Georgetown University

29 April 2011

“A Formal Approach to Digital Libraries - The 5S Framework: Societies, Scenarios, Spaces, Structures, Streams”

by Edward A. Fox

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

Ed Fox, Director, Virginia Tech Digital Library Research Laboratory
Executive Director, Networked Digital Library of Theses & Dissertations
Chair, Steering Committee, ACM/IEEE Jt. Conf. on Digital Libraries
Member, Board of Directors, Computing Research Association

Digital	Information	Libraries	Retrieval
Chair	Committee	Conference	Workshop
Education	Electronic	Dissertations	Theses
Research	Science	Systems	Technology
Open	Initiative	International	Multimedia
Access	Archive	Intelligent	Interactive
Knowledge	Learning	SIGIR	Tutorial

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, …

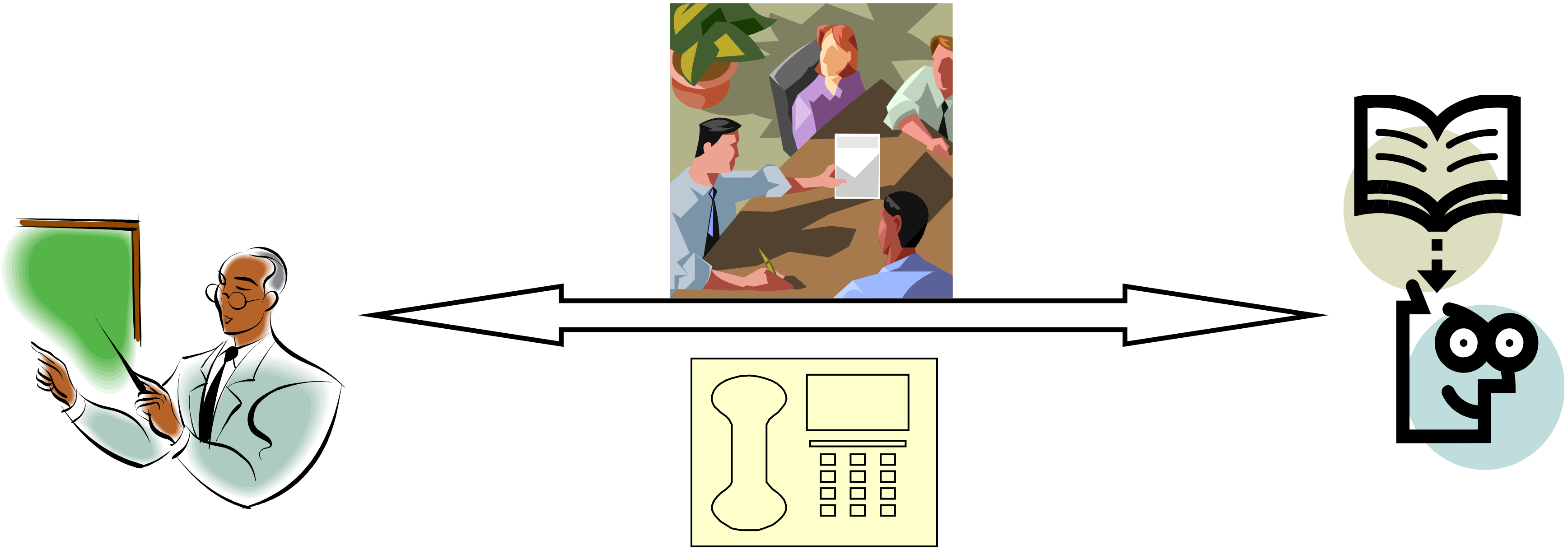
Selected DL Projects

- 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

Outline

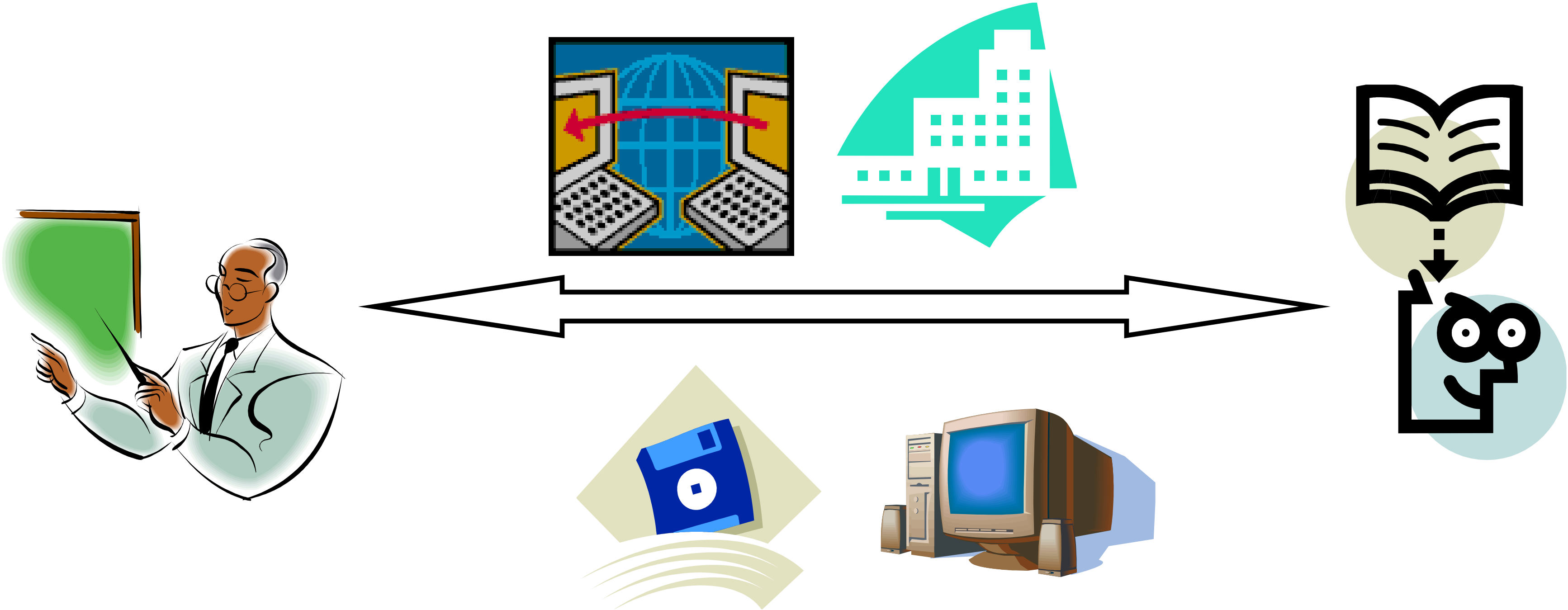
- **An informal overview of digital libraries**
- An informal view of 5S
- A formal perspective of 5S
- What has been done with 5S
- Future plans

