Jet Propulsion Laboratory
California Institute of Technology

Screen Text Modifications

- Examples and instructions were added for author and title entries.
- Syntax help. Cataloging Manual for Dublin Core / Joshua Phillips



Jet Propulsion Laboratory
California Institute of Technology

Testing

- Functional testing will be in 2 phases:
  - Out of the box DSpace, version 1.2.1
  - Modified version DSpace 1.2.2 beta 1
- Usability Testing—Library staff
  - Browse, search, print functions
  - Record submissions and edit
  - DSpace-admin screen edits
  - Authentication and authorization
- Load and Performance Testing—
  - Concurrent user testing (not yet done for either search or workflow)
  - Do not plan to run load test programs.

Jet Propulsion Laboratory
California Institute of Technology

Testing (2)

- Browser Testing
  - Netscape 7+
  - Internet Explorer 6+
  - Safari tbd
  - Mozilla tbd
  - Firefox tbd
- Platform Testing
  - Mac System 10
  - PC Windows XP, 2000
  - Sun Solaris
- OAI Harvesting by NTRS

Jet Propulsion Laboratory
California Institute of Technology

Documentation

- Cataloging Manual for Dublic Core / by Joshua Phillips [draft]
- System administration procedures [not yet done]
- Desk procedures / by Barbara Amago.

Jet Propulsion Laboratory
California Institute of Technology

Operations: Customer Help

- Provide Online help
  - DSpace has online help for users
  - Cataloging manual will be available online.
- Provide E-mail and telephone contacts on the web site
- Provide full help service 8x5 Mon - Fri
- Coordinated with JPL EIS and DNS Help Desk Managements for referral of help calls to JPL Library staff.

Jet Propulsion Laboratory
California Institute of Technology

Security

- Application level
  - Records are protected with login requirement
  - Different authorization groups exist for read, write, and delete
- System level
  - TRS/DSpace is part of our Division's group IT Security Plan
  - Subscribe to CERT Advisory and NASA Incident Report Center notification

Jet Propulsion Laboratory
California Institute of Technology

Change Control

- Proposed changes to DSpace will be collected, approved or rejected, and implemented in a systematic manner.
- A Change Control Board will be appointed to decide on proposed changes.
- Changes requiring funding beyond the approved budget will be submitted to Library, Archives, and Records management.
- Major changes will be tested in a test environment.

Jet Propulsion Laboratory
California Institute of Technology

Metrics and Reporting

- Metrics and reports will be collected by the Team lead as requested by Management.
- Basic suite of metric and reports
  - WebTrends Log analyzer Monthly Statistics
  - Number of new records.

Jet Propulsion Laboratory
California Institute of Technology

Communication Plan

- Elements (We've just begun)
- Define the objective-What to we need to accomplish and when
- Determine the audience(s)—who are we trying to reach
- Identify potential barriers—what misunderstandings might occur
- Customize the message—Inside JPL; outside JPL
- Schedule when the message will be delivered—we do not want to be overwhelmed
- Select the communication vehicles—print, electronic, in person
- Solicit feedback—Inside JPL; outside JPL
- Establish metrics—how will we know we met our objective

Jet Propulsion Laboratory
California Institute of Technology

Questions

- Bottom line—when do you go to operations?

  May or June if the creek don't rise.