Grid & Data Preservation

Boon Low
boon.low@ed.ac.uk
System Development, EGEE Training
National e-Science Centre

Topics

- Digital curation and UK Digital Curation Centre
- General preservation issues
- Preservation and data grid
- DSpace + SRB project

Digital curation: a definition

- The actions needed to maintain digital research data and other materials over their life-cycle, for current and future generations.
- These actions include digital archiving and preservation, and good practice in data creation and management.
- Also, providing the capacity for adding value to data to generate new sources of information and knowledge.

Why a national centre?

"Long-term curation and preservation of digital resources is seen as a challenge which is difficult if not impossible for individual institutions to resolve on their own due to the complexity and scale of the challenges involved."

- JISC circular, 6/03

"Scientists and researchers across the UK generate increasingly vast amounts of digital data, with further investment in digitisation and purchase of digital content and information. The scientific record and the documentary heritage created in digital form are at risk from technology obsolescence and by the fragility of digital media."

- JISC press release, 3/04

DCC

Digital Curation Centre

- Established in 2004 under JISC/EPSRC funding
- Continuing quality improvement in data curation & digital preservation practice
  - Initial focus: data as evidence for scholarly conclusions
  - wider remit: scholarly communication & e-Learning
- Working with data repositories, rather than being a data centre
- Centre of excellence in research & service
  - Programmes to address wider issues of data curation
  - Evaluation of tools, standards and policies
  - Focal point for digital curators with repository of tools and technical information
- Connecting communities *via Associates Network*
  - universities & research institutes
  - scientific data tradition & document tradition
  - international & cross-sectoral

DCC

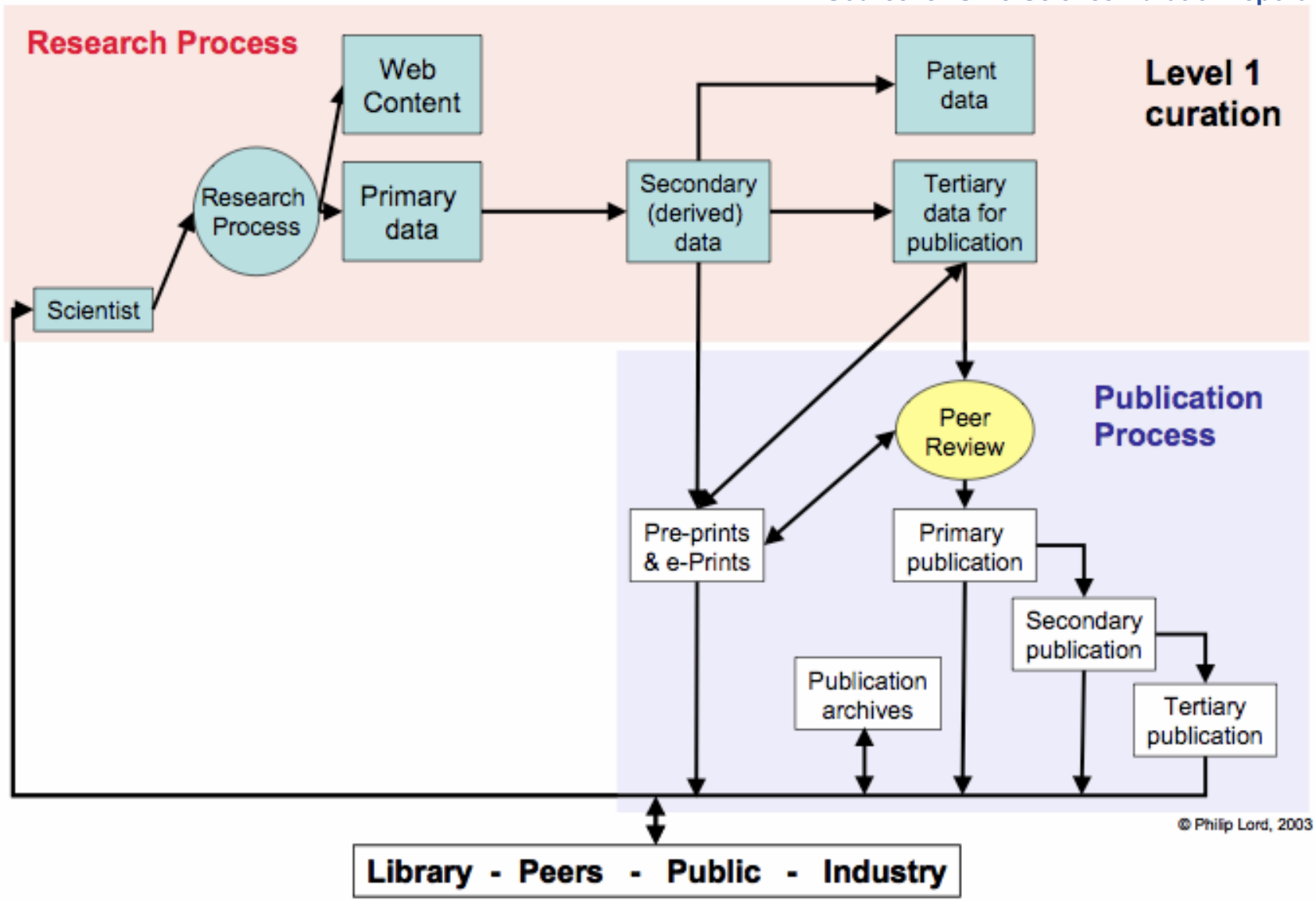
DCC people (some of them…)

- Management & Co-ordination
  - Director Chris Rusbridge (University of Edinburgh)
- Community Support & Outreach
  - Led by Dr Liz Lyon (UKOLN, University of Bath)
- Service Definition & Delivery
  - Led by Professor Seamus Ross (HATII [ERPANET], University of Glasgow)
- Development
  - Led by Dr David Giaretta (Astronomical Software & Services, CCLRC)
- Research
  - Led by Professor Peter Buneman (Informatics, University of Edinburgh)

