1ˢᵗ Canadian ETD & Open Repositories Workshop

May 10-11, 2010   Carleton University, Ottawa

"Opening and Expanding Digital Library Services"

by Edward A. Fox

- fox@vt.edu   http://fox.cs.vt.edu
- Dept. of Computer Science, Virginia Tech
- Blacksburg, VA 24061 USA

Acknowledgements

- Mentors (Licklider, Kessler, Salton)
- Virginia Tech, CS, Digital Library Research Laboratory
- NSF and other sponsors
- Students, colleagues, co-investigators
- Monika Akbar, Yinlin Chen, Spencer Lee, Venkat Srinivasan, Seungwon Yang, …
- Boots Cassel, Gary Marchionini, Jeffrey Pomerantz, Barbara Wildemuth, Andrea Kavanaugh, Naren Ramakrishnan, Steve Sheetz, Don Shoemaker, …

Part 1 – Selected DL Projects

- Digital Library Curricular Resources
  - NSF IIS-0535057 & 0535060
- CTRnet (Crisis, Tragedy & Recovery Net)
  - NSF IIS-0916733
- Ensemble (Computer Science Education)
  - NSF DUE-0840719
- Digital Preserve
  - NSF IIS-0910183 & 0910465
  - http://slurl.com/secondlife/Digital%20Preserve /140/126/29

DL Curric. Project - 1

- NSF awards to VT and UNC-CH
- CS and LIS

- Project server: http://curric.dlib.vt.edu/

- Wikiversity: http://en.wikiversity.org/wiki/Curriculum_on _Digital_Libraries

DL Curric. Project - 2

- **Module 1-b: History of digital libraries and library automation**
- **Module 2-c: File Formats, Transformation, and Migration**
- **Module 3-b: Digitization**
- **Module 4-b: Metadata**
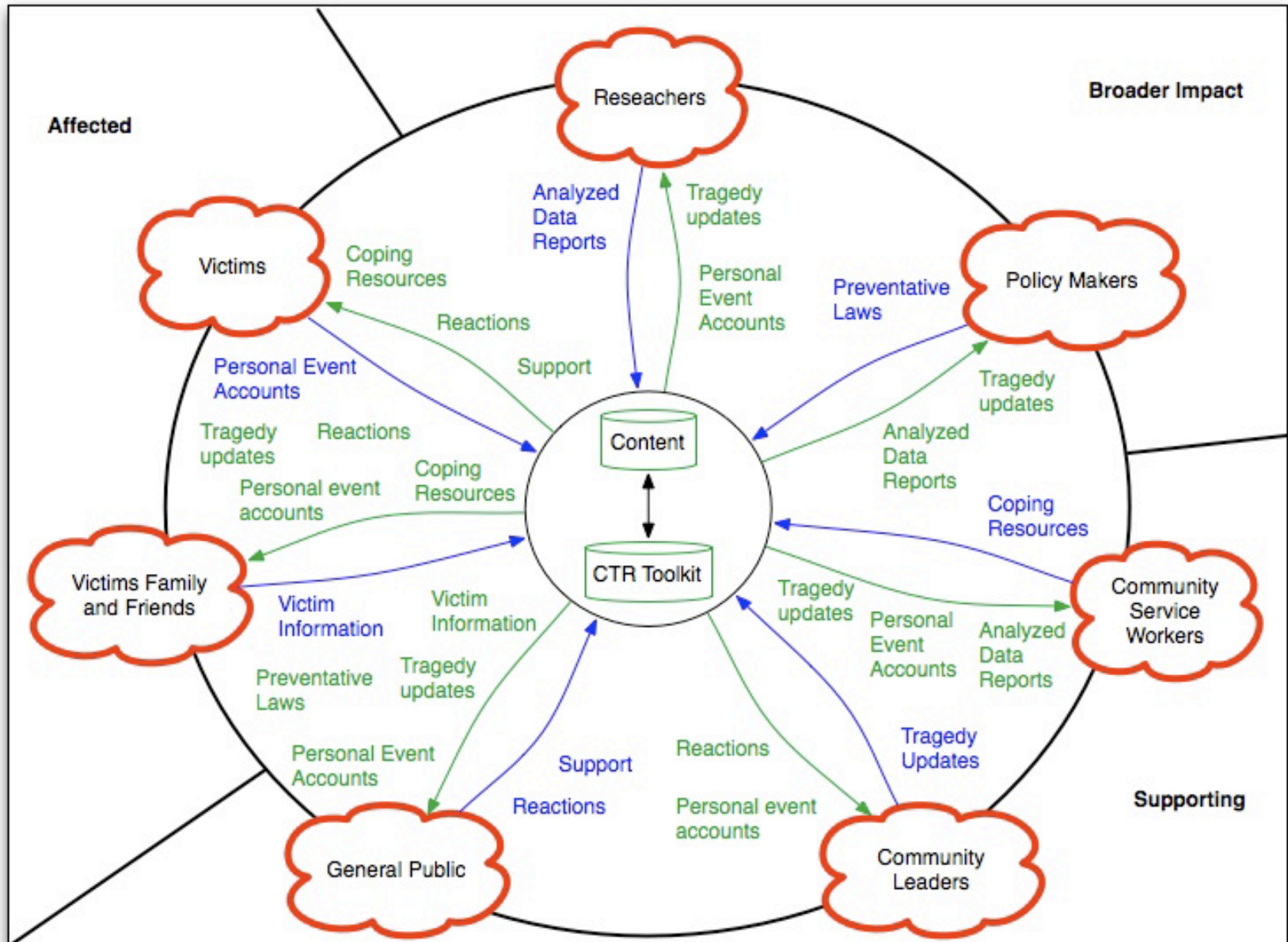- **Module 5-a: Architecture overviews**

DL Curric. Project - 2

- **Module 5-b: Application software**
- **Module 5-d: Protocols**
- **Module 6-a: Information needs/relevance**
- **Module 6-b: Online information seeking behaviors and search strategies**
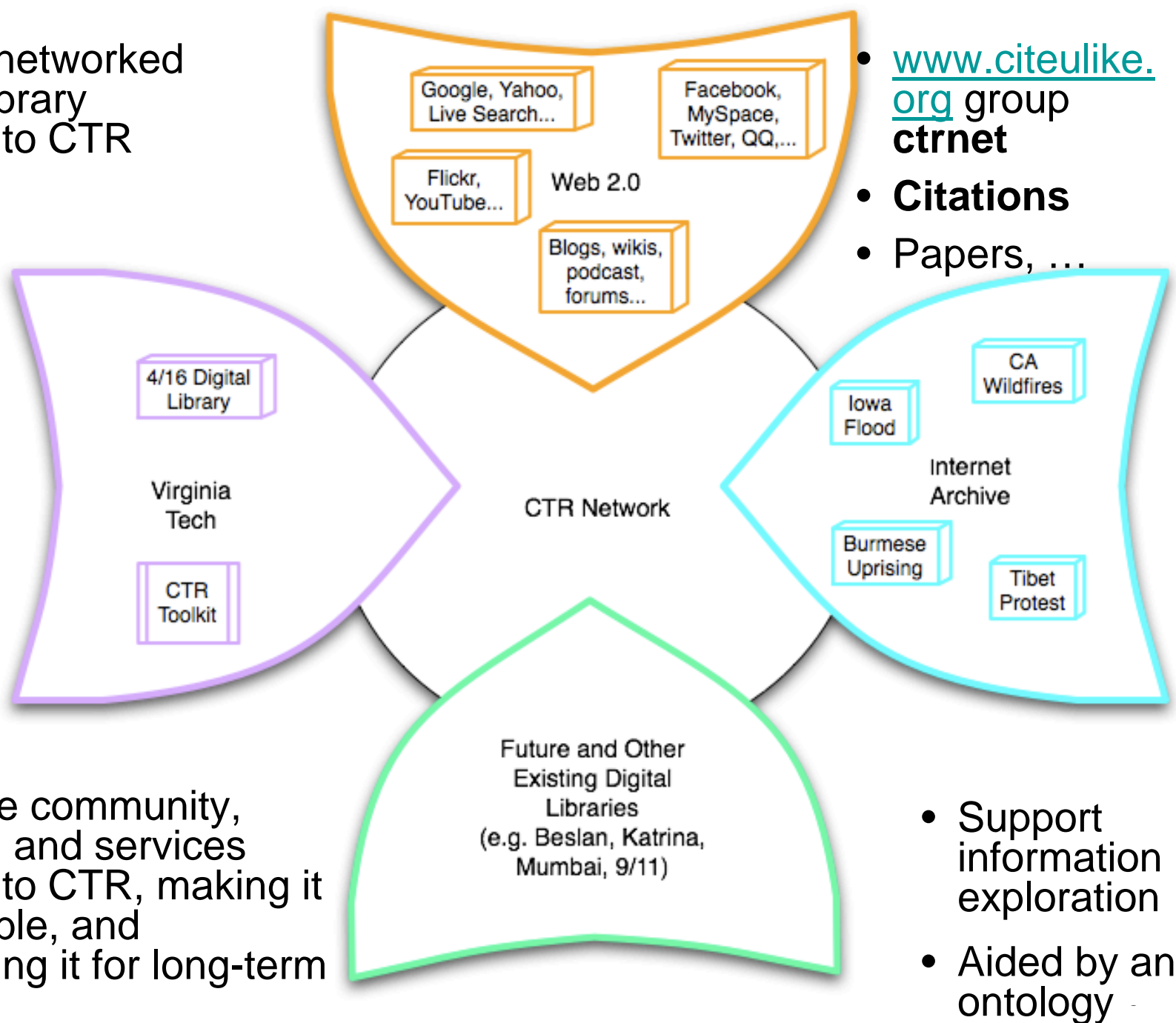- **Module 6-d: Interaction design and usability assessment**

DL Curric. Project - 3

- **Module 7-b: Reference Services**
- **Module 7-g: Personalization**
- **Module 8-b: Web Archiving**
- **Module 9-c: Digital library evaluation, user studies**

CTR stakeholders

- Build a networked digital library relating to CTR

Google, Yahoo, Live Search...

Facebook, MySpace, Twitter, QQ,...

Flickr, YouTube...

Web 2.0

Blogs, wikis, podcast, forums...

- www.citeulike.org group **ctrnet**
- **Citations**
- Papers, …

4/16 Digital Library

Virginia Tech

CTR Toolkit

CTR Network

Iowa Flood

CA Wildfires

Internet Archive

Burmese Uprising

Tibet Protest

Future and Other Existing Digital Libraries (e.g. Beslan, Katrina, Mumbai, 9/11)

- Integrate community, content, and services relating to CTR, making it accessible, and preserving it for long-term reuse

- Support information exploration

- Aided by an ontology

Haiti Photographs, Content Based Image Retrieval Evaluation
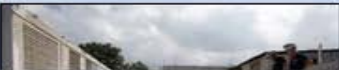
Goals for Ontology for CTR
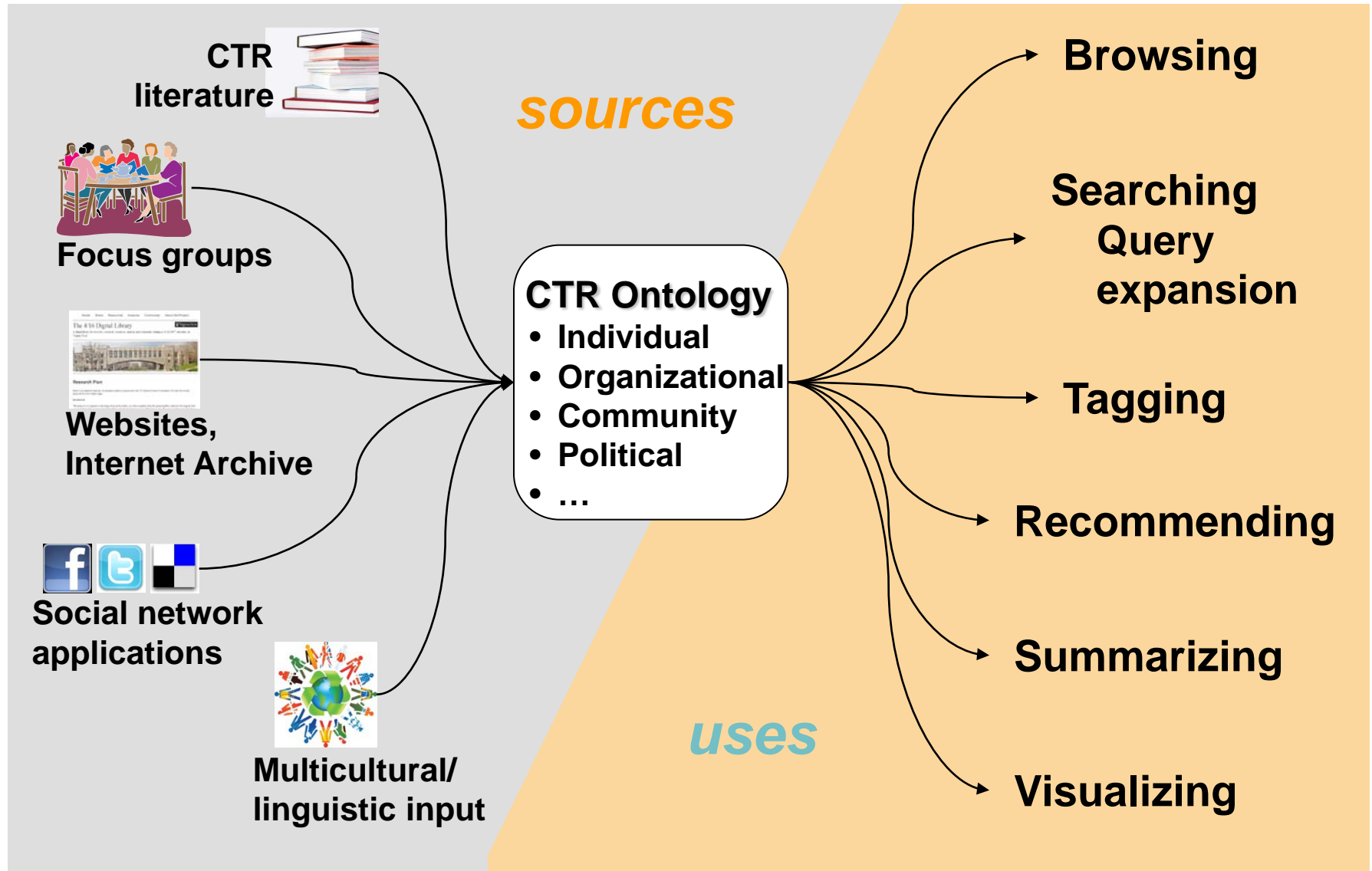
Preliminary Data Analysis

