



THE NATIONAL SCIENCE DIGITAL LIBRARY

# Preservation Environments



Reagan Moore  
Charlie Cowart  
San Diego Supercomputer Center  
October 13, 2003

# Preservation Strategy

- Based on use of data grids to manage technology evolution
- Crawl URLs listed in NSDL central repository
  - Recursively retrieve to a depth of 10
  - Map internal URL links to logical name
  - Aggregate digital entities into containers
  - Archive in HPSS at SDSC, or at another site

# Preservation Approach

- Provide mechanisms to:
  - Create archival context for your content
    - Context is preservation metadata (provenance, administrative, descriptive, structural, behavioral)
    - Content is the submitted digital entity
  - Assert authenticity - the consistency between the context and the content
- Provide mechanisms to manage:
  - Persistent identifiers
  - Technology evolution (encoding standard, storage repositories, information repositories, access methods)

# Persistent Archive Collections

- Build collections based on date crawled
- For each collection, use separate folder to hold digital entities associated with the original URL
  - Typically 100 digital entities per URL
- What preservation metadata should be provided within the archive?
  - Administrative, descriptive, structural, behavioral

# A Few Statistics on NSDL Content

Drawn from the SDSC Crawl (April 03, 4 Links Deep)

received correctly	—	1,530,206
no data received	—	51
see other	—	5
forbidden	—	311
file not found	—	38,386
internal server error	—	946
application error	—	15
service temp. overloaded	—	8
WIMS User Error	—	1
Gone	—	1
unused	—	1
redirection w/out location	—	1
<b>total digital entities</b>	<b>—</b>	<b>1,569,932</b>
<b>error percentage</b>	<b>—</b>	<b>2.53%</b>

# Encoding Formats Present in Archive

<b>Digital Entity Type</b>	<b>Number of files</b>
html	331557
gif	157891
jpg	136445
xml	21528
txt	17433
pdf	9369
css	4073
doc	862
asp	819
ppt	161
xls	15

CSS - Cascading Style Sheet

ASP - Microsoft Active Server Page

# Current Crawl - to a Depth of 10

- Expect 25 days to complete
  - 93,000 URLs excluding arXiv.org
  - Limited retrieval to a request every 2 seconds
- Retrieved in current crawl
  - 20,023 URLs
  - 2,584,142 digital entities (129 entities per URL)
  - 137.2 Gbytes (53 kBytes per entity)
- Can increase retrieval rate by decreasing time between requests

# Technical Questions on Preservation

- What types of preservation are of most interest for your collection?
- What type of authenticity information is needed for your material (dates, versions, audit trails)?
- What types of data formats do you expect to use for the material, that can be viewed and displayed in the future?
- Is your project providing alternate mechanisms to ensure preservation of their material?



# Technical Questions on Preservation

- Long-term Host for collection
- Where does your material go at the end of the project?
- Is the NSDL archive of web crawls appropriate for long-term preservation of your material?
- How should the NSDL link to your long-term preservation system to ensure access to the material?
- Should a copy of the crawl reside on your resources?

# mySRB Interface to the Persistent Archive

**MySRB**  
[MySRB Login Page](#)  
[Use Ticket-based Access](#)  
[Change MySRB Password](#)  
[Forgot MySRB Password](#)  
[Register As New User](#)  
[MySRB Online Help](#)

**MySRB Login**  
SRB User Name:   
SRB User Domain Name:   
SRB User Password:   
SRB Port Number:   
SRB Host:

Please exit SRB once you have finished

Session will timeout in:  minutes

---

**MySRB - A Quick Introduction**

- **What is MySRB?** MySRB is a web-based **browse and search** interface to the [Storage Resource Broker](#) (SRB) developed at the San Diego Supercomputer Center ([SDSC](#)). The SRB facilitates information sharing by allowing users (1) to access files stored on heterogeneous resources including disks, tapes and databases on different machines through logically organized catalogs; and (2) to manage and share data collections in a secure manner.
- **Who can use MySRB?** Currently, MySRB is restricted to users who have computer accounts at SDSC, and who have [registered for MySRB](#). If you are not at SDSC and would like to use MySRB please contact [srb@sdsc.edu](mailto:srb@sdsc.edu).
- **Can I use MySRB from anywhere?** Yes, except during registration. Once you are registered user of MySRB, you can access it from anywhere via web browsers. When registering, the browser must be running inside the SDSC firewall (i.e., IP address of 132.249.x.x).
- **What about security?** MySRB uses secure-http (https) protocol using 128-bit RSA authentication. As an extra degree of protection, your browser receives a unique session key when you login; once its time limit (default 60 minutes) expires, you must login again in order to continue.
- Additional Information can be found at: [MySRB Online Help Page](#).
- Please send enquiries or complaints to [srb@sdsc.edu](mailto:srb@sdsc.edu).