DCC

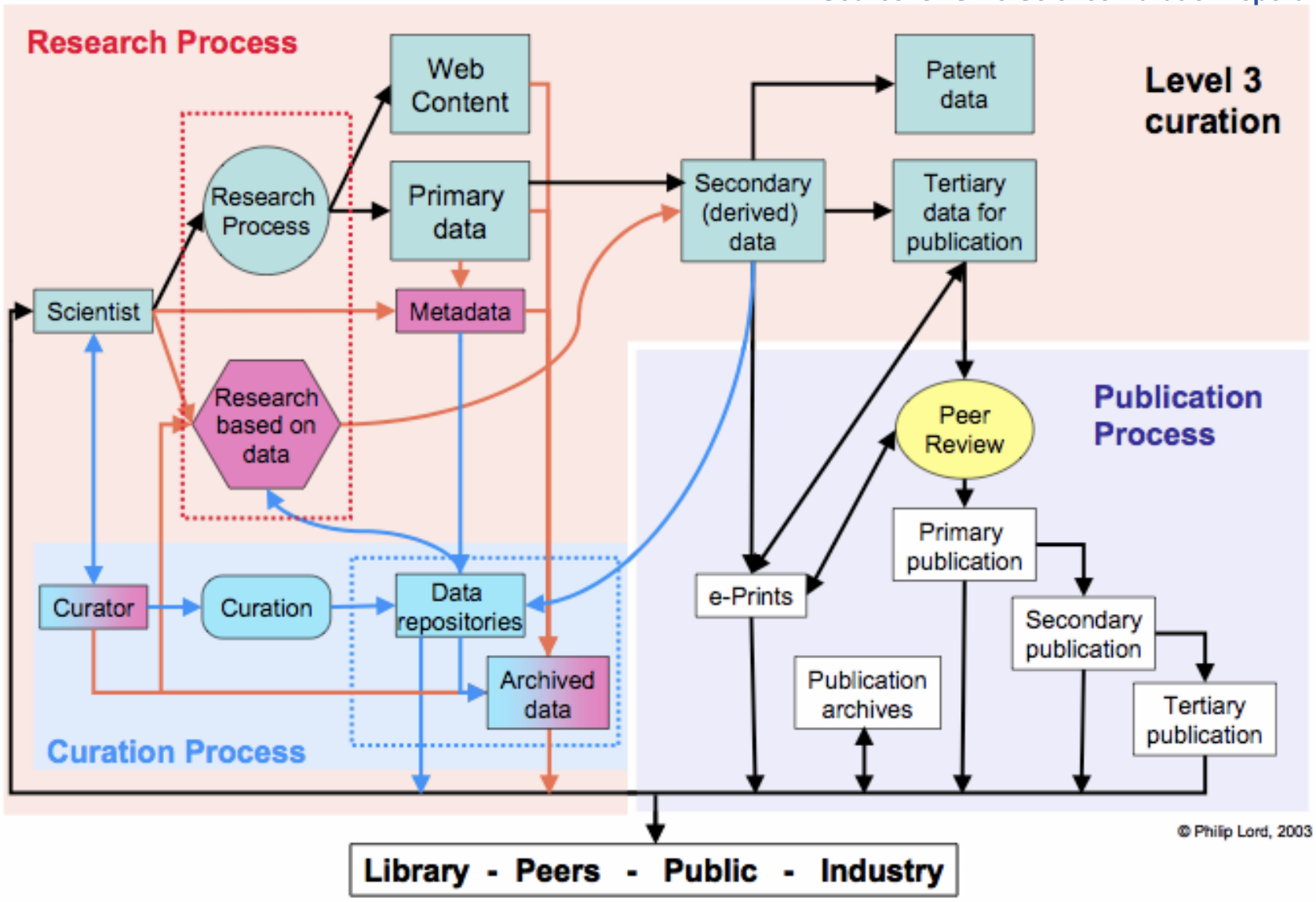
Evolving curation picture

Source: JCSR e-Science Curation report



Evolving curation picture

Source: JCSR e-Science Curation report



Preservation

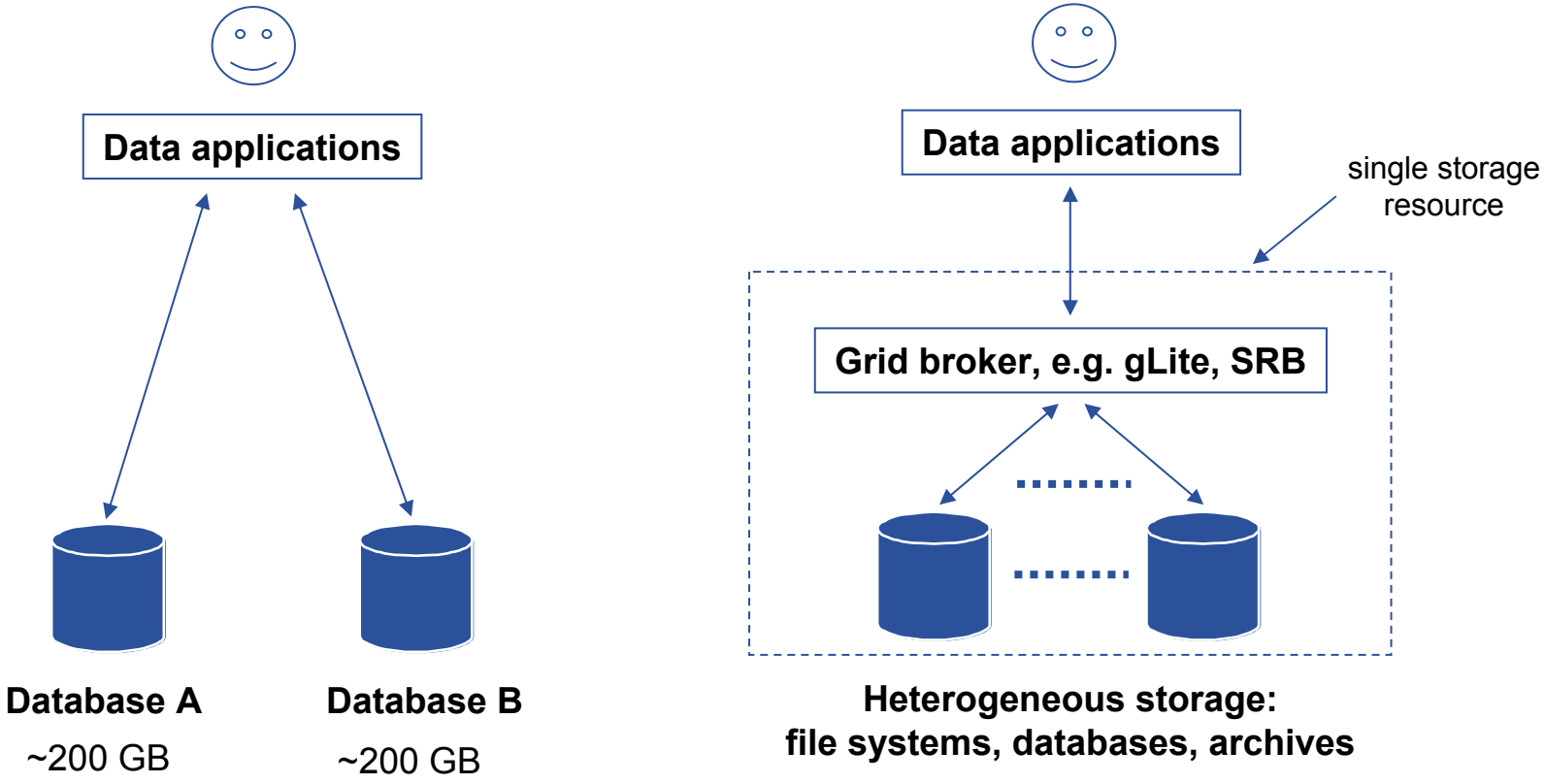
- **Technology changes needs to be addressed to ensure the long termed usage of archives**
- **Changes may stem from applications, OS environments, database systems, hardware and the encoding format of data**

- **Some approaches for preservations:**
  - Emulation: recreating the application in new technology environment while preserving the original data
  - Migration: preserving usability instead of the original data, by transforming it into usable format suitable for new software, technology
  - Preserving data and application contexts such as schema / dtds, or operations applied on data
- **Involves the maintenance of preservation metadata, e.g:**
  - descriptive, authenticity, structural
- **Manages content (the data to be archived) and context (metadata)**

