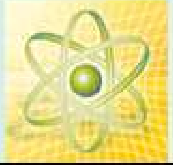
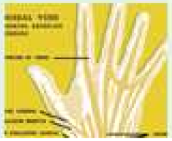


NSDL

THE NATIONAL SCIENCE DIGITAL LIBRARY

NSDL: OAI and a large-scale digital library

Carl Lagoze, Cornell University
NSDL Director of Technology
lagoze@cs.cornell.edu



What is the NSDL?

- NSF program to move science, math, engineering education in the US to digital age
  - http://www.ehr.nsf.gov/ehr/due/programs/nsdl/
- Over 80 independent grants exploring NSDL goals
  - http://comm.nsdlib.org
- Focused effort to develop and model infrastructure for science education on the web.
  - http://cinews.comm.nsdlib.org/cgi-bin/wiki.pl
- A production digital library
  - http://www.nsdl.org

Short History of the NSDL

1996	**Vision articulated by NSF's Division of Undergraduate Education**
1997	**National Research Council workshop**
1998	**Preliminary grants through Digital Libraries Initiative 2**
1998	**SMETE-Lib workshop**
1999	**NSDL Solicitation**
2000	**6 Core Integration demonstration projects + 23 others funded**
2001	**1 large Core Integration System project funded**
2002	**More than 80 independent projects funded**
2003	**Core Integration funding fixed until 2006**

NSF Grant Structure

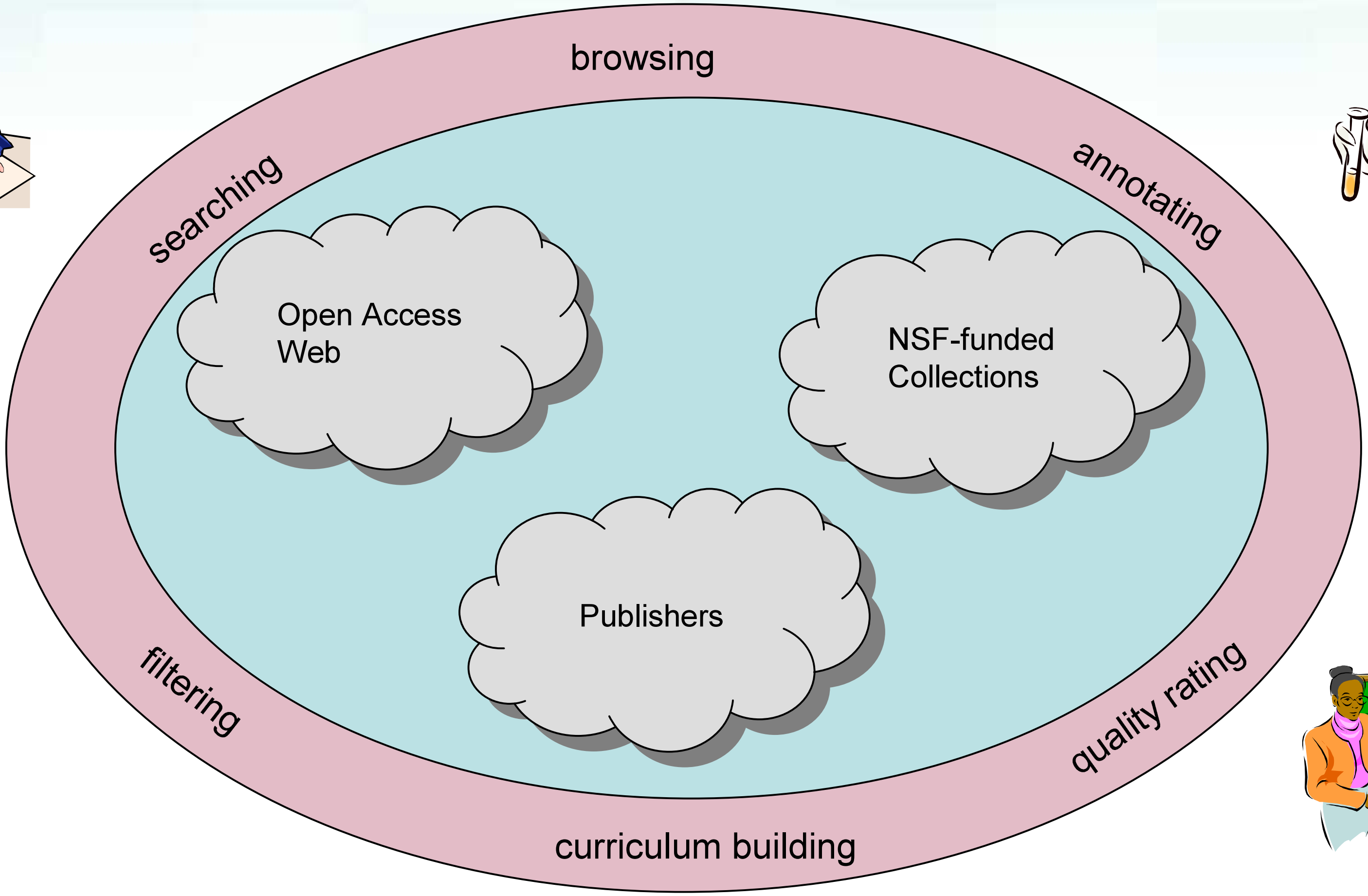
http://www.nsf.gov/pubs/2002/nsf02054/nsf02054.html