Revise seeds if poor preliminary data

Collect Seeds	Crawl	Index Data	Pass Along

- Index crawl data from Heritrix
- Use NutchWax to preliminarily analyze seed quality
- Send ARC files on for Story-telling
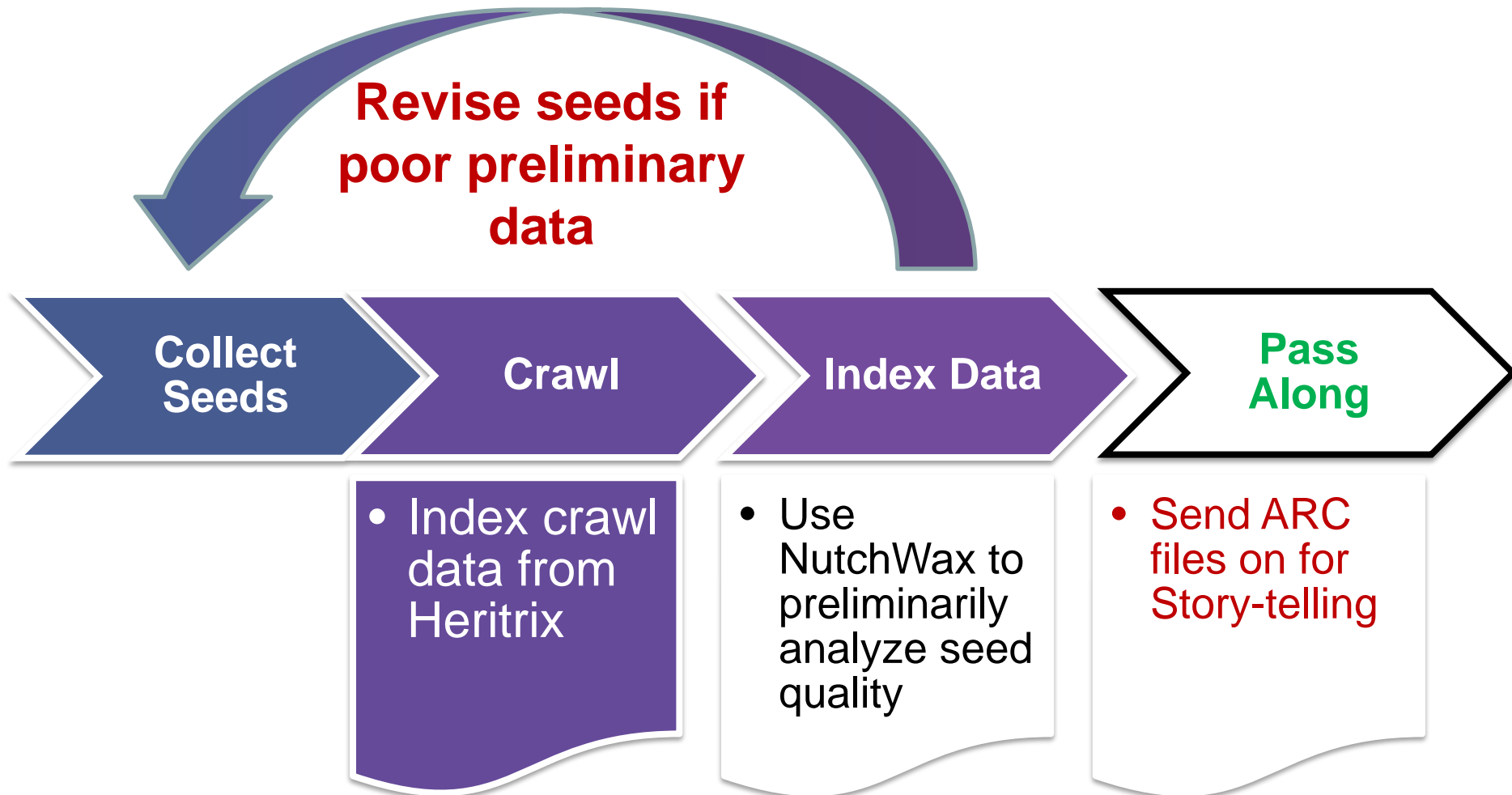
Data Filtering and Storytelling

Crawling

Preprocessing

- Extracting Text
- Basic Text Cleanup

Classification

- Supervised learning methods
- Evaluation
- Classifying new data

Storytelling

- Generating stories
- Visualization
- Story analysis

Computing Communities

WebCAT TECH

FOCES CATSpace

CS1

Social network services

twitter

facebook.

Ensemble Portal

Drupal

Blog Browse Search

Forum Submit RSS

Tools

Walden's Path

VKB SI

Fedora

Storage

Computing Resources

Syllabus

SWENET AlgoViz

CSTC

CITIDEL

CSTA

SECOND LIFE

Ensemble in Second Life

http://slurl.com/secondlife/Educators%20Coop%204/66/236/28
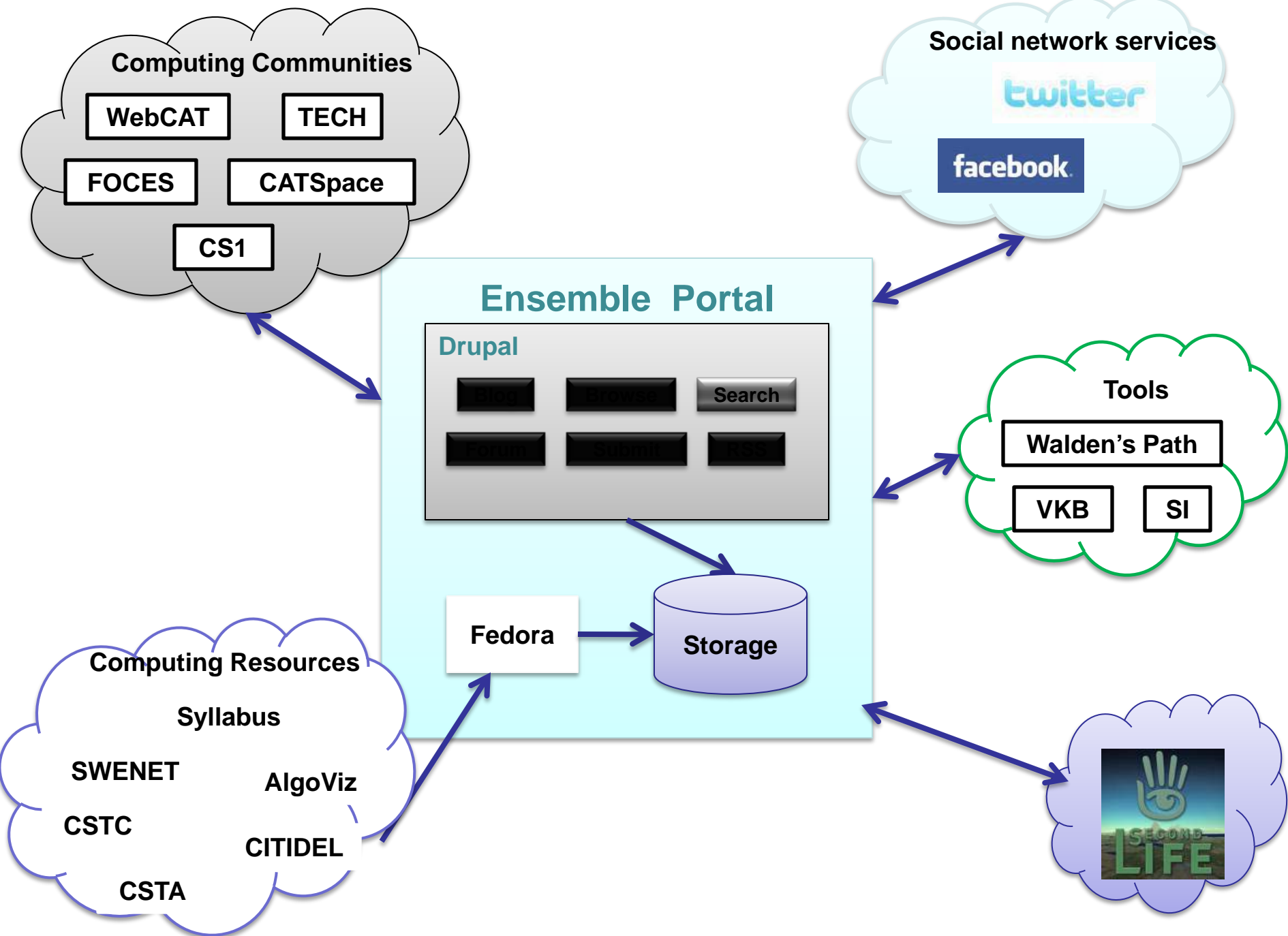
The Ensemble Pavilion offers:
• teleports to other computing sites in Second Life like the *Digital Preserve*
• hyperlinks to related computing websites
• RSS readers with feeds from computing and computing education blogs
• membership in the Ensemble Computing group in Second Life, Facebook, and Twitter

www.computingportal.org

Selected Digital Preserve Personnel
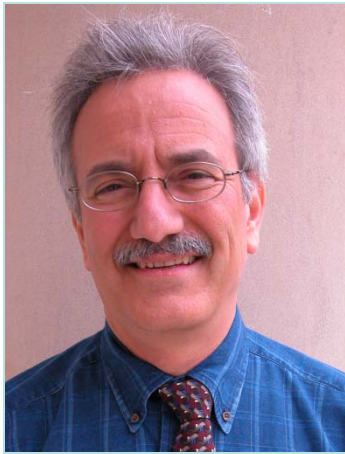
Gary Octagon
Gary Marchionini

EdFox Rieko
Edward Fox

zamfir Paule
Spencer Lee

mantruc Martian
Javier Velasco-Martin

Uma Aldrin
Uma Murthy

Krad Proto
Seungwon Yang

DP areas

Poster Building

- 18 posters on display
- Poster view tips
- Video screen

Cafe

- Beverages
- Screens
- Discussion areas

Part 2 – Basic DL Concepts

- Digital Library Scope
- OAI
  - Harvesting
  - Repositories
- Space-related Perspectives of Computing
  - Distributed
  - Cloud …
- 5S

DL Scope

- Institutional repositories
- Open archives
- Electronic/virtual libraries
- Content management systems
- Courseware management systems
- Personal information management systems
- Cloud/ubiquitous/… computing

Synchronous Scholarly Communication
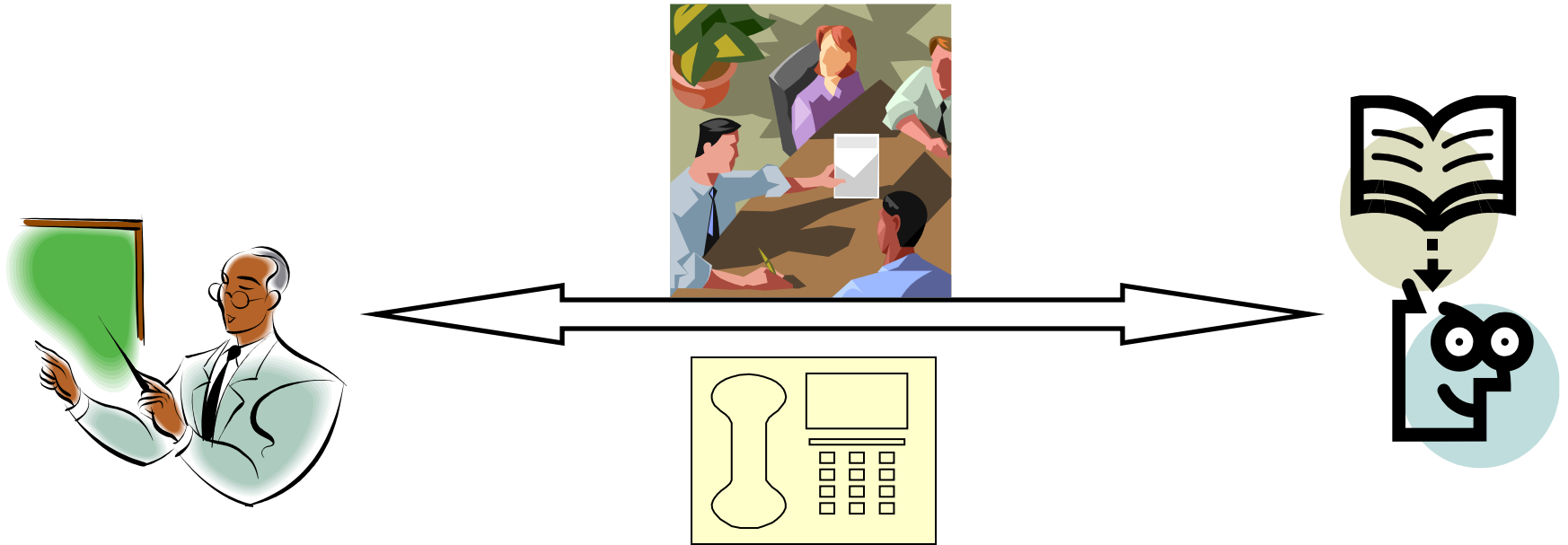
Same time, Same or different place