Preservation environment & grid

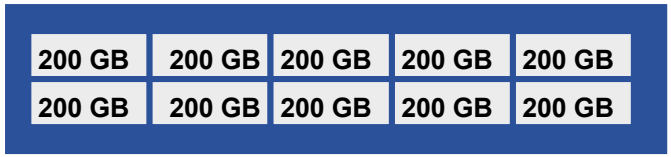
- **Involves extracting data from its creation and application contexts and storing them in a preservation environment**
- **A preservation environment can be built upon the grid infrastructure**
- **Data grid provides mechanisms to manage the evolution of technology infrastructure**
- **Grid middleware such as the SRB can be used to provide abstraction capabilities, for example:**
  - Logical name space for files stored in distributed locations
  - Storage repository abstraction
- **For additional data grid capabilities, see:**
  - Documentation of SRB project
  - http://www.sdsc.edu/srb/Pappres/Pappres.html

Storage repository abstraction

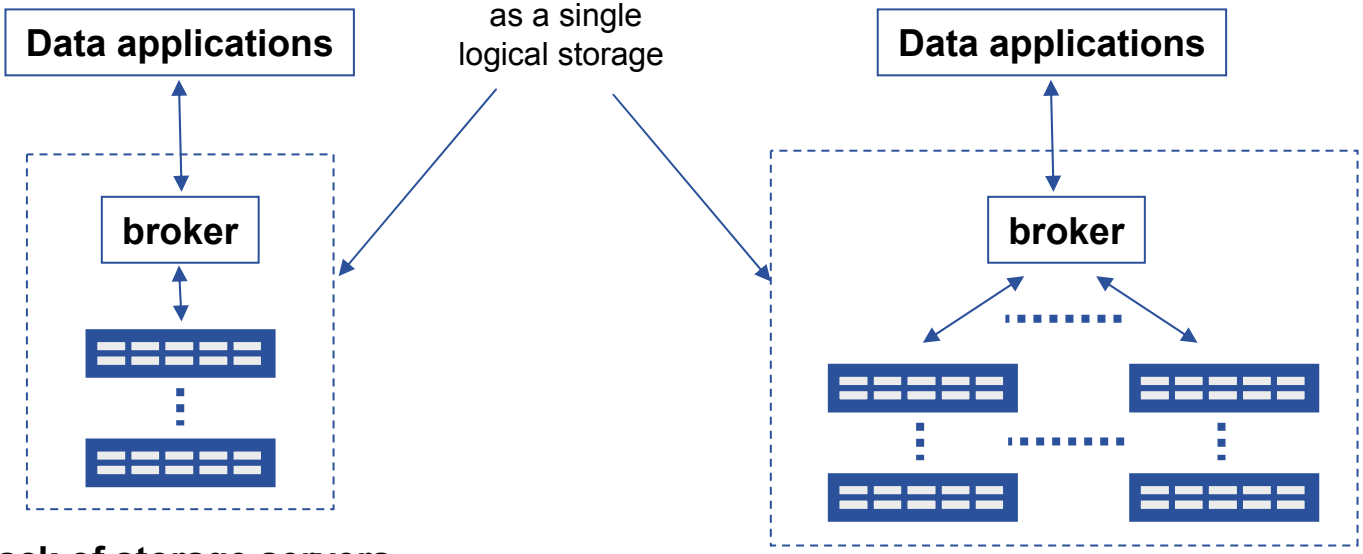


Data grid topology

“Grid Bricks”, grid storage building blocks on dedicated storage server
e.g. 10 x 200GB drives = 2 Terabytes



Data grid -
as a single
logical storage



Rack of storage servers

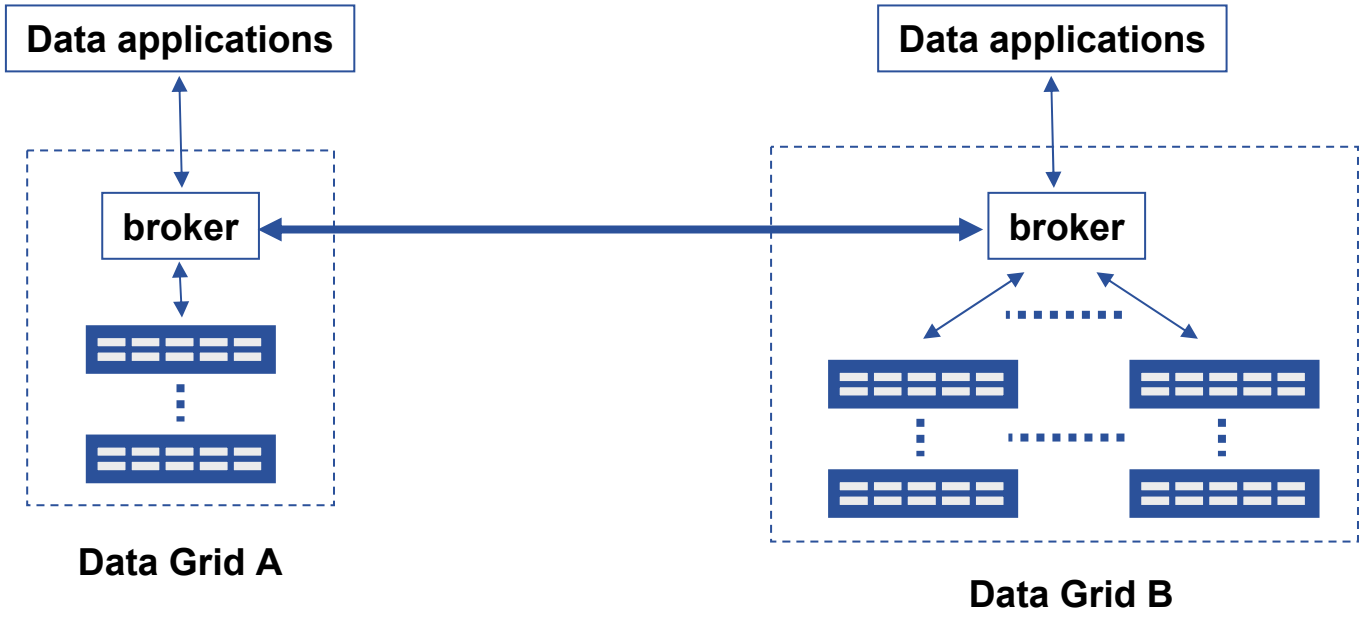
e.g. 5 x 2TB = 10 Terabytes
storage servers