- Collections
  - Develop and maintain content
- Services
  - For users, collection providers, core integration
- Targeted research
- Core Integration
  - Organizational, economic, technical
  - $US5M of total $US25M total budget

NSDL CI Technical Organization

- A collaborative project
  - University Corporation for Atmospheric Research - **Dave Fulker**
  - Cornell University - **William Arms**
  - Columbia University - **Kate Wittenberg**
- With additional partners
  - Eastern Michigan University
  - Syracuse University
  - U Mass-Amherst
  - UC-Santa Barbara
  - UC-San Diego (Supercomputer Center)
- Director of Technology - **Carl Lagoze**

Building service and knowledge layers over a variety of resources for a variety of users



browsing

searching

annotating

Open Access Web

NSF-funded Collections

Publishers

filtering

quality rating

curriculum building

How Big might the NSDL be?

All branches of science, all levels of education, very broadly defined:

Five year targets

1,000,000 different users

10,000,000 digital objects

10,000 to 100,000 independent sites

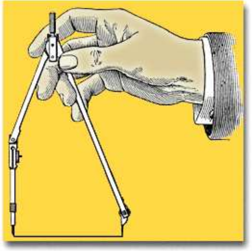
Core Integration Philosophy

It is possible to build a **very large** digital library with a small staff.