Asynchronous, Digital Library Mediated Scholarly Communication
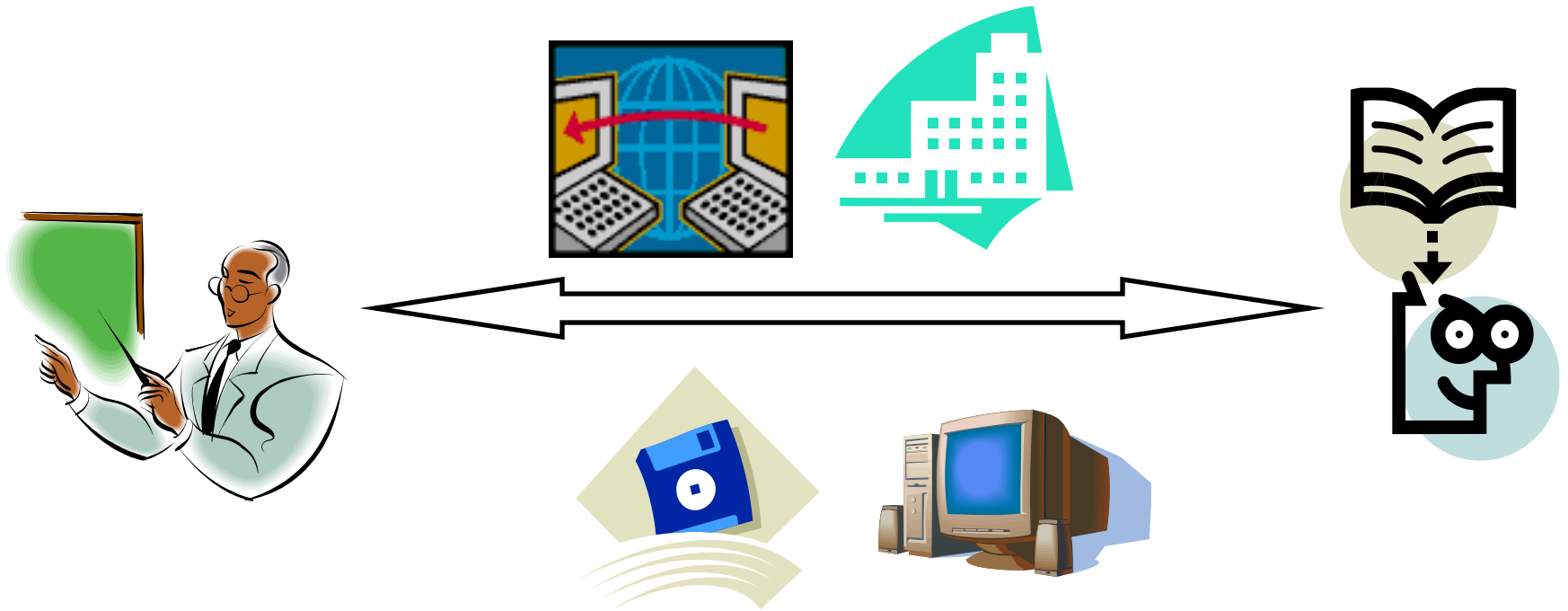
Different time and/or place

Information Life Cycle

Creation

Active

Social Context

Authoring
Modifying

Using
Creating

Organizing
Indexing

Retention
/ Mining

Accessing
Filtering

Storing
Retrieving

Semi-
Active

Utilization

Distributing
Networking

Inactive

Searching
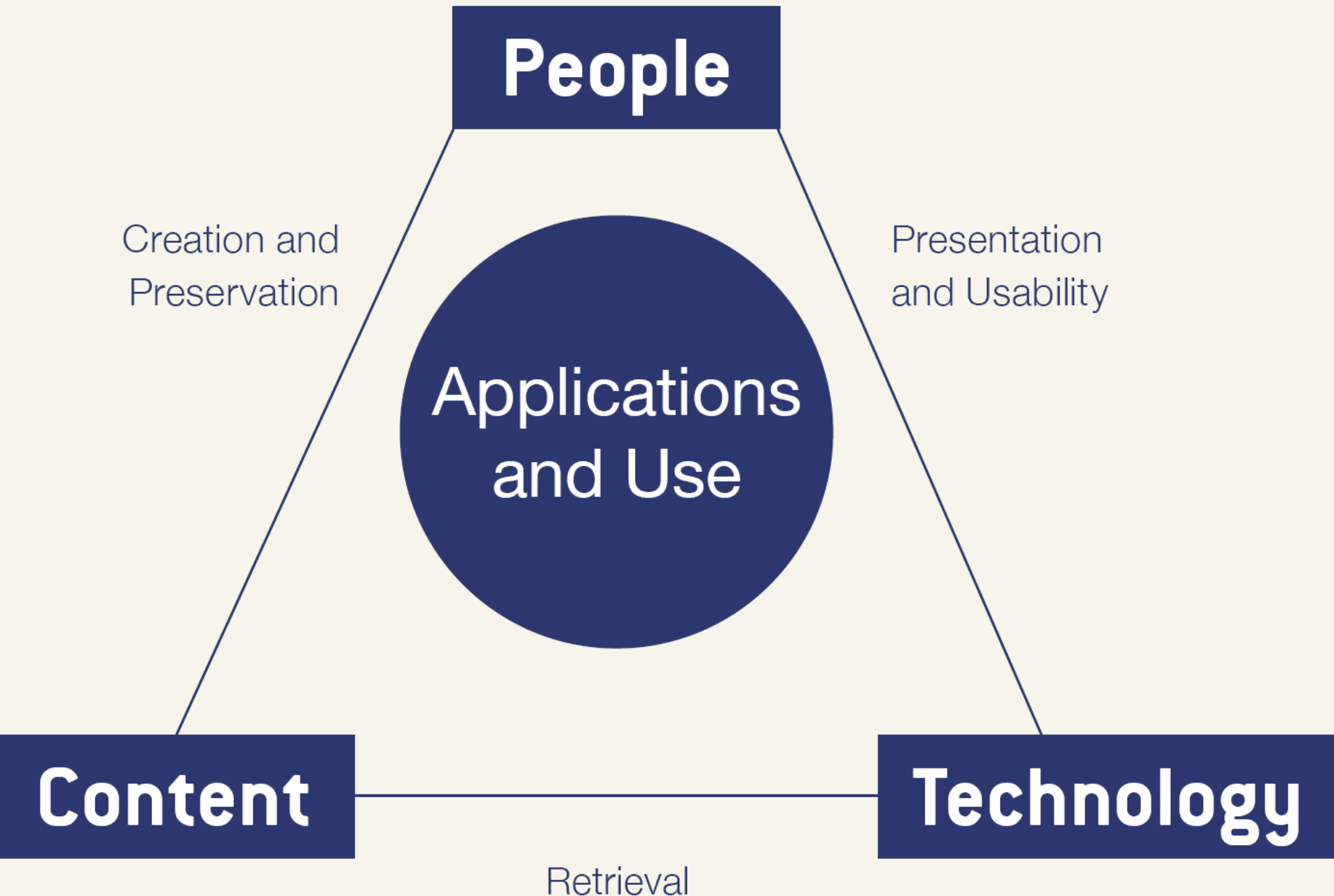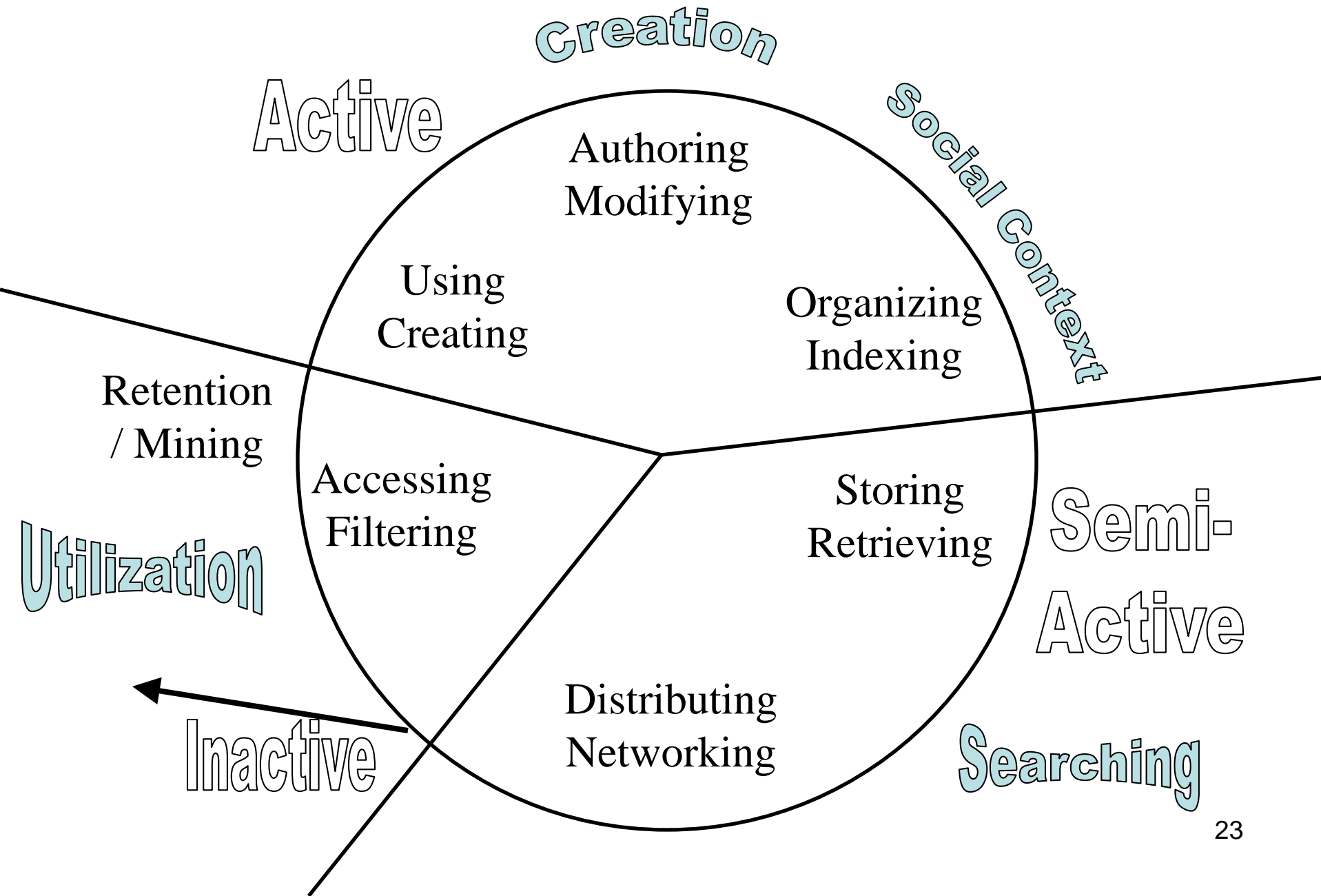
Quality and the Information Life Cycle