Multiple storage server racks (in a room)

e.g. 5 x 10TB = 50 Terabytes
storage servers

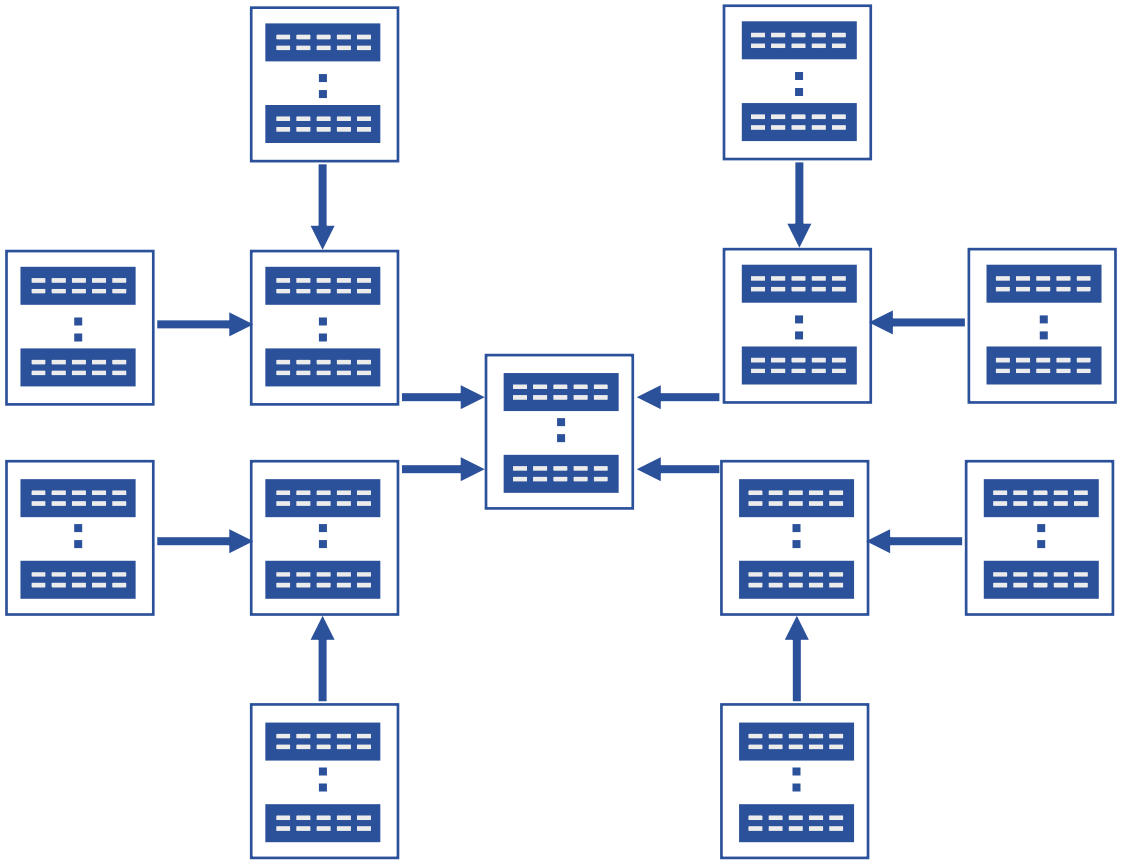
Data grids federation

- **Federation provides mechanisms to organise and manage data on multiple data grids, to extend storage capacity**
- **Interactions among grids is facilitated by the brokers**
- **There various approaches in data grids federations, e.g.:**
  - Applications can share data on Grid A and Grid B as an aggregated data storage
  - Data on a grid can also be replicated automatically on another grid



Data grids federation

- large scale federation, e.g. “snow-flake” federation approach



Federation approaches

Zone SRB	Zone Organization	Zone interaction control	Consistency Management	User Connection Point to access files	Data Access Control Setting	Metadata synchroni-zation	Resource sharing	User-ID sharing between zones
	Zones	Zones	Collections	Files	Files	Metadata	Resources	User names
Free Floating Zones	Peer-to-Peer	Local Admin	User-specified data publication	From home zone	User set access controls	User controlled synchronization	None	None
Occasional Interchange	Peer-to-Peer	Local Admin	User specified	From home zone	User set access controls	User controlled synchronization	None	Partial
Replicated Data Zones	Peer-to-Peer	Local Admin	User-specified replication	From home zone	User set local access controls	User controlled synchronization	Partial	Partial, user establishes own accounts
Resource Interaction	Peer-to-Peer	Local Admin	User-specified replication	From home zone	User set access controls	None	Partial shared resource for replication	Partial
User and Data Replica Zones	Peer-to-Peer	Local Admin	User-specified replication	From home zone	System set access controls	System controlled complete synchronization	Partial	Complete
Replicated Catalog	Peer-to-Peer	Local Admin	System managed name conflict resolution	From any zone	System replicated access controls	System controlled complete synchronization	All zones share resources	Complete
Snow Flake Zones	Hierarchical	Local Admin	System managed replication in hierarchy of zones	From home zone	System set access controls	System controlled partial synchronization	None	One
Master-Slave Zones	Hierarchical	Super Admin	System-managed replication to slave	From home zone	System set access controls	System controlled partial synchronization	None	One
Archival zones	Hierarchical	Super Admin	System-managed versioning to parent zone	From home zone	System set access controls	System controlled complete synchronization	None	Complete
Nomadic Zones	Hierarchical	Local Admin	User-managed replication to parent zone	From home zone	User set access controls	User controlled synchronization	Partial	One

See “Data grids federation” http://www.sdsc.edu/srb/Pappres/Pappres.html

Example: DSpace + SRB project

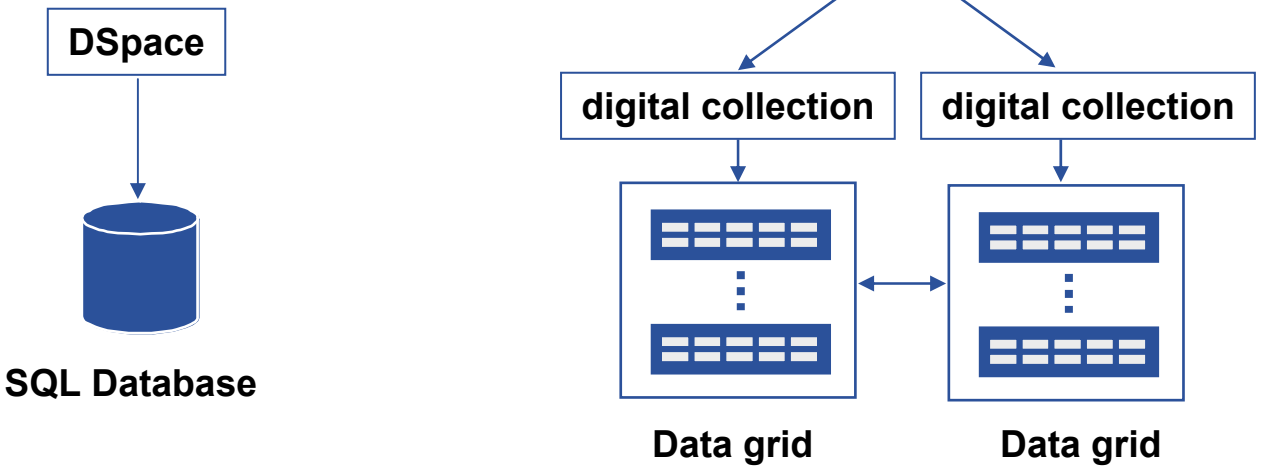
- **DSpace is an open source digital library system providing:**
  - Content/metadata management
  - Collection/user/communities administration
  - Digital content ingestion (batch upload)
  - Indexing, search and discovery
  - Dissemination services (alerting)
  - OAI Harvesting
  - Web UI and API for cross application context development