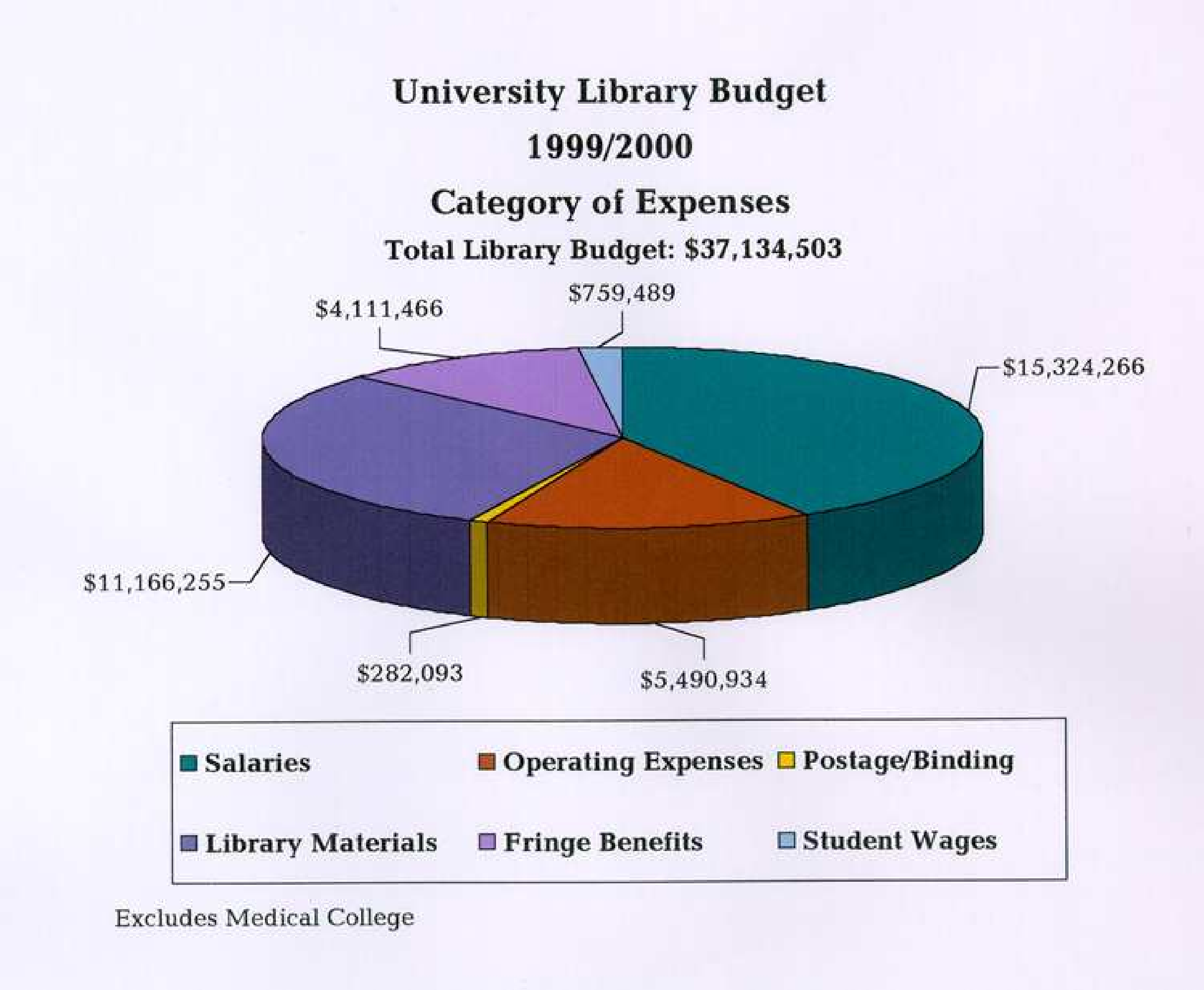
But ...

Every aspect of the library must be planned with scalability in mind.

Some compromises will be made.



Perspective on the Budget



University Library Budget

1999/2000

Category of Expenses

Total Library Budget: $37,134,503

Excludes Medical College

Resources for Core Integration

	Core Integration
Budget	$4-6 million
Staff	25 - 30
Management	Diffuse

How can a small team, without direct management control, create a very large-scale digital library?

NSDL technical mantras

- **Aggregation rather than collection**
  - Core integration team will not manage any collections
- **Spectrum of interoperability**
  - Accommodate diversity of participation models
  - Open interfaces and standards permitting plug in of array of value-added services
- **One library many portals**
  - Accommodate multiple quality and selection metrics
  - Tailor presentation of content and nature of services to audience needs
- **Open toolkit of software and services for library building**

Spectrum of interoperability

Level	*Agreements*	*Example*
Federation	Strict use of standards (syntax, semantic, and business)	AACR, MARC Z 39.50
Harvesting	Digital libraries expose metadata; simple protocol and registry	Open Archives metadata harvesting
Gathering	Digital libraries do not cooperate; services must seek out information	Web crawlers and search engines