DLs Shorten the Chain to

Roles

Author — Teacher — User — Digital Library

Reader — Editor — Learner — Reviewer — Librarian
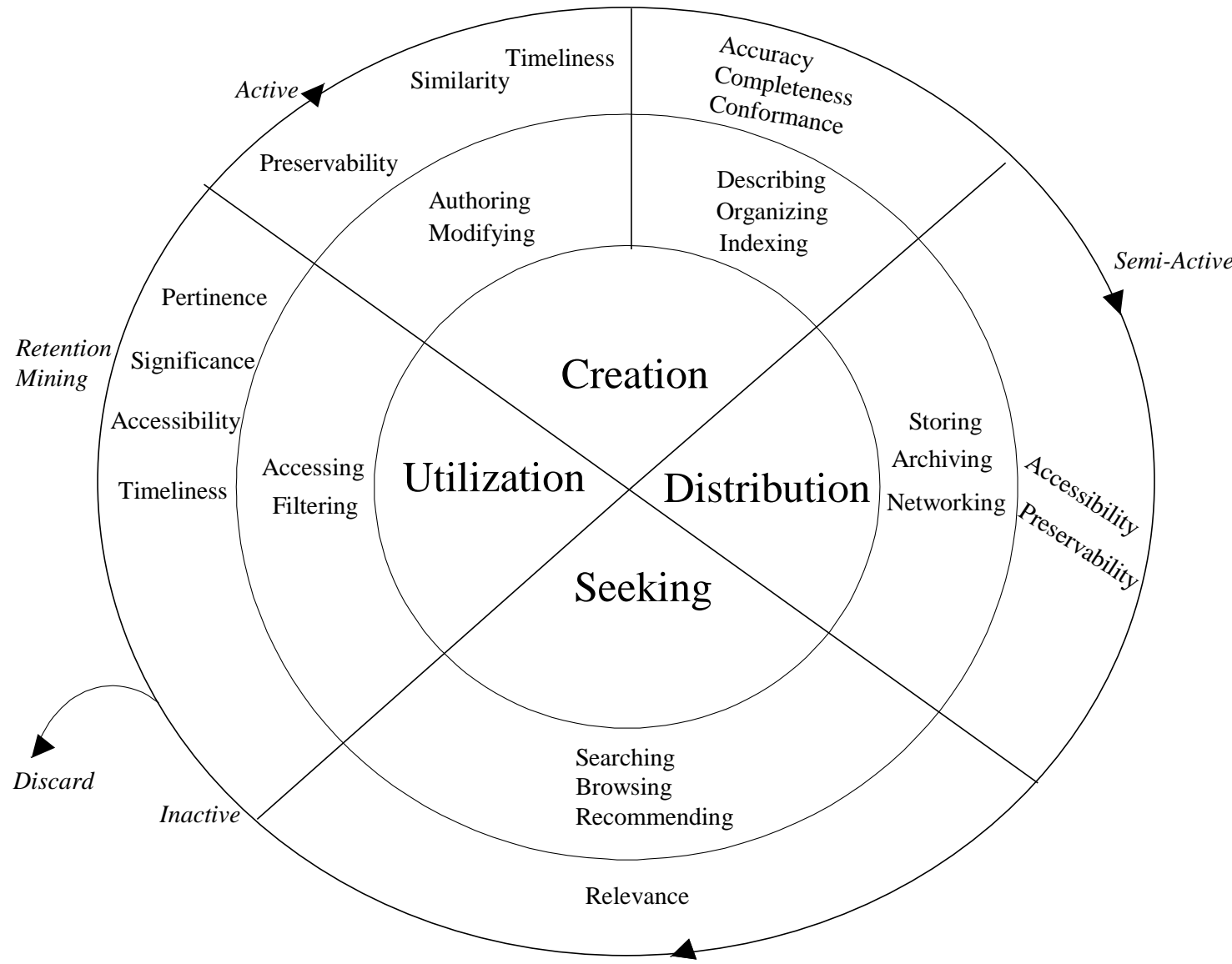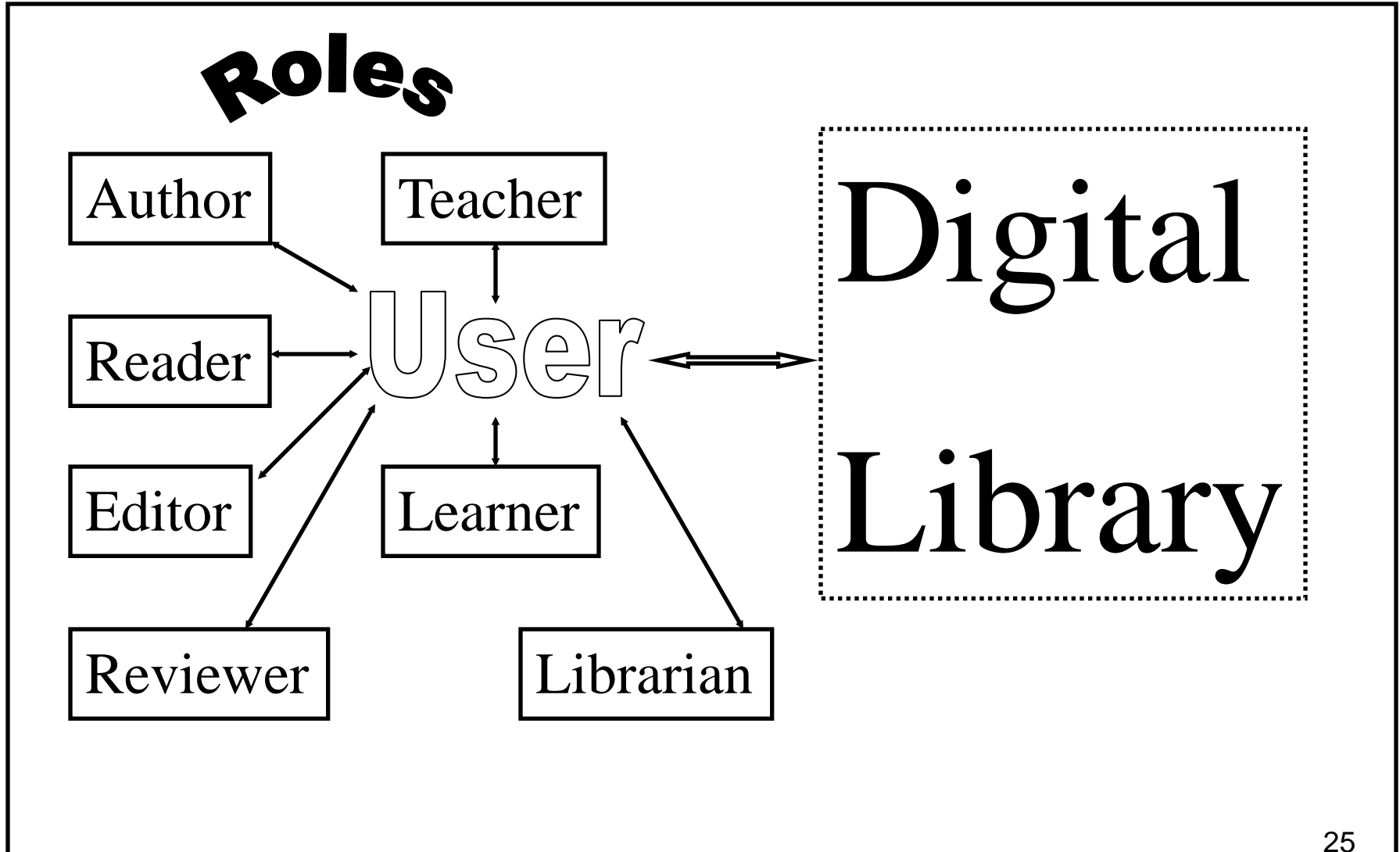
Degree of Structure

Web   DLs   DBs

$\longleftrightarrow$

Chaotic   Organized   Structured

Example of Granularity of Information Structure

Word level

Phrase level

Sentence level

Passage level

Document level
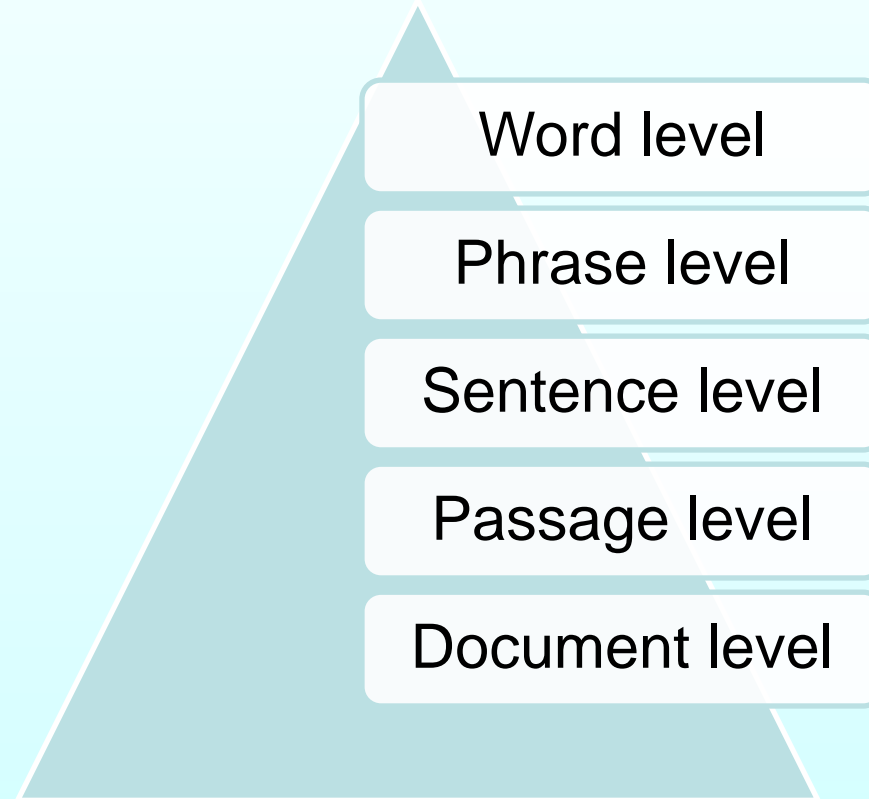
**Example of Structural Level
of Text Information**

ETD Logical Hierarchy
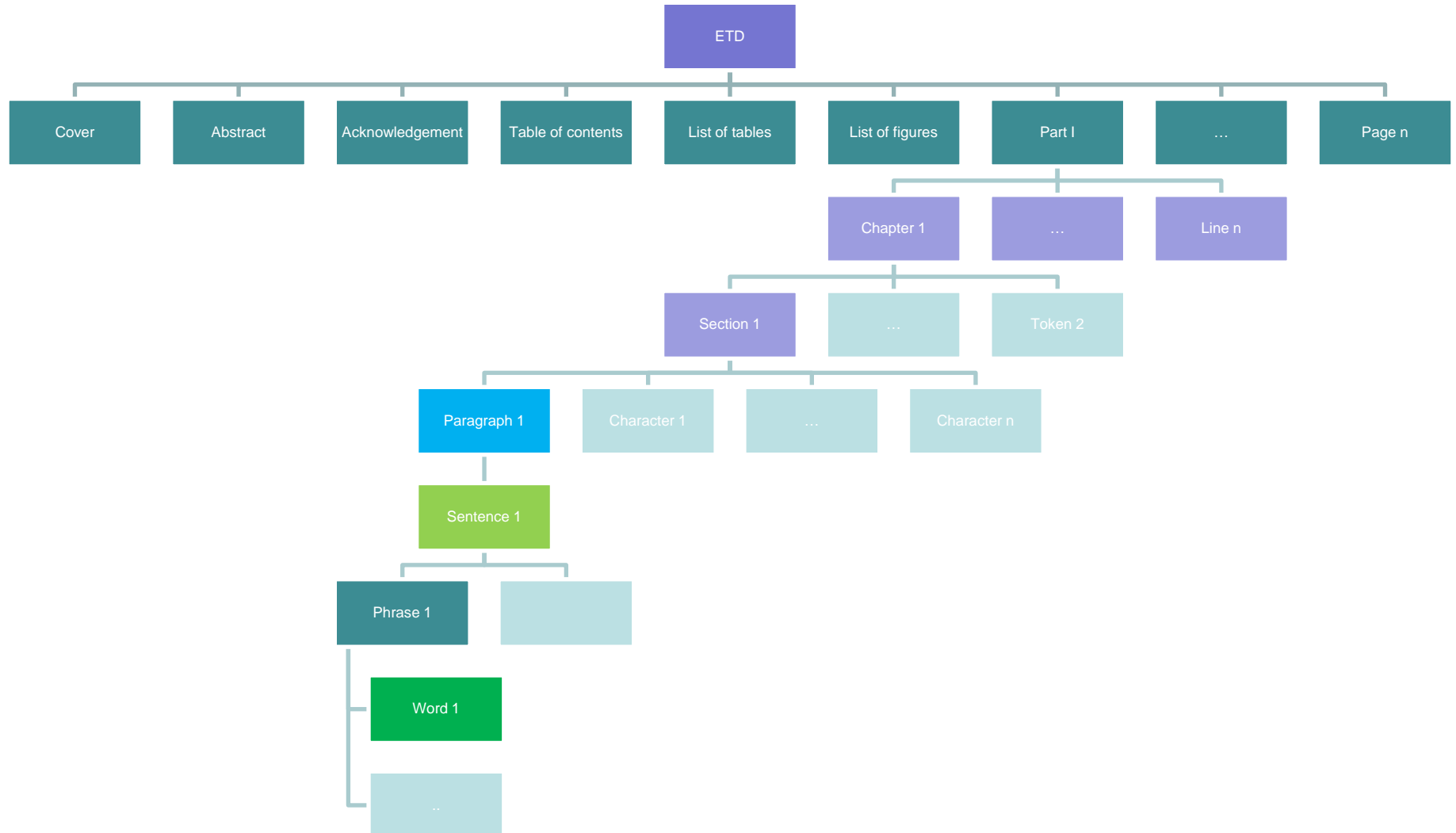
OAI = Technical Umbrella for Practical Interoperability…