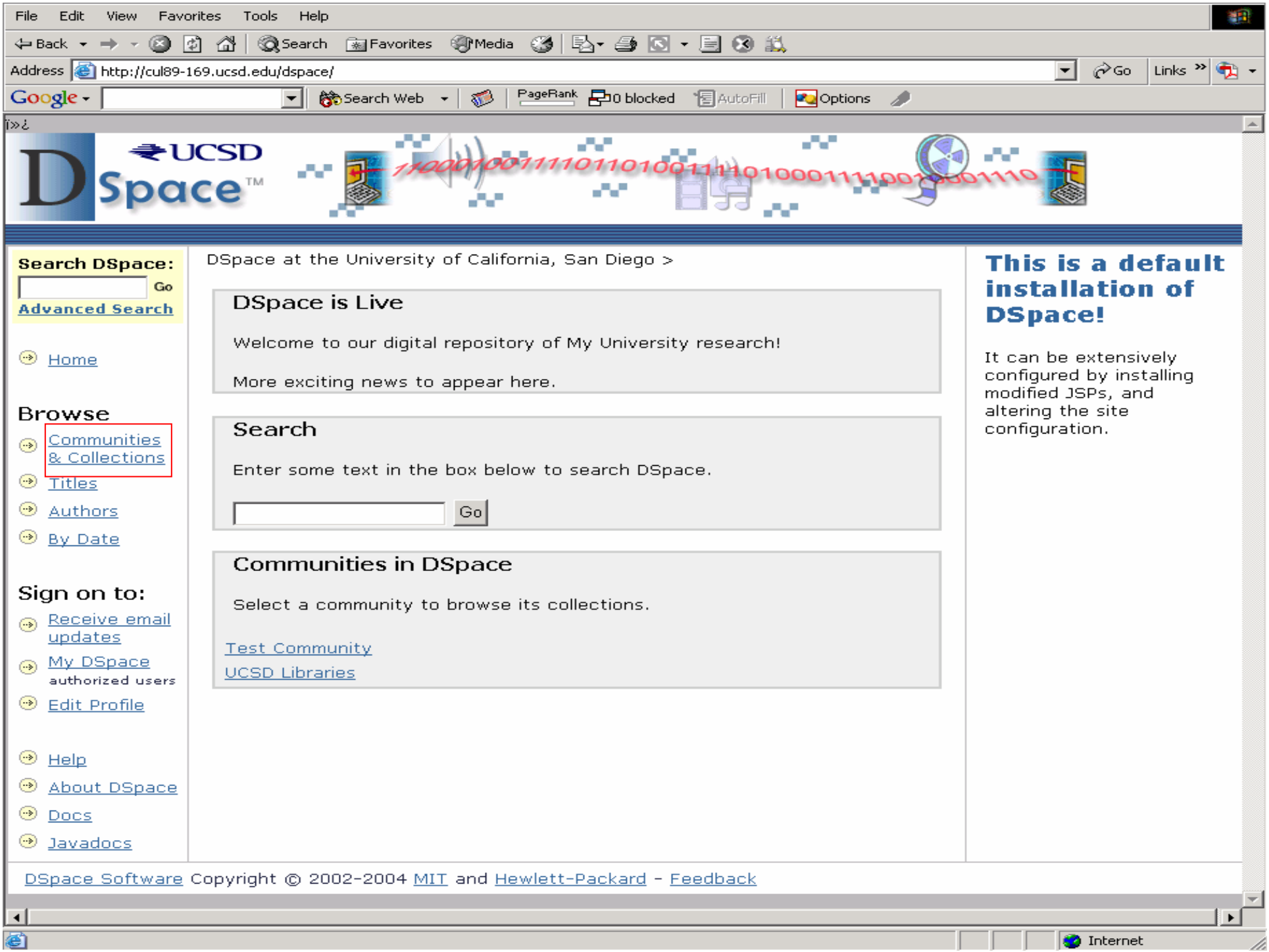
  Jointly developed by:
  - MIT Libraries (MIT)
  - Hewlett-Packard (HP)

- **DSpace + SRB (Storage Resource Broker) is a project by:**
  - San Diego Super Computing Center (SDSC)
  - MIT Libraries (MIT)
  - UC San Diego Libraries (UCSD)
  - US National Archives and Records Administration (NARA)

Example: DSpace + SRB project

- Goal is to extends DSpace storage capability by using data grid, in addition the existing SQL database system
- Replace DSpace file system calls with access calls to data grid
- Uses METS based Archival Information Package (AIP)




File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://cul89-169.ucsd.edu/dspace/
Go Links
Google Search Web PageRank 0 blocked AutoFill Options
UCSD DSpace™
Search DSpace:
Go
Advanced Search
Home
Browse
Communities & Collections
Titles
Authors
By Date
Sign on to:
Receive email updates
My DSpace
authorized users
Edit Profile
Help
About DSpace
Docs
Javadocs
DSpace at the University of California, San Diego >
DSpace is Live
Welcome to our digital repository of My University research!
More exciting news to appear here.
Search
Enter some text in the box below to search DSpace.
Go
Communities in DSpace
Select a community to browse its collections.
Test Community
UCSD Libraries
This is a default installation of DSpace!
It can be extensively configured by installing modified JSPs, and altering the site configuration.
DSpace Software Copyright © 2002-2004 MIT and Hewlett-Packard - Feedback
Internet

Search DSpace:
Go
Advanced Search

- Home

Browse

- Communities & Collections
- Titles
- Authors
- By Date

Sign on to:

- Receive email updates
- My DSpace authorized users
- Edit Profile

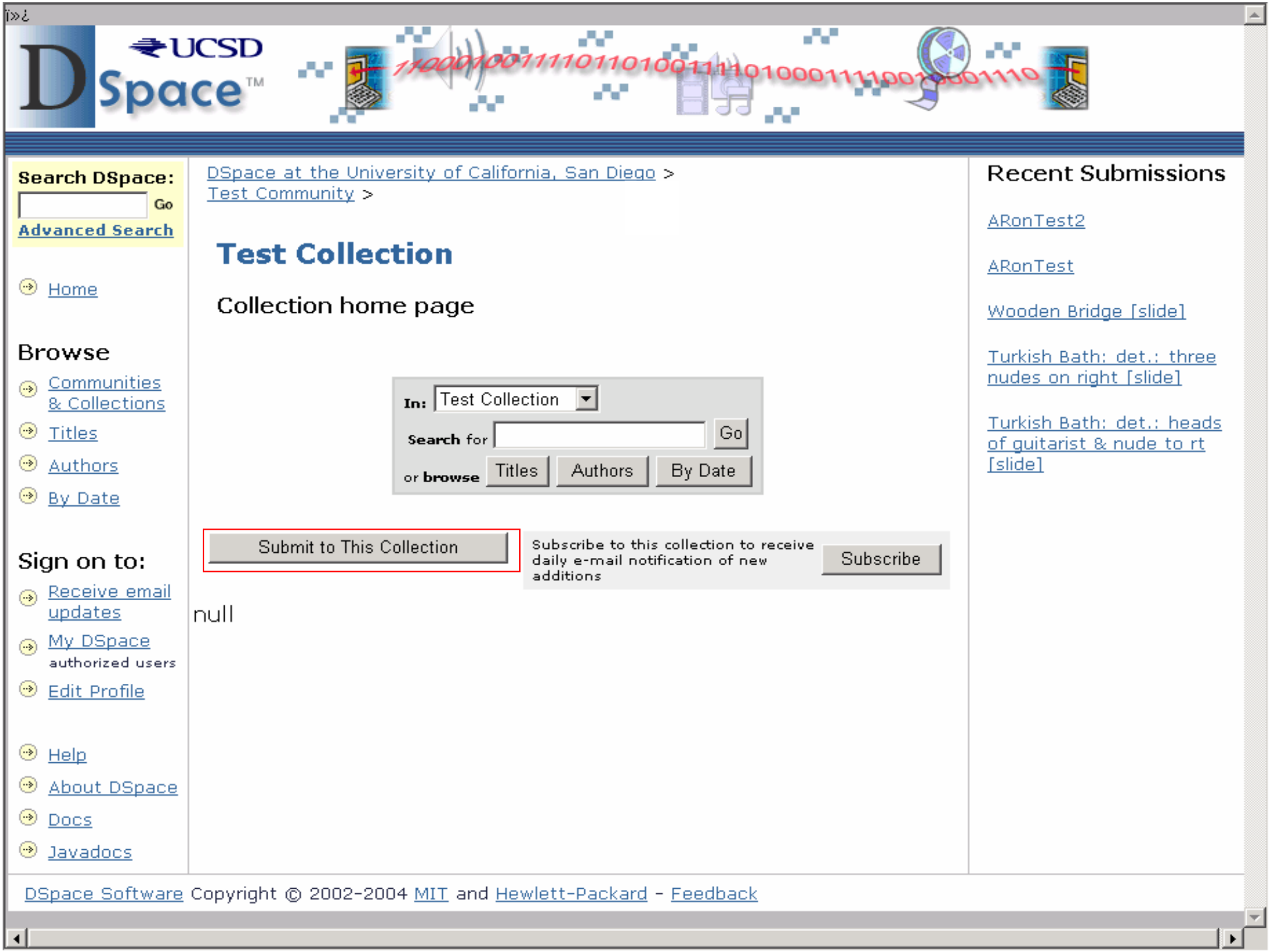
- Help
- About DSpace
- Docs
- Javadocs

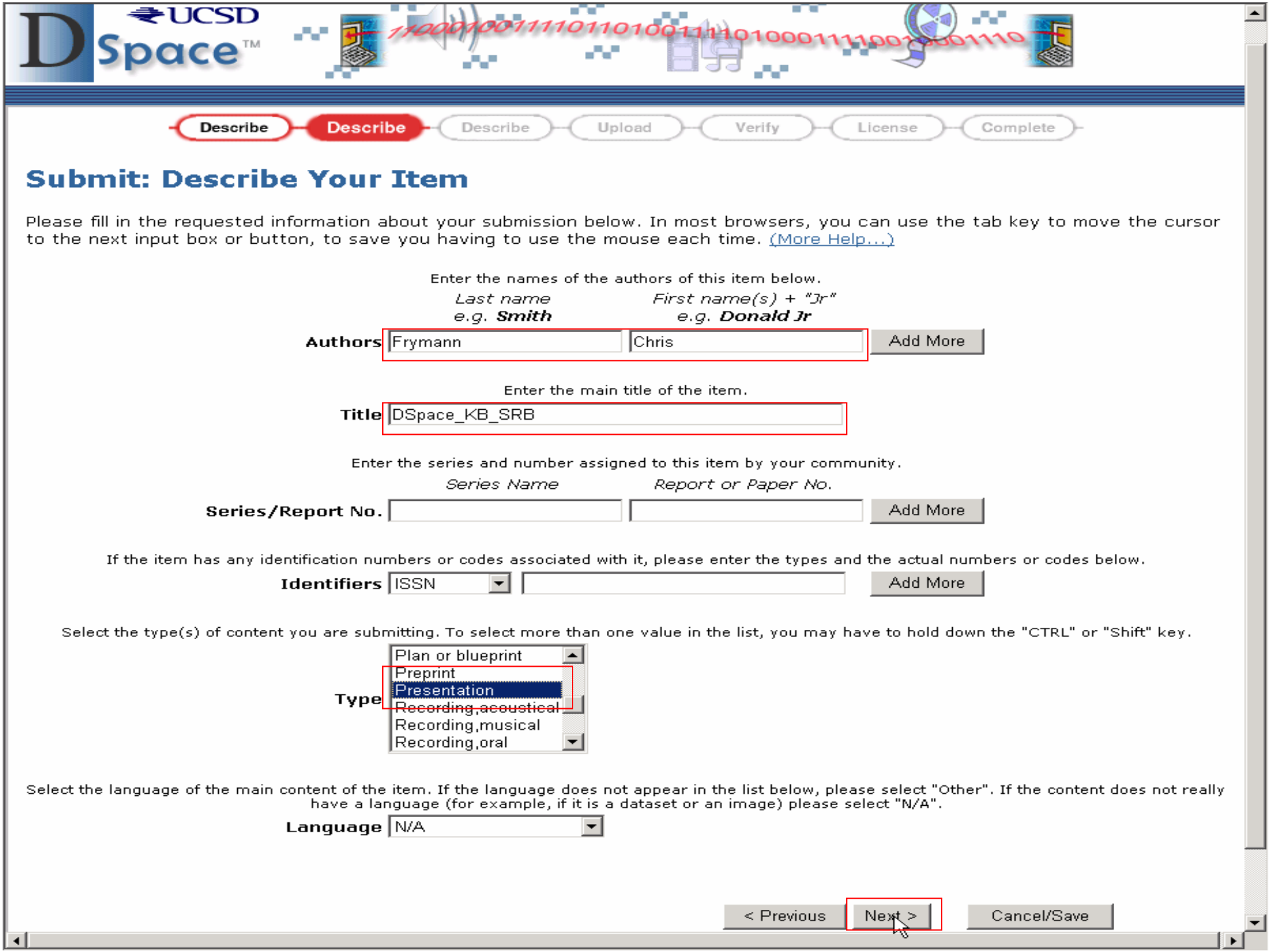
DSpace at the University of California, San Diego >