Translating to first release goals

- **This is a big task that no one has done before!**
- **Work on the priorities**
  - Focus on one point on spectrum of interoperability
    - Metadata harvesting
    - Incorporate NSF funded collections and selected other collections
  - Leverage existing (or at least emerging) technologies and protocols
    - OAI, uPortal, Shibboleth, SDLIP, InQuery
  - Provide reliable base level services
    - Search and Discovery, Access Management, User Profiles, Exemplary Portals, Persistence
- **Plant some seeds for the future**
  - Machine-assisted metadata generation
  - Automated collection aggregation
  - Web gathering strategies

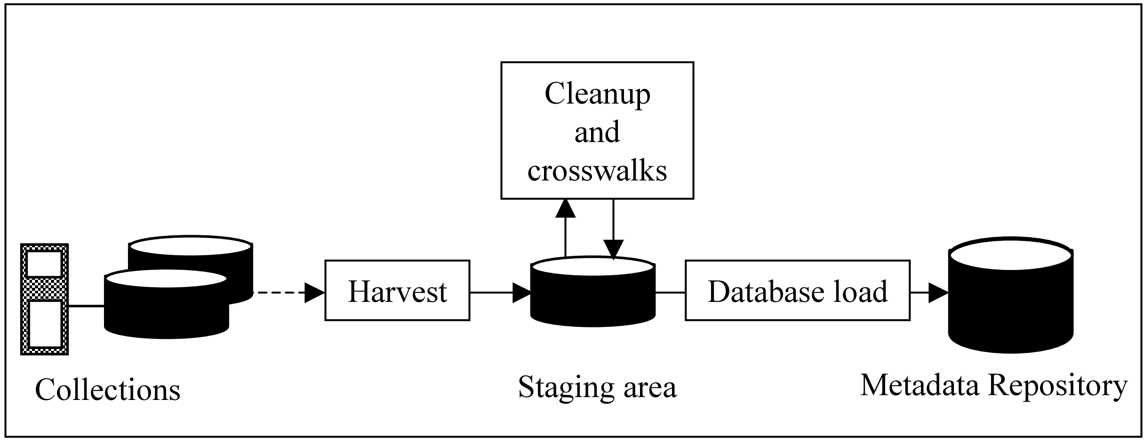
Metadata Repository

- **Central storage of all metadata about all resources in the NSDL**
  - Defines the extent of NSDL collection
  - Metadata includes collections, items, annotations, etc.
- MR main functions
  - Aggregation
  - Normalization
  - redistribution
- **Ingest of metadata by various means**
  - Harvesting, manual, automatic, cross-walking
- **Open access to MR contents for service builders via OAI-PMH**