Metadata Harvesting

Museums

Reference Libraries

Publishers

E-Print Archives
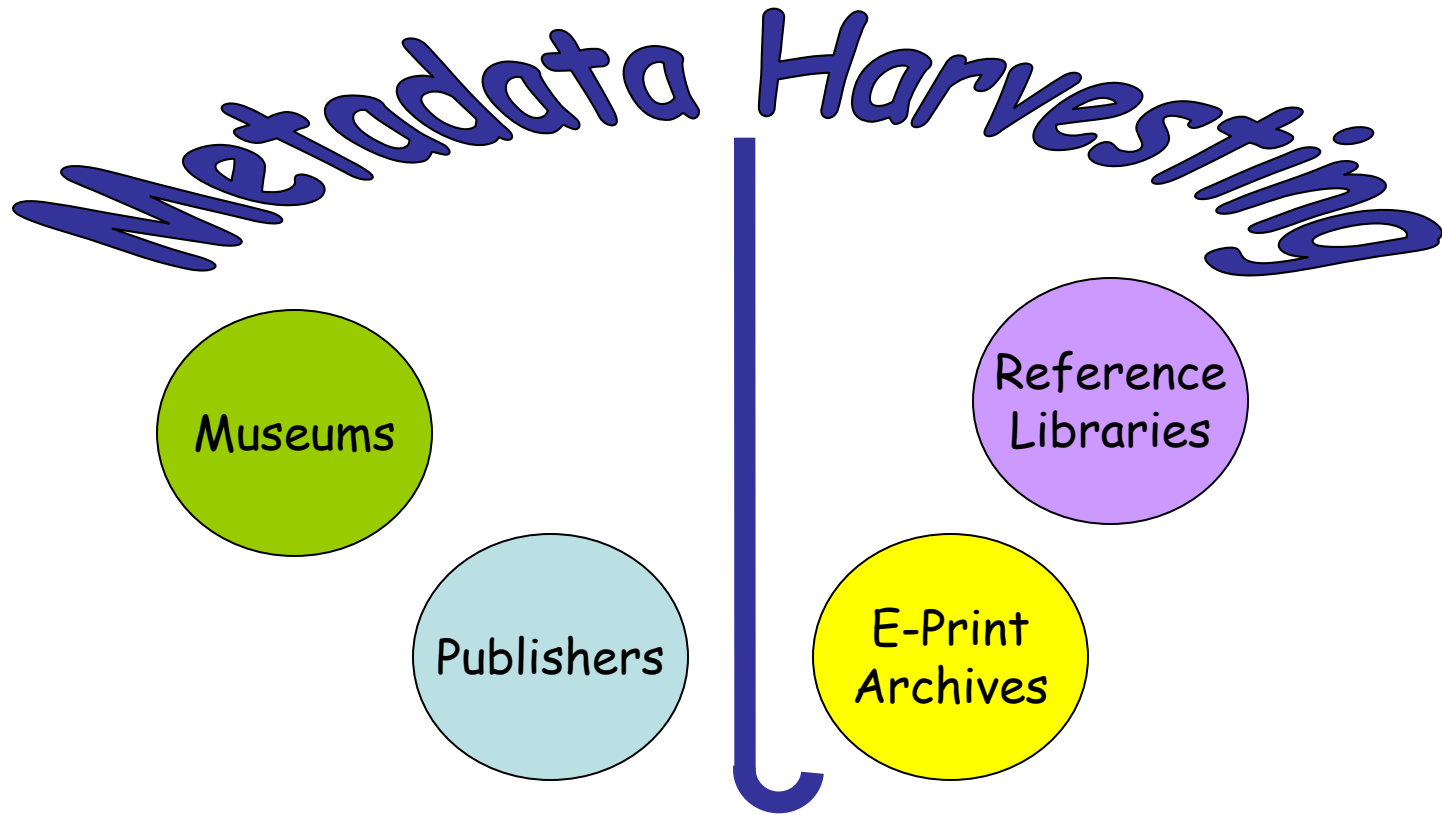
…that can be exploited by different communities

OAI – Repository Perspective

Glossary:
DC=Dublin Core
MDO=Metadata Object
DO=Digital Object

Required: Protocol

Set Structure

URI Scheme

Required: DC

| MDO | MDO | MDO | MDO |
| MDO | MDO | MDO | MDO |

| DO | DO | DO | DO |
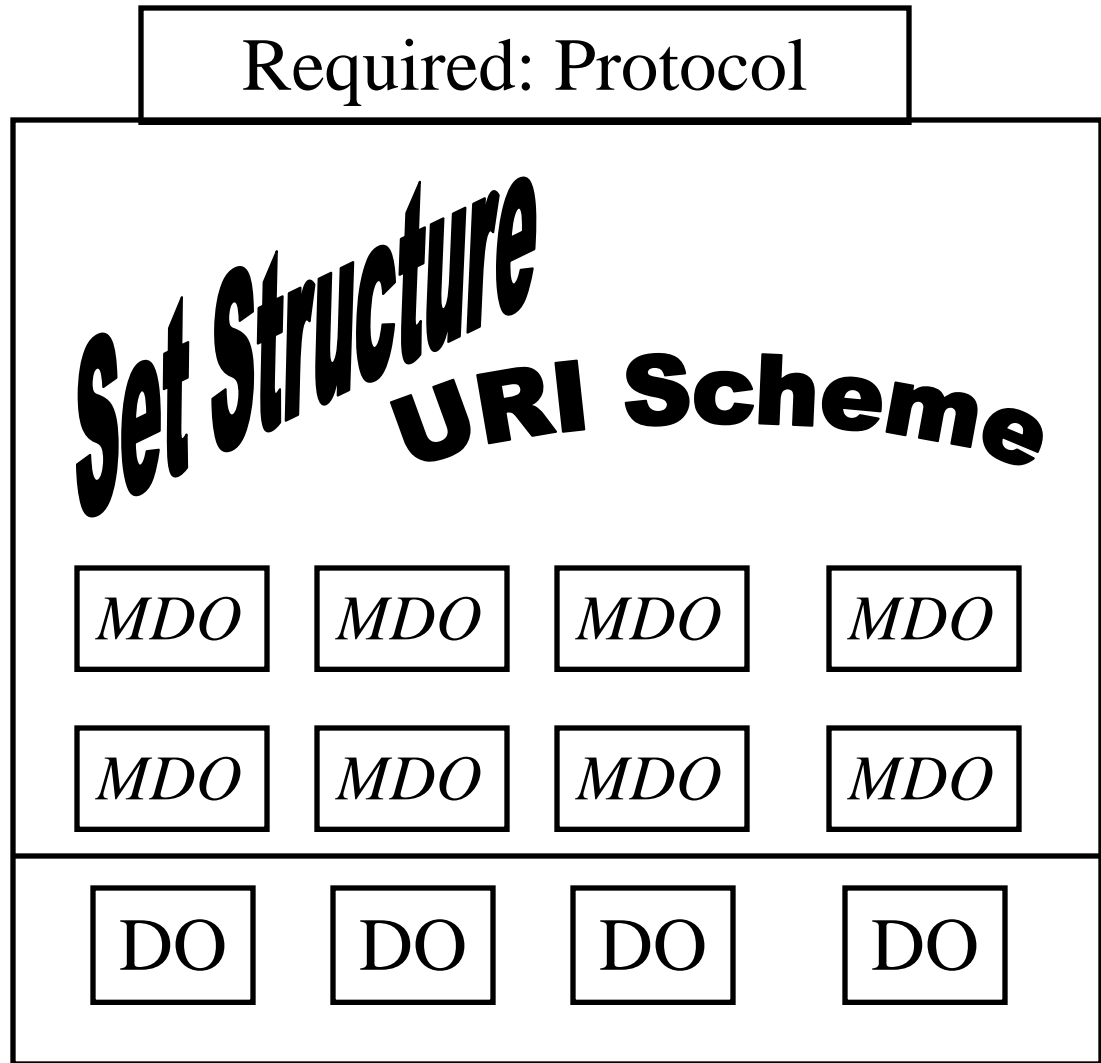
The World According to OAI

Service Providers

Discovery | Current Awareness | Preservation

Metadata harvesting

Data Providers

31
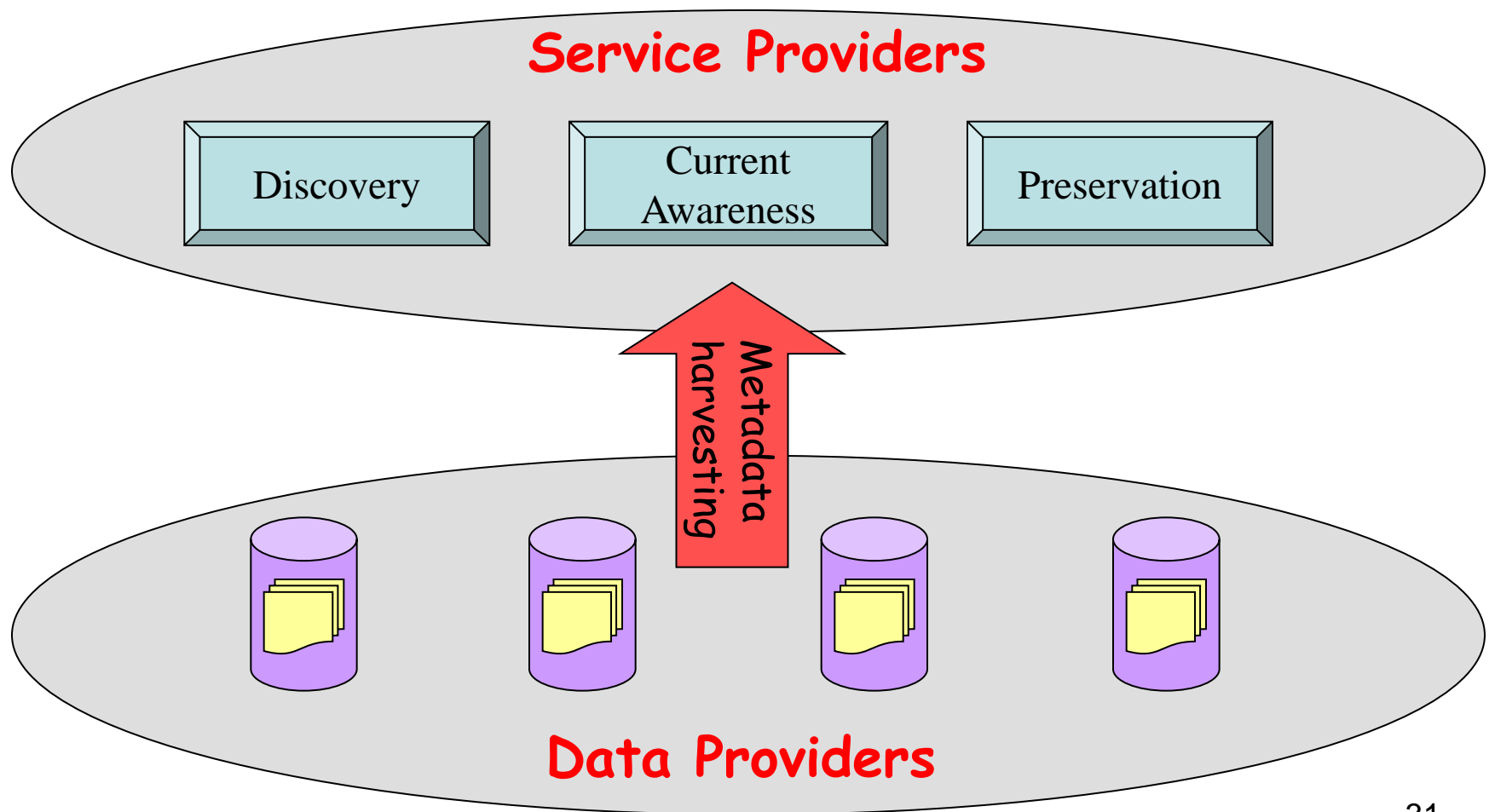
Space-related Computing
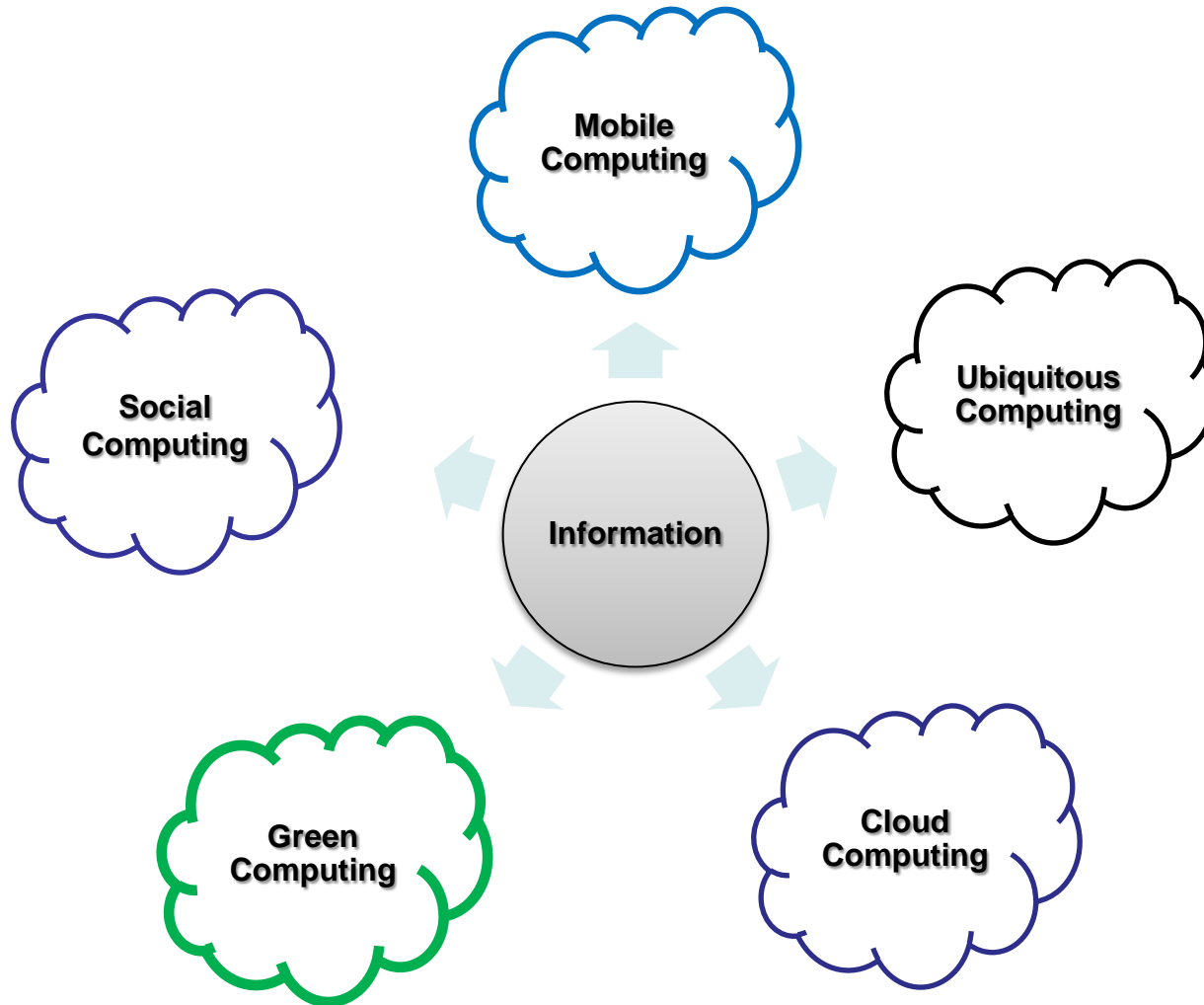
5S Layers