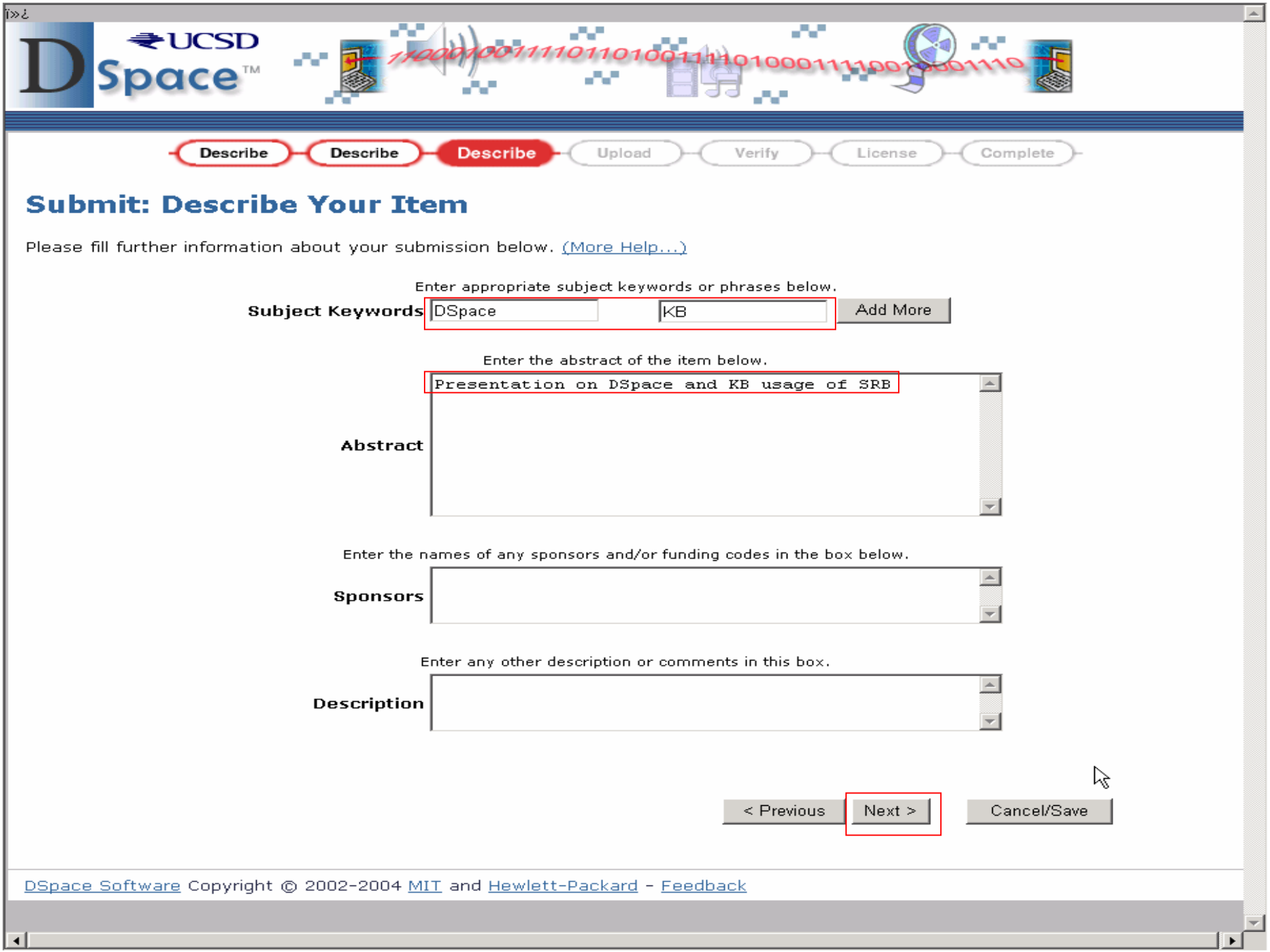
Communities and Collections

Shown below is a list of communities and the collections and sub-communities within them. Click on a name to view that community or collection home page.

- **Test Community**
  - Test Collection
- **UCSD Libraries**


UCSD
DSpace™
Search DSpace:
Go
Advanced Search
Home
Browse
Communities & Collections
Titles
Authors
By Date
Sign on to:
Receive email updates
My DSpace
authorized users
Edit Profile
Help
About DSpace
Docs
Javadocs
DSpace at the University of California, San Diego >
Test Community >
Test Collection
Collection home page
In: Test Collection
Search for
Go
or browse
Titles
Authors
By Date
Submit to This Collection
Subscribe to this collection to receive daily e-mail notification of new additions
Subscribe
null
Recent Submissions
ARonTest2
ARonTest
Wooden Bridge [slide]
Turkish Bath: det.: three nudes on right [slide]
Turkish Bath: det.: heads of guitarist & nude to rt [slide]
DSpace Software Copyright © 2002-2004 MIT and Hewlett-Packard - Feedback


UCSD
DSpace™
Describe
Describe
Describe
Upload
Verify
License
Complete
Submit: Describe Your Item
Please fill in the requested information about your submission below. In most browsers, you can use the tab key to move the cursor to the next input box or button, to save you having to use the mouse each time. (More Help...)
Enter the names of the authors of this item below.
Last name e.g. Smith
First name(s) + "Jr" e.g. Donald Jr
Authors
Frymann
Chris
Add More
Enter the main title of the item.
Title
DSpace_KB_SRB
Enter the series and number assigned to this item by your community.
Series Name
Report or Paper No.
Series/Report No.
Add More
If the item has any identification numbers or codes associated with it, please enter the types and the actual numbers or codes below.
Identifiers
ISSN
Add More
Select the type(s) of content you are submitting. To select more than one value in the list, you may have to hold down the "CTRL" or "Shift" key.
Type
Plan or blueprint
Preprint
Presentation
Recording,acoustical
Recording,musical
Recording,oral
Select the language of the main content of the item. If the language does not appear in the list below, please select "Other". If the content does not really have a language (for example, if it is a dataset or an image) please select "N/A".
Language
N/A
< Previous
Next >
Cancel/Save


UCSD
DSpace™
Describe
Describe
Describe
Upload
Verify
License
Complete
Submit: Describe Your Item
Please fill further information about your submission below. (More Help...)
Enter appropriate subject keywords or phrases below.
Subject Keywords
DSpace
KB
Add More
Enter the abstract of the item below.
Presentation on DSpace and KB usage of SRB
Abstract
Enter the names of any sponsors and/or funding codes in the box below.
Sponsors
Enter any other description or comments in this box.
Description
< Previous
Next >
Cancel/Save
DSpace Software Copyright © 2002-2004 MIT and Hewlett-Packard - Feedback

Describe | Describe | Describe | Upload | Verify | License | Complete

Submit: Upload a File

Please enter the name of the file on your local hard drive corresponding to your item. If you click "Browse...", a new window will appear in which you can locate and select the file on your local hard drive. (More Help...)

Netscape users please note: By default, the window brought up by clicking "Browse..." will only display files of type HTML. If the file you are uploading isn't an HTML file, you will need to select the option to display files of other types. Instructions for Netscape users are available.

Please also note that the DSpace system is able to preserve the content of certain types of files better than other types. Information about file types and levels of support for each are available.

Document File: d Settings\Chris Frymann\Desktop\DS_KB.ppt Browse...

< Previous | Next > | Cancel/Save

Describe | Describe | Describe | Upload | Verify | License | Complete

Submit: File Uploaded Successfully

Your file was successfully uploaded.

Here are the details of the file you have uploaded. Please check the details before going to the next step. (More Help...)