Synchronous Scholarly Communication



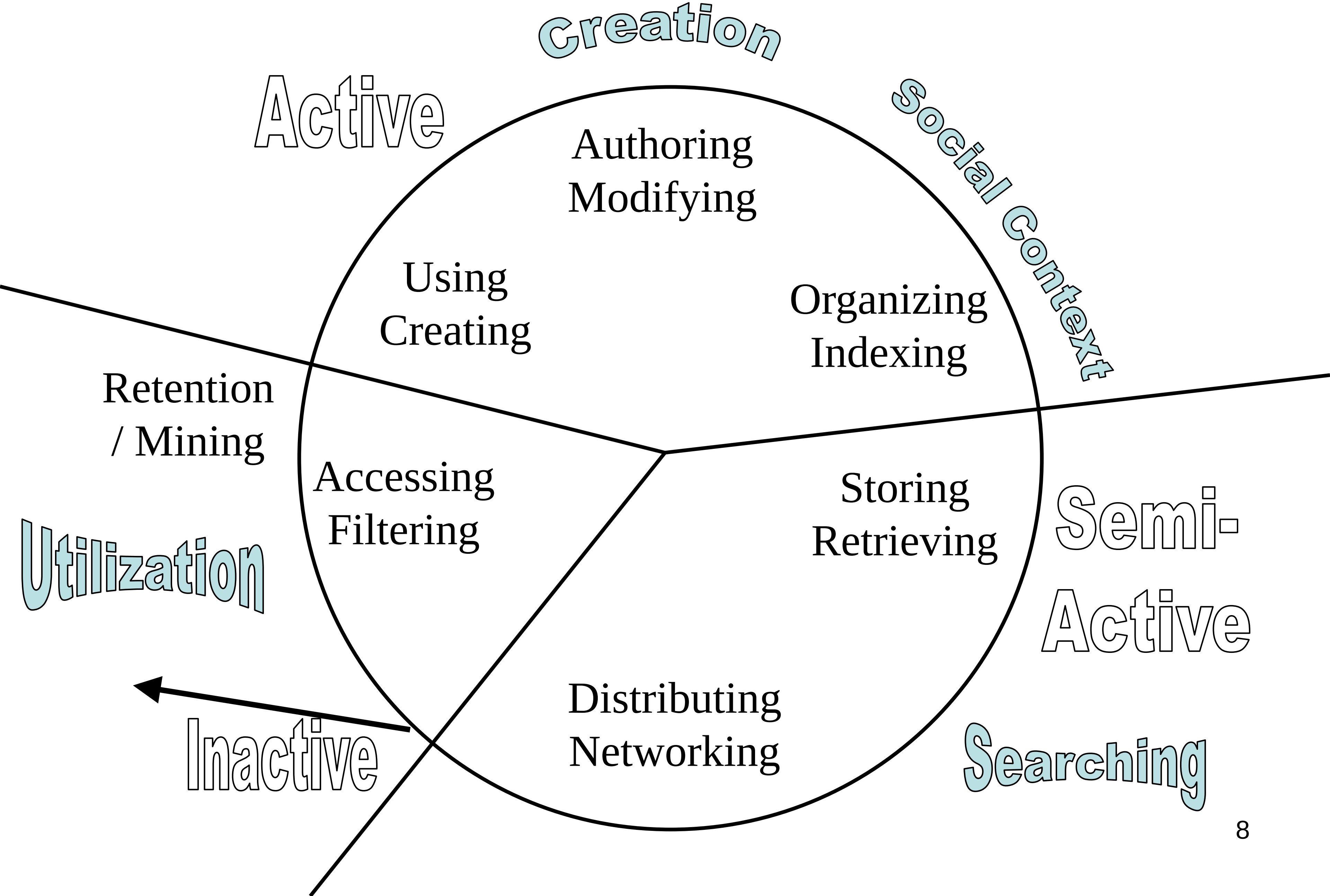
Same time, Same or different place