Societies

Scenarios

Spaces

Structures

Streams

5Ss

Ss	Examples	Objectives
Streams	Text; video; audio; image	Describes properties of the DL content such as encoding and language for textual material or particular forms of multimedia data
Structures	Collection; catalog; hypertext; document; metadata	Specifies organizational aspects of the DL content
Spaces	Measure; measurable, topological, vector, probabilistic	Defines logical and presentational views of several DL components
Scenarios	Searching, browsing, recommending	Details the behavior of DL services
Societies	Service managers, learners, teachers, etc.	Defines managers, responsible for running DL services; actors, that use those services; and relationships among them

34

5S Contextualized

- <u>Societies</u>/communities/users served
- <u>Scenarios</u>/services supported
- Management of physical/conceptual/ feature <u>spaces</u>
- Use of <u>structures</u>/organizational devices
- <u>Streams</u> of content and communication

5S and DL formal definitions and compositions (April 2004 TOIS)

relation (d. 1)

measurable(d.12), measure(d.13), probability (d.14), vector (d.15), topological (d.16) spaces

language (d.5)

sequence (d. 3)   graph (d. 6)

function (d. 2)

sequence (d. 3)

tuple (d. 4)*

state (d. 18)

event (d.10)

5S

grammar (d. 7)

streams **(d.9)**   structures **(d.10)**   spaces **(d.18)**   scenarios **(d.21)**   societies (d. 24)

structured stream (d.29)

services **(d.22)**

structural metadata specification (d.25)

descriptive metadata specification (d.26)

(d.34)indexing service

browsing service (d.37)

searching service (d.35)

digital object (d.30)

hypertext (d.36)

transmission (d.23)

collection (d. 31)

metadata catalog (d.32)

digital library (minimal) **(d. 38)**

repository (d. 33)

36

Streams

Structures

image

contains

text

video

contains

audio

is_version_of/
cites/links_to

digital
object

describes

metadata
specifications

∈

Collection

describes

Catalog

Index

∈

stores

Repository

Measurable

Topological

is_a

is_a

Measure

is_a

Vector

Metric

employs
produces

Probabilistic

Spaces

employs
produces

employs
produces

Societies

inherits_from/includes

Service
Manager

runs

Service

∈

uses

⊂

extends

Scenario

participates_in

Actor

reuses

Content / People

precedes
happens_before

contains

recipient

association

event

operation

Scenarios

executes

redefines
invokes

Extending 5S

- Higher DL Constructs
  - Collections
  - Catalogs
  - Repositories and Archives
  - Systems
  - Case Studies
- Specialized views and services

Streams **Structures** **Spaces** **Scenarios** **Societies**

community

user role

services

user model

structured stream

structural metadata specification

descriptive metadata specification

digital object

metadata catalog

hypertext

browsing

indexing

searching

user

personalization

collection

repository

Minimal DL

collaboration

image stream

feature vector

base document

superimposed document

image descriptor

composite image descriptor

structured feature vector

superimposed structure

mark

subdocument

image content description

image object

presentation channel

view in context

image descriptor metadata catalog

image collection

image digital object

visualization

CBIR service

complex object structure

complex object

Formal Theory/ Metamodel

5SGraph

5SLGen

DL XML Log

5S $\longrightarrow$ 5SL $\Rightarrow$ OO Classes Workflow $\Rightarrow$ Components $\longrightarrow$ DL Evaluation

Requirements $\longrightarrow$ **Analysis** $\longrightarrow$ **Design** $\longrightarrow$ **Implementation** $\longrightarrow$ **Test**

Tools/Applications

Society Centered

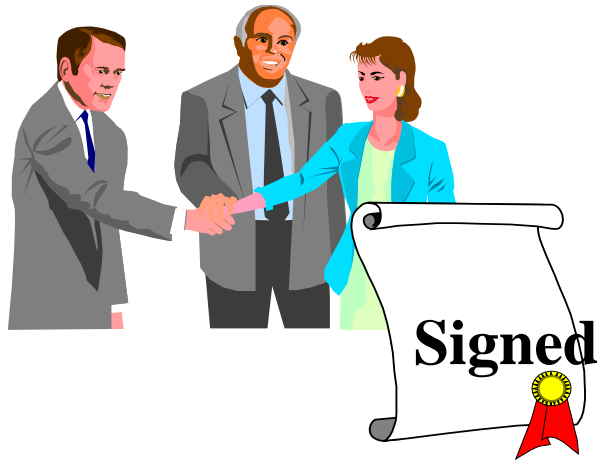- Society, community, group, user
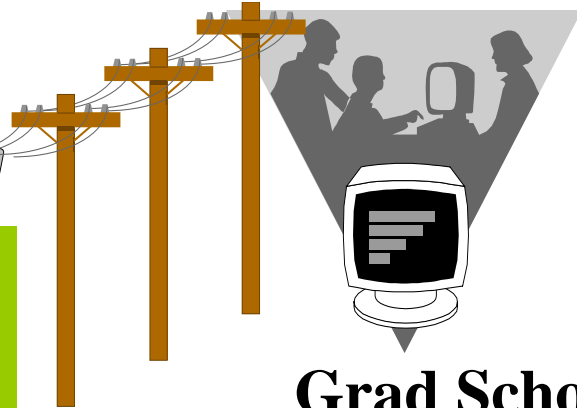- Web 2.0, Social networking
- Computer-supported cooperative work
- User modeling
  - Authors, committee/peers, readers
- Economics / culture
  - Free: but who actually pays, how, implications
  - Low cost: prepaid, but what of preservation
  - Repository hierarchy: group, institution, nation

Student Gets Committee Signatures and Submits ETD

Signed

Approval form
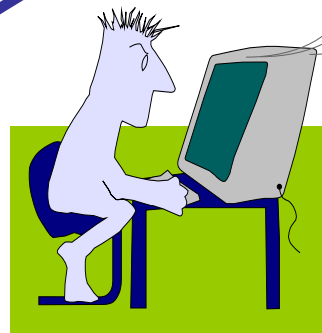
Grad School

Library Catalogs ETD, Access is Opened to the New Research

NDLTD

WWW

Digital library access control
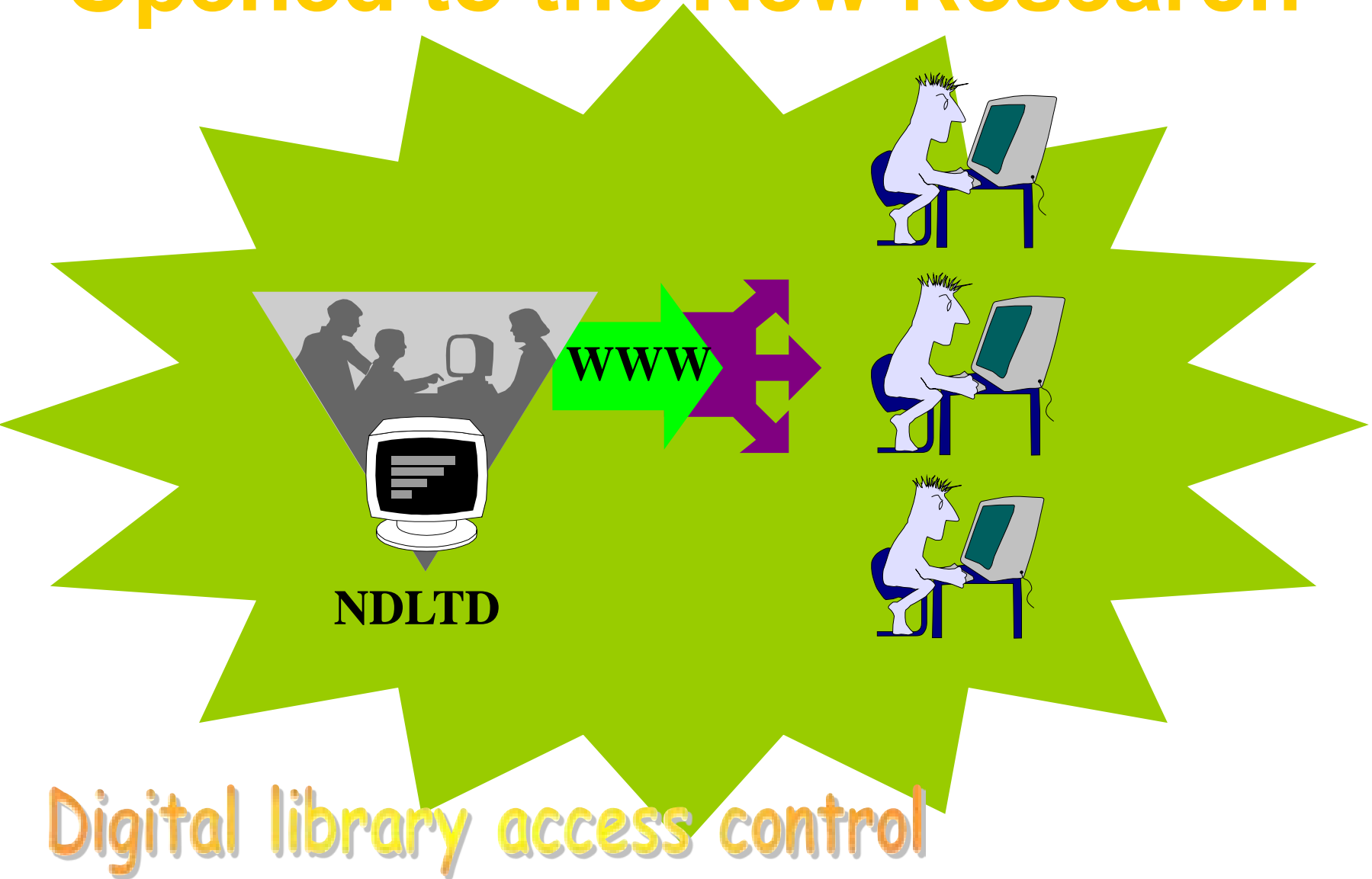
Content Centered

- Genre
  - Gray literature
  - Report, courseware
  - Posters, demos, tutorials, panels, debates
- Format
- Presentation
- Preservation

Part 3 – Services Centered

- Taxonomy
- Interoperability, integration, packaging
  - HTML5
- Collaboration, annotation, recommending
- Indexing, CBIR
- Categorizing, browsing
- Roles of librarians

Infrastructure Services			Information Satisfaction Services
Repository-Building		***Add Value***	
<u>Creational</u>	<u>Preservational</u>		
Acquiring Cataloging Crawling (focused) Describing Digitizing Federating Harvesting Purchasing Submitting	Conserving Converting Copying/Replicating Emulating Renewing Translating (format)	Annotating Classifying Clustering Evaluating Extracting Indexing Measuring Publicizing Rating Reviewing (peer) Surveying Translating (language)	Browsing Collaborating Customizing Filtering Providing access Recommending Requesting Searching Visualizing

DL.Org Functionality WG Dagobert Soergel – Sci. Lead: Functions where Interoperability is important

Behind the scene	For users
Feature extraction	Federated search
Classification / clustering	Incorporating content from other places on the fly
Sharing authority files	Display and visualization
Log file analysis	Timelines
Sharing user profiles	Maps
Harvesting , aggregating	Playing videos
Shared storage and backup	Same look-and-feel browse

Sub-functions of *search*

Quick Search	Advanced Search
Enter a query and click search	Enter a query and click search
Enter keywords or phrases for selected field	Enter keywords or phrases for selected fields
Limit results to	Select keyword from a list
Search subscribed titels	Select Boolean operator (explicit)
Clear	Define phrase match (explicit)
	Clear
	Search within results
	Limit results to (preselection)
	Sort by (preselection)
	Select display options
	Display X results per page
	Display search history

Sub-functions of *annotate*

Select object to be annotated
    (need to indicate selection method)

Mark region in the object
    (many different methods depending on the object)

Select type of annotation
    (highlight, mark with special meaning, text, image,
sound)