File	Size	File Format
DS_KB.ppt	552448 bytes	Microsoft Powerpoint (known)

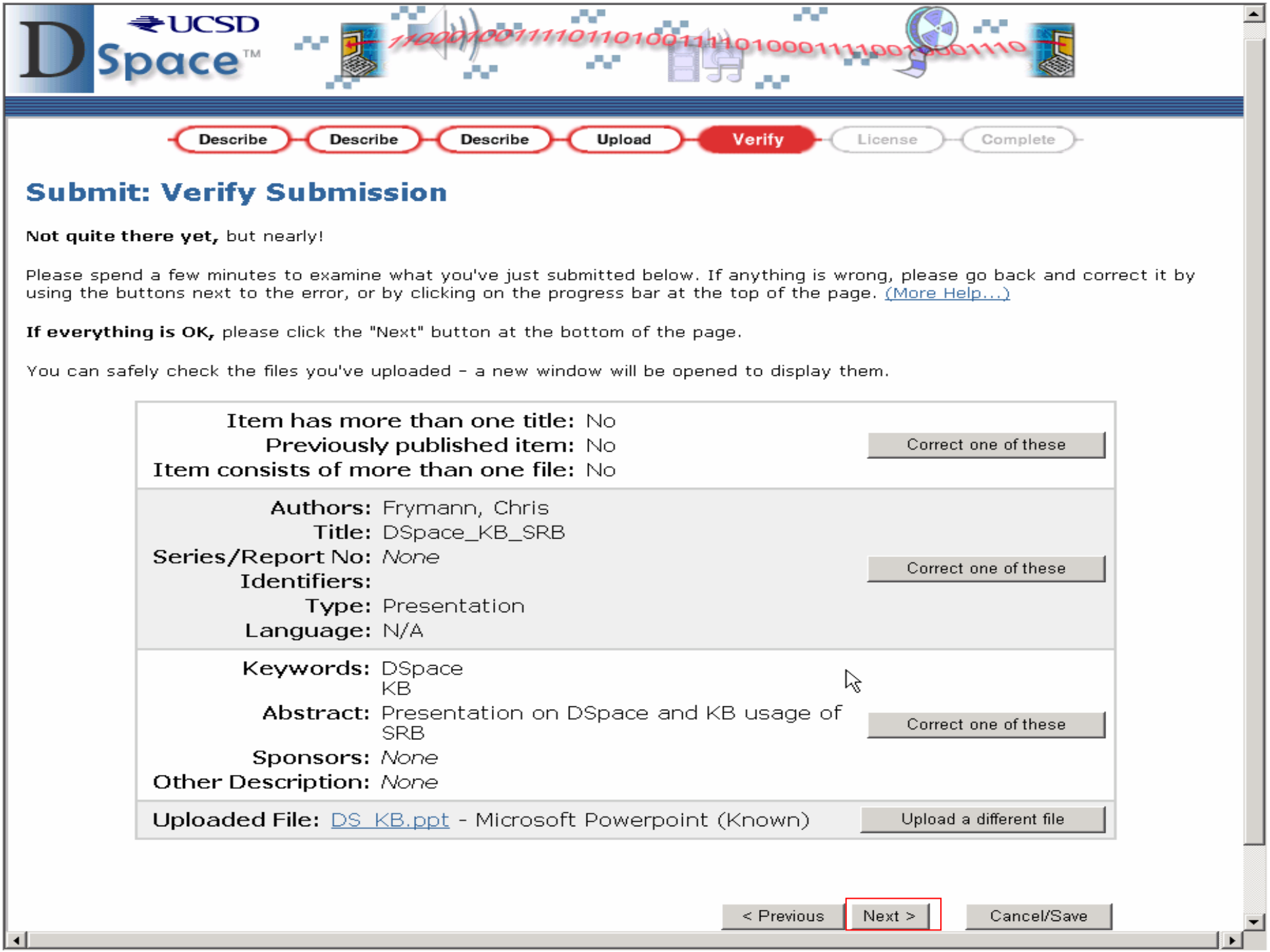
Click here if this is the wrong format

Click here if this is the wrong file

You can verify that the file has been uploaded correctly by:

- Clicking on the filename above. This will download the file in a new browser window, so that you can check the contents.
- The system can calculate a checksum you can verify. Click here for more information. Show checksums

< Previous | Next > | Cancel/Save


UCSD
DSpace™
Describe
Describe
Describe
Upload
Verify
License
Complete
Submit: Verify Submission
Not quite there yet, but nearly!
Please spend a few minutes to examine what you've just submitted below. If anything is wrong, please go back and correct it by using the buttons next to the error, or by clicking on the progress bar at the top of the page. (More Help...)
If everything is OK, please click the "Next" button at the bottom of the page.
You can safely check the files you've uploaded - a new window will be opened to display them.
Item has more than one title: No
Previously published item: No
Item consists of more than one file: No
Correct one of these
Authors: Frymann, Chris
Title: DSpace_KB_SRB
Series/Report No: None
Identifiers:
Type: Presentation
Language: N/A
Correct one of these
Keywords: DSpace
KB
Abstract: Presentation on DSpace and KB usage of SRB
Sponsors: None
Other Description: None
Correct one of these
Uploaded File: DS_KB.ppt - Microsoft Powerpoint (Known)
Upload a different file
< Previous
Next >
Cancel/Save

Describe | Describe | Describe | Upload | Verify | License | Complete

Submit: Grant DSpace Distribution License

There is one last step: In order for DSpace to reproduce, translate and distribute your submission worldwide, your agreement to the following terms is necessary. Please take a moment to read the terms of this license, and click on one of the buttons at the bottom of the page. By clicking on the "Grant License" button, you indicate that you grant the following terms of the license. (More Help...)

Not granting the license will not delete your submission. Your item will remain in your "My DSpace" page. You can then either remove the submission from the system, or agree to the license later once any queries you might have are resolved.

NOTE: PLACE YOUR OWN LICENSE HERE
This sample license is provided for informational purposes only.

NON-EXCLUSIVE DISTRIBUTION LICENSE

By signing and submitting this license, you (the author(s) or copyright owner) grants to the University of California, San Diego (UCSD) the non-exclusive right to reproduce, translate (as defined below), and/or distribute your submission (including the abstract) worldwide in print and electronic format and in any medium, including but not limited to audio or video.

You agree that UCSD may, without changing the content, translate the submission to any medium or format for the purpose of preservation.

You also agree that UCSD may keep more than one copy of this submission for purposes of security, back-up and preservation.

You represent that the submission is your original work, and that you have the right to grant the rights contained in this license. You also represent that your submission does not, to the best of your knowledge, infringe upon anyone's copyright.

If the submission contains material for which you do not hold copyright, you represent that you have obtained the unrestricted permission of the copyright owner to grant UCSD the rights required by this license, and that such third-party owned material is clearly identified and acknowledged


UCSD
DSpace™
Logged in as drlittle@ucsd.edu (Logout)
Search DSpace:
Go
Advanced Search
Home
Browse
Communities & Collections
Titles
Authors
By Date
Sign on to:
Receive email updates
My DSpace
authorized users
Edit Profile
Help
About DSpace
Docs
Javadocs
DSpace at the University of California, San Diego >
DSpace is Live
Welcome to our digital repository of My University research!
More exciting news to appear here.
Search
Enter some text in the box below to search DSpace.
Go
Communities in DSpace
Select a community to browse its collections.
Test Community
UCSD Libraries
This is a default installation of DSpace!
It can be extensively configured by installing modified JSPs, and altering the site configuration.
DSpace Software Copyright © 2002-2004 MIT and Hewlett-Packard - Feedback

For further information

Curation, preservation, data grid
http://www.dcc.ac.uk
http://www.sdsc.edu/srb/Pappres/Pappres.html

DSpace + SRB project:
http://dspace.org
http://libnet.ucsd.edu/nara/
http://wiki.dspace.org/DspaceSrbIntegration