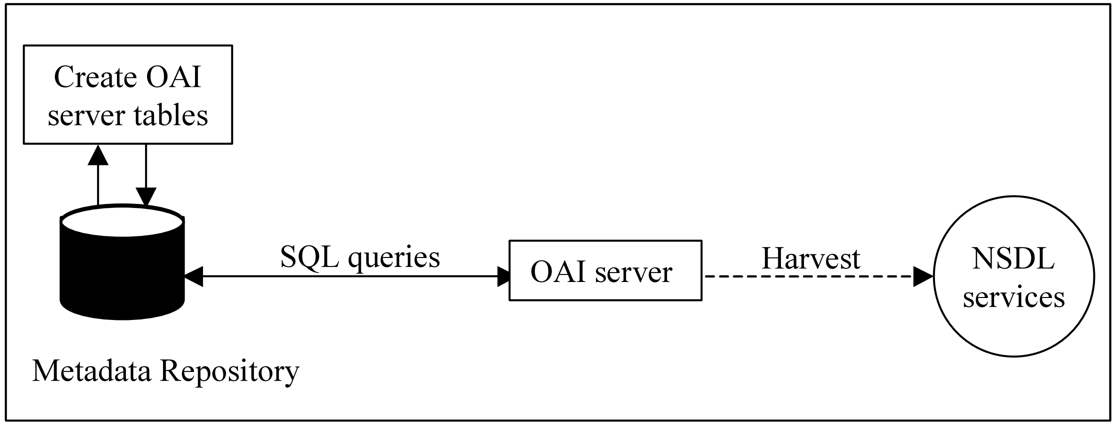
Metadata Strategy

- Collect and redistribute any native (XML) metadata format
- Provide crosswalks to Dublin Core from eight standard formats
  - Dublin Core, DC-GEM, LTSC (IMS), ADL (SCORM), MARC, FGCD, EAD
- Concentrate on collection-level metadata
- Use automatic generation to augment item-level metadata

Importing metadata into the MR



Exporting metadata from the MR



Simple Metadata-Based Services:

The recognition of common elements among a set of core Library services (initially Exhibits News, Annotation, Equivalence, and My Site), led the NSDL Team to create a model for the development and implementation of services that could be based on simple extensions to standard Metadata Records. Services that fit this model are known as Simple Metadata-Based Services, or SiMBaS.