If text, image, sound
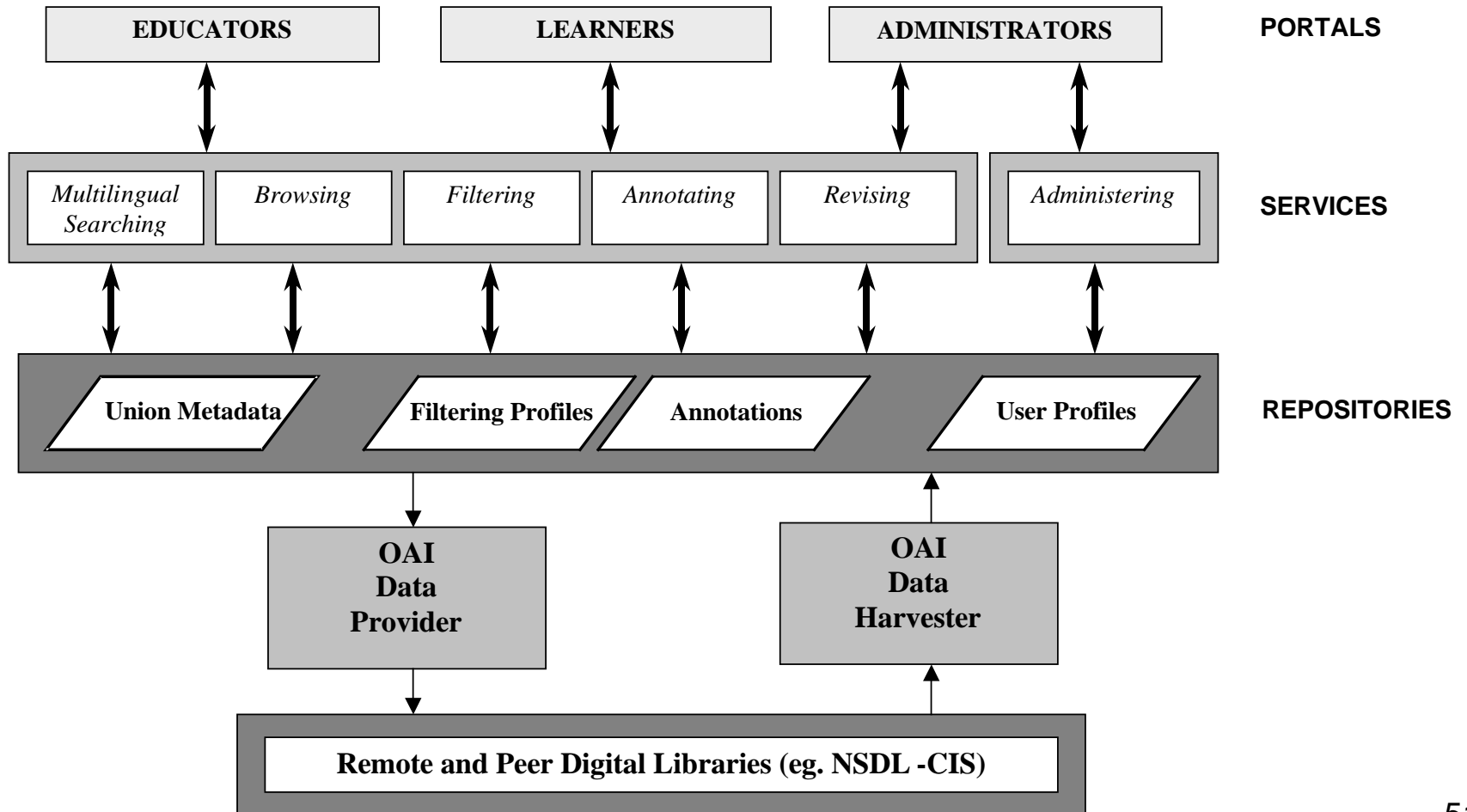
    Specify relationship to object to be annotated

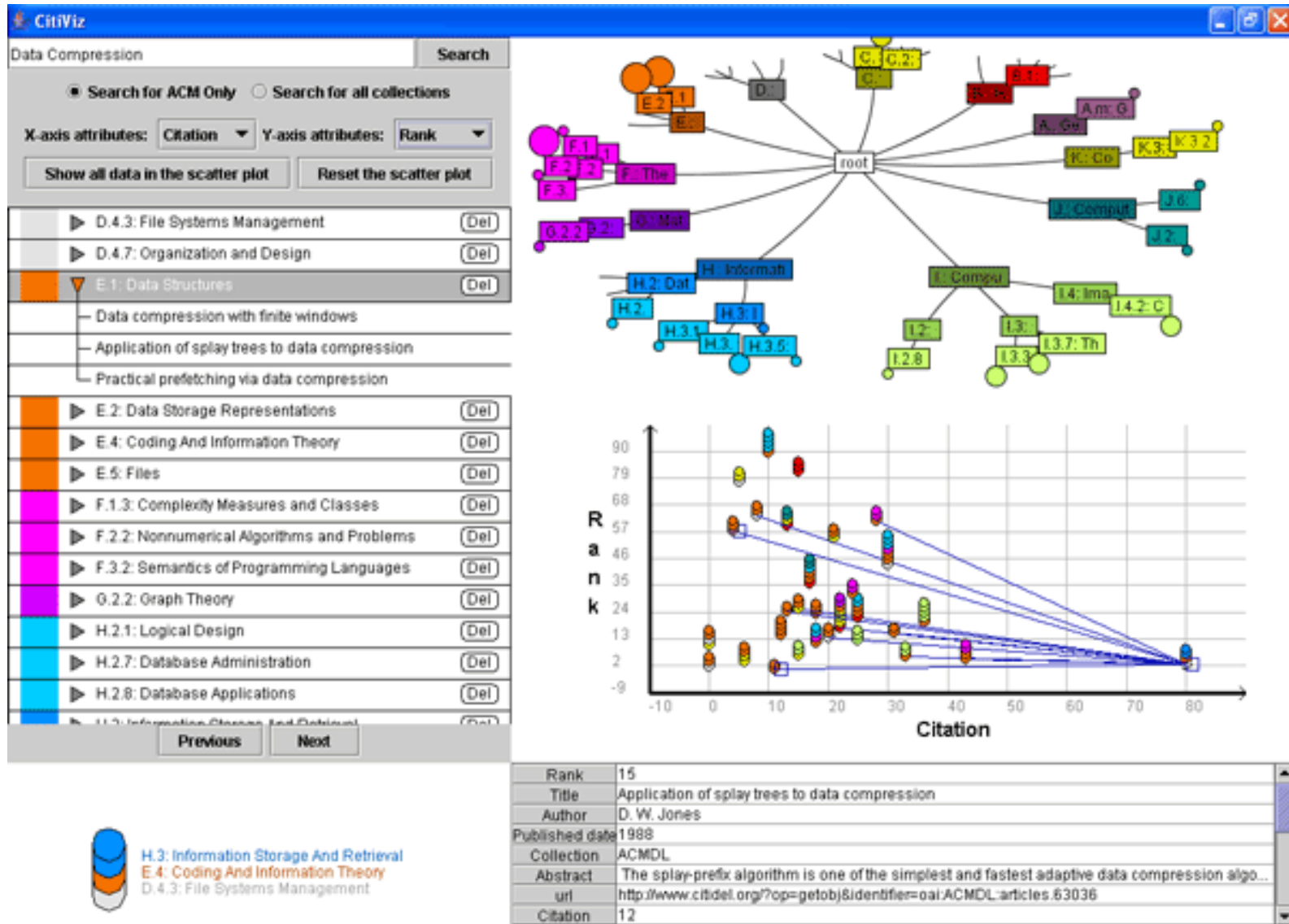   Select or create the annotating object
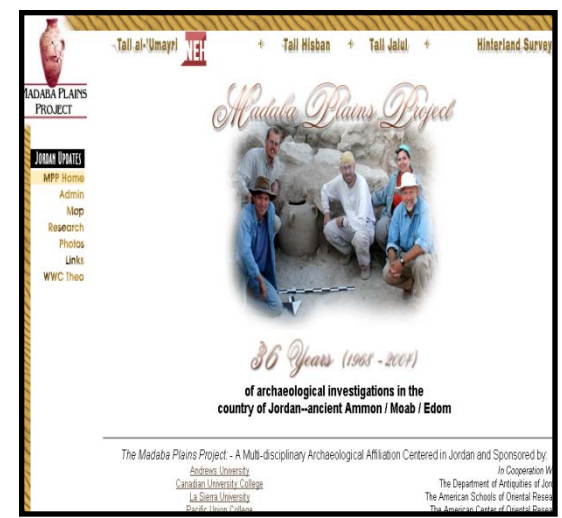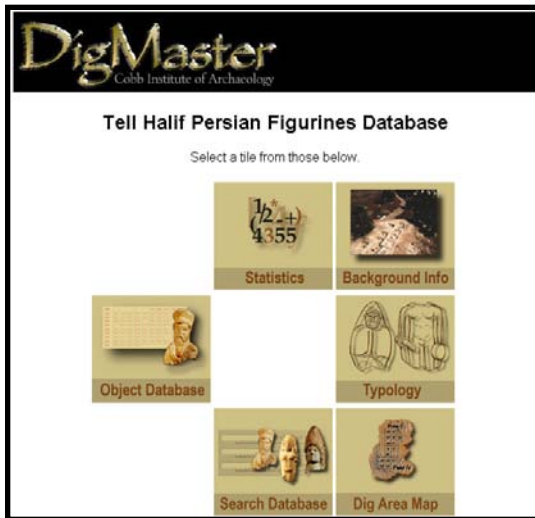   (possibly specifying a region

Annotating within one system

Annotating across systems

Digital library architecture for local and interoperable CITIDEL services

EDUCATORS	**LEARNERS**	**ADMINISTRATORS**

PORTALS

Multilingual Searching	*Browsing*	*Filtering*	*Annotating*	*Revising*	*Administering*

SERVICES

Union Metadata	Filtering Profiles	Annotations	User Profiles

REPOSITORIES

OAI Data Provider

OAI Data Harvester

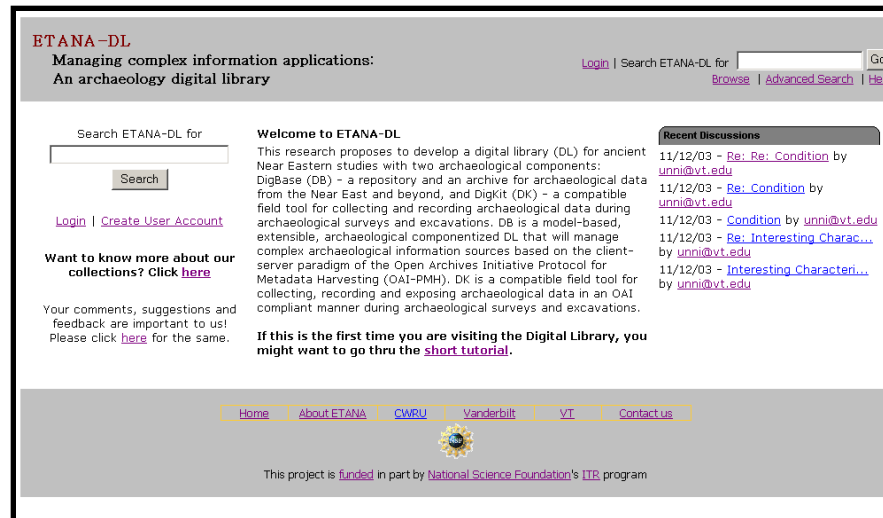**Remote and Peer Digital Libraries (eg. NSDL -CIS)**
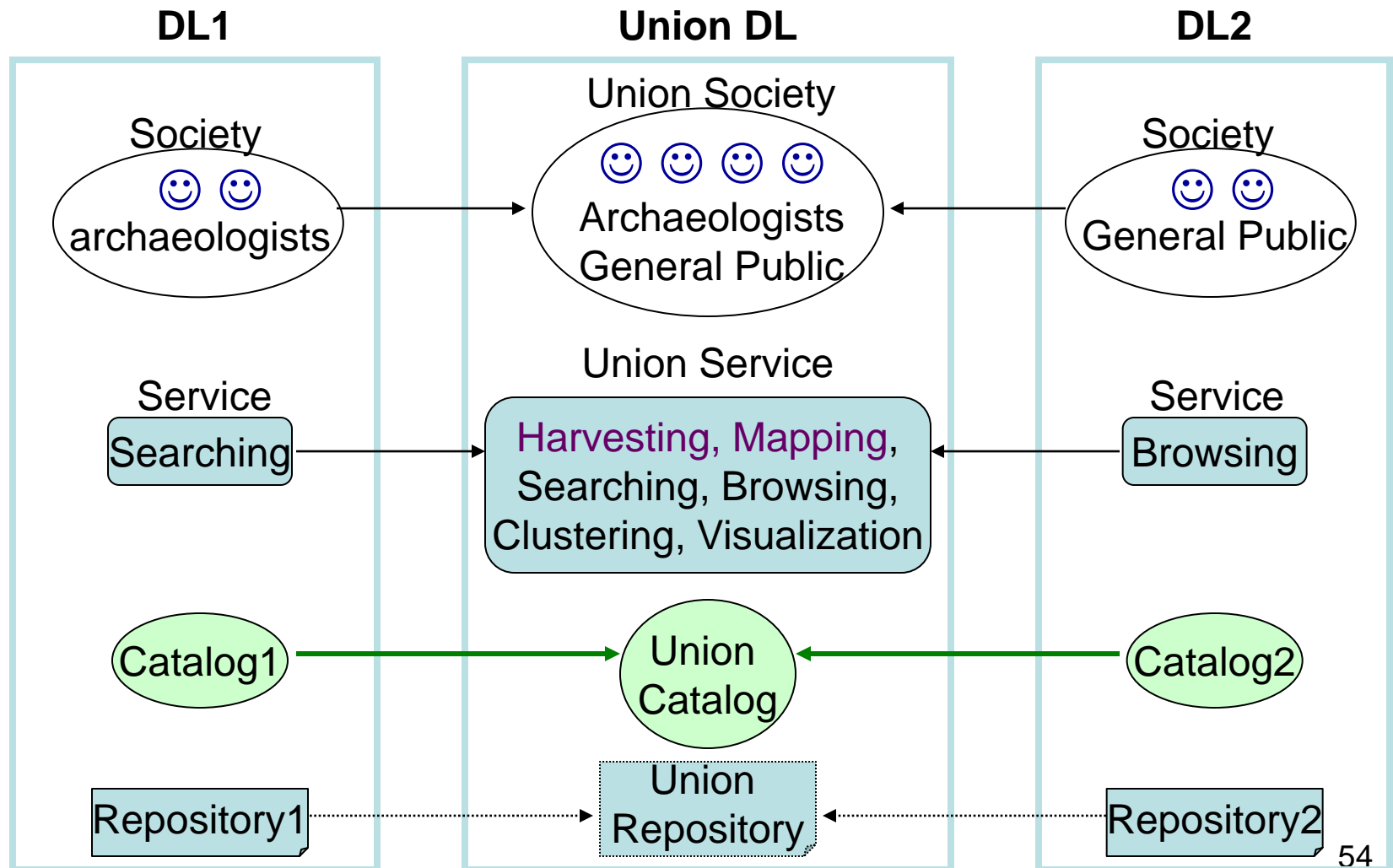
51