# Archive Collection Hierarchy

- NSDL collection
  - List of Time-based snapshots
    - List of URLs
      - List of entities per URL
- /home/nsdl
  - 2003-06-10T13:53:37z/
    - oai:nsdl.org:GROW:70/
      - oai.nsdl.org.GROW.70.html

## View All Metadata

Collection: **oai:nsdl.org:GROW:70**  
 Parent Collection: **/home/nsdl.sdsc/2003-06-10T13:53:37Z**  
 Owner: **nsdl@sdsc**

More Metadata Found

[Explore SRB](#)[Ingest File](#)[Create File](#)[Register File](#)[Register Directory](#)[Register URL](#)[Register SQL](#)[Register ORBData](#)[Register Command](#)[Make Collection](#)[Make Container](#)[Browse Query](#)[Other Info](#)[Online Help](#)[Exit MySRB](#)

</home/nsdl.sdsc/2003-06-10T13:53:37Z/oai:nsdl.org:GROW:70>

Move Up To Collection:

Function	Data Name	Creation Time	Owner	Replica Number	Version Number	Size	Data Type	Resource	In Container
Get File	<a href="#">back.gif</a>	2003-06-12-15.14.49	nsdl@sdsc	0	0	1607	gif image	ux-srbbrick1	Yes
Get File	<a href="#">fao.css</a>	2003-06-12-15.13.07	nsdl@sdsc	0	0	1570	generic	ux-srbbrick1	Yes
Get File	<a href="#">next.gif</a>	2003-06-12-15.14.35	nsdl@sdsc	0	0	1491	gif image	ux-srbbrick1	Yes
Get File	<a href="#">oai:nsdl.org:GROW:70.html</a>	2003-10-10-17.31.20	nsdl@sdsc	0	0	270	html	ux-srbbrick1	Yes
Get File	<a href="#">oai:nsdl.org:GROW:70.html</a>	2003-10-10-17.32.24	nsdl@sdsc	0	0	270	generic	ux-srbbrick1	No
Get File	<a href="#">oai:nsdl.org:GROW:70.xml</a>	2003-06-12-15.12.39	nsdl@sdsc	0	0	2413	generic	ux-srbbrick1	Yes
Get File	<a href="#">output.txt</a>	2003-06-12-15.12.53	nsdl@sdsc	0	0	4043	ascii text	ux-srbbrick1	Yes
Get File	<a href="#">toc.gif</a>	2003-06-12-15.14.29	nsdl@sdsc	0	0	1438	gif image	ux-srbbrick1	Yes
Get File	<a href="#">top.gif</a>	2003-06-12-15.14.42	nsdl@sdsc	0	0	1601	gif image	ux-srbbrick1	Yes
Get File	<a href="#">w2598e.jpg</a>	2003-06-12-15.13.14	nsdl@sdsc	0	0	6985	jpeg image	ux-srbbrick1	Yes
Get File	<a href="#">w2598e00.gif</a>	2003-06-12-15.15.37	nsdl@sdsc	0	0	1458	gif image	ux-srbbrick1	Yes
Get File	<a href="#">w2598e00.htm</a>	2003-06-12-15.13.00	nsdl@sdsc	0	0	17842	html	ux-srbbrick1	Yes
Get File	<a href="#">w2598e00.jpg</a>	2003-06-12-15.15.03	nsdl@sdsc	0	0	18761	jpeg image	ux-srbbrick1	Yes

Internet zone



MYSRB V8.0

Back Forward Stop Refresh Home AutoFill Print Mail

Address: <https://srb.npacl.edu/cgi-bin/demo/mysrb2.cgi?function=full&dataname=oai.nsd.org.GROW.70.html&collection=/home/nsdl.sdsc/2003-06-10T13:53:37Z/oai:nsdl.org:GROW:70> go

Live Home Page Apple Apple Support Apple Store .Mac Mac OS X Microsoft MacTopia Office for Macintosh MSN

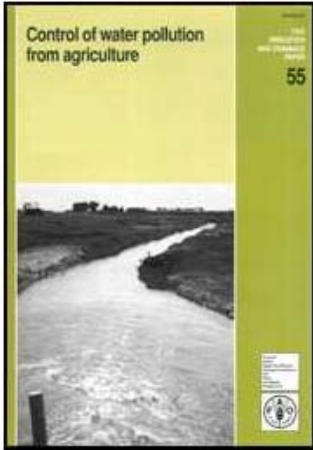
## View All Metadata

Data Object: **oai.nsd.org.GROW.70.html**  
Parent Collection: **/home/nsdl.sdsc/2003-06-10T13:53:37Z/oai:nsdl.org:GROW:70**  
Owner: **nsdl@sdsc**

**NSDL**  
THE NATIONAL SCIENCE DIGITAL LIBRARY

The National Science Digital Library's **Archived Version** is the snapshot we took of the page as we last checked its availability.

## Control of water pollution from agriculture - FAO irrigation and drainage paper 55



[Table of Contents](#)

Internet zone