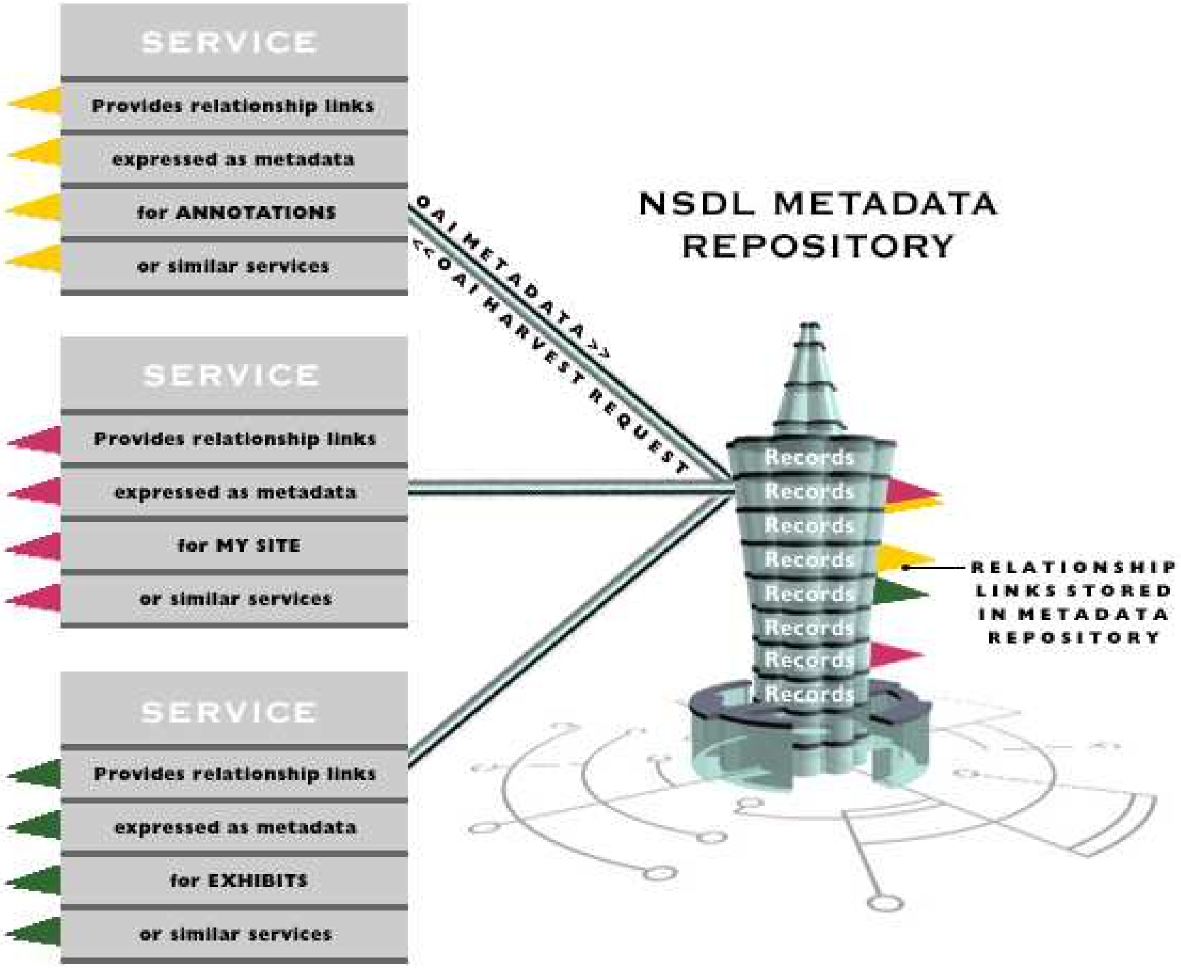
SIMBaS Characteristics

- Services provide metadata records for harvesting by MR
- Metadata records may include typed relationship links to each other or to pre-existing Metadata Records in the MR.
- Example relationship links
  - Collections->items.
  - Annotation metadata record->item-level metadata record.

A Model for Simple Metadata-Based Services

NSDL services that will utilize Simbas:

- **Annotations**
  (Review Service)
- **My Site**
  (Personal Content Creation Component)
- **Exhibits**
- News
- Equivalence



SERVICE

Provides relationship links

expressed as metadata

for ANNOTATIONS

or similar services

SERVICE

Provides relationship links

expressed as metadata

for MY SITE

or similar services

SERVICE

Provides relationship links

expressed as metadata

for EXHIBITS

or similar services

OAI METADATA>>

<<OAI HARVEST REQUEST

NSDL METADATA REPOSITORY

Records

Records

Records

Records

Records

Records

Records

Records

RELATIONSHIP LINKS STORED IN METADATA REPOSITORY

Searching

What to Index?

When possible, full text indexing is excellent, but full text indexing is not possible for all materials (non-textual, no access for indexing).

Comprehensive metadata is an alternative, but available for very few of the materials.