Example of Union Service: CitiViz

ETANA.org

Architecture of a Union DL (ETANA.org)

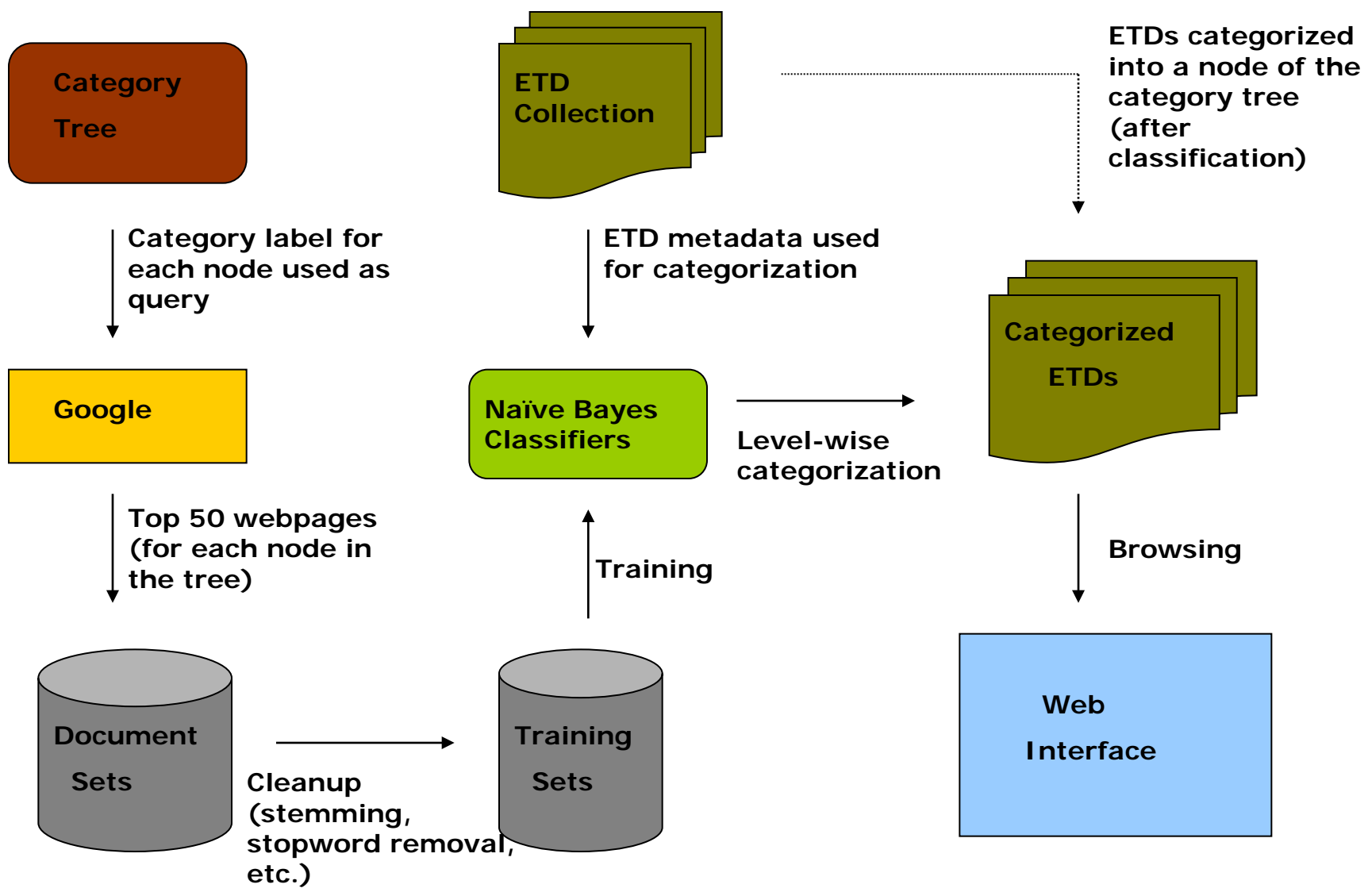# Union Catalog Integration
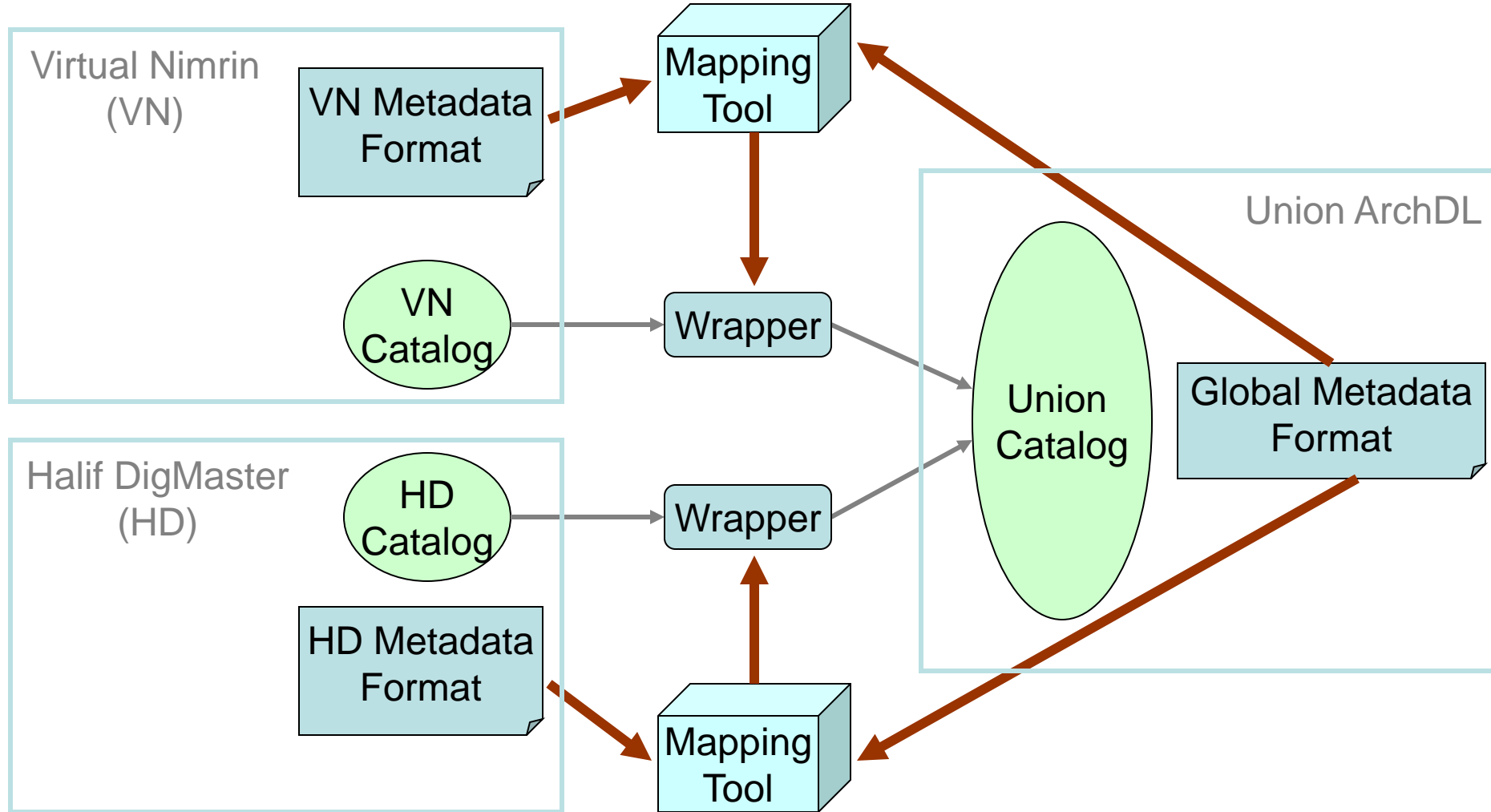
HTML5 Structuring Flowchart

PDF ETD

MP3

HTML

PDF2Text/ HTML converter

TXT/ HTML

TXT/ HTML

ETD structure analyzer

Multimedia file link extractor

Multimedia file source extractor

Tagged TXT

Tagged TXT

Tagged MM Source

HTML5 tag set

Text/ Grammar

HTML5 Converter

HTML5 ETD
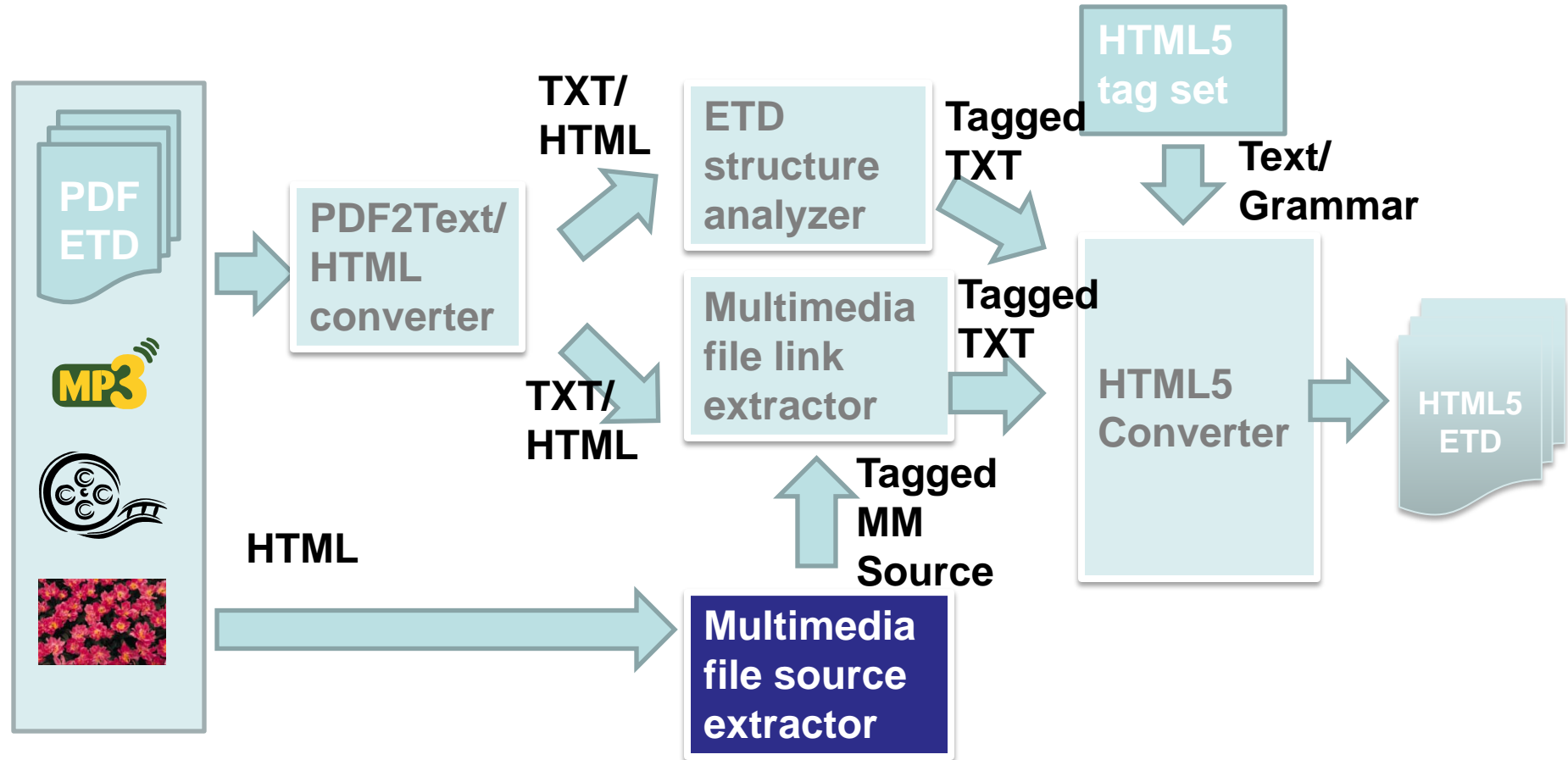
ETD Classification: Algorithm Pipeline

Category Tree

ETD Collection

ETDs categorized into a node of the category tree (after classification)

Category label for each node used as query

ETD metadata used for categorization

Google

Naïve Bayes Classifiers

Categorized ETDs

Level-wise categorization

Top 50 webpages (for each node in the tree)

Training

Browsing

Document Sets

Cleanup (stemming, stopword removal, etc.)

Training Sets

Web Interface

Digital Librarians

- **Community oriented**
- **Collection management**
- **Customized services**

- **Principles:**
  - Openness
  - Expansion
    - Interoperation, integration, communitization

Summary

- Selected DL Projects
- Basic DL Concepts
- Services Centered

- Openness
- Expansion

- Questions and Comments?
- http://fox.cs.vt.edu/talks/2010/