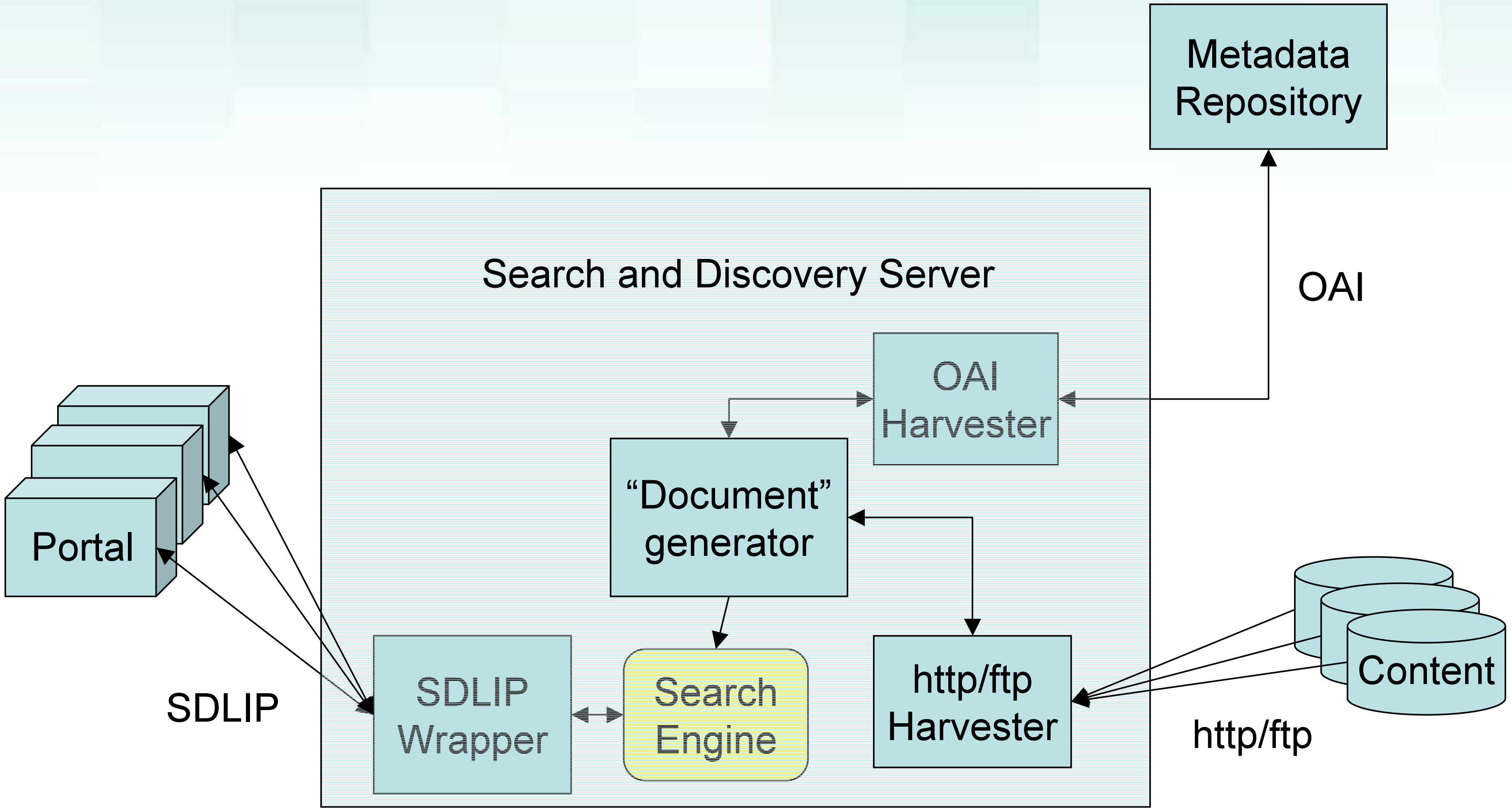
What Architecture to Use?

Few collections support an established search protocol (e.g., Z39.50)

Search system general features

- Implement a query language that includes most features that are common in commercial and Web search engines.
- Periodically harvest the MR (via OAI-PMH) to incorporate the latest changes in the library.
- Allow search on resources’ metadata as well as textual content, when available.
- Communication with portals is done via the Simple Digital Library Interoperability Protocol (SDLIP).

Search Architecture



Persistent Archive for the NSDL

- Provide a persistent copy of the resources identified in the NSDL repository
  - Provide a mechanism to retrieve prior versions of resources
- Verify availability of on-line digital resources that have presence in MR

Persistent Archive Approach

- Use data grid technology to:
  - Implement a persistent logical name space for registering resources
  - Manage archiving of modules on distributed storage systems
- Use OAI harvesting to extract metadata from the NSDL repository
- Crawl the web to retrieve resources
- Provide OAI interface for reporting validation results
- Manage the persistent archive through a separate information repository

Experience thus far

- OAI – low barrier?
  - Sets
  - Identifiers
- XML flakiness
- Limitations of basic Dublin Core
- Metadata quality and trust
- Resource granularity

Closing Thoughts

- We have only just begun!
- Automation is key to scalability
  - Metadata generation
  - Longevity/preservation
  - Quality and selection
  - Collection development
- The NSDL needs to be more that *data*
  - Knowledge
  - Curricula
  - Community
  - collaboration