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

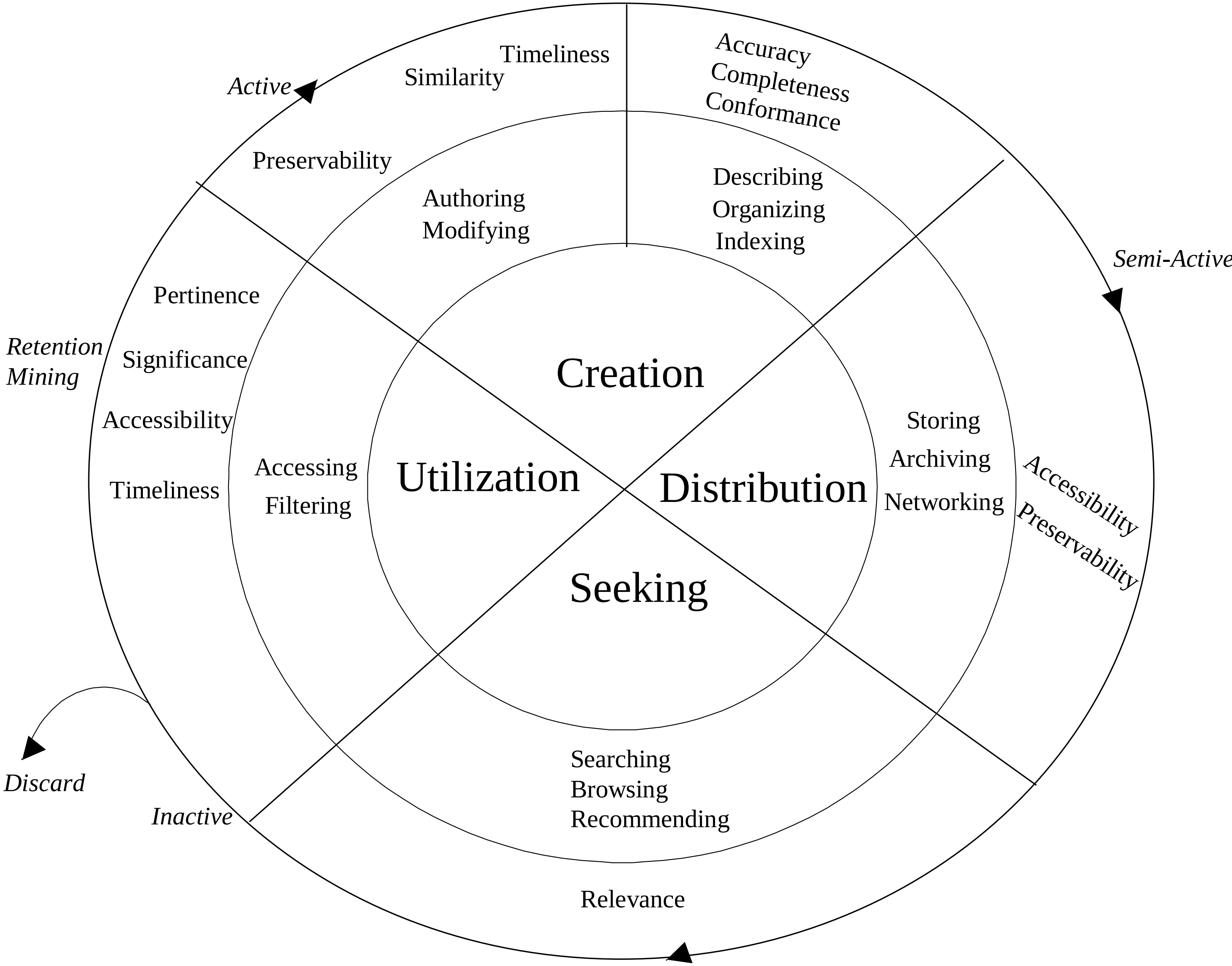
Utilization

Distributing
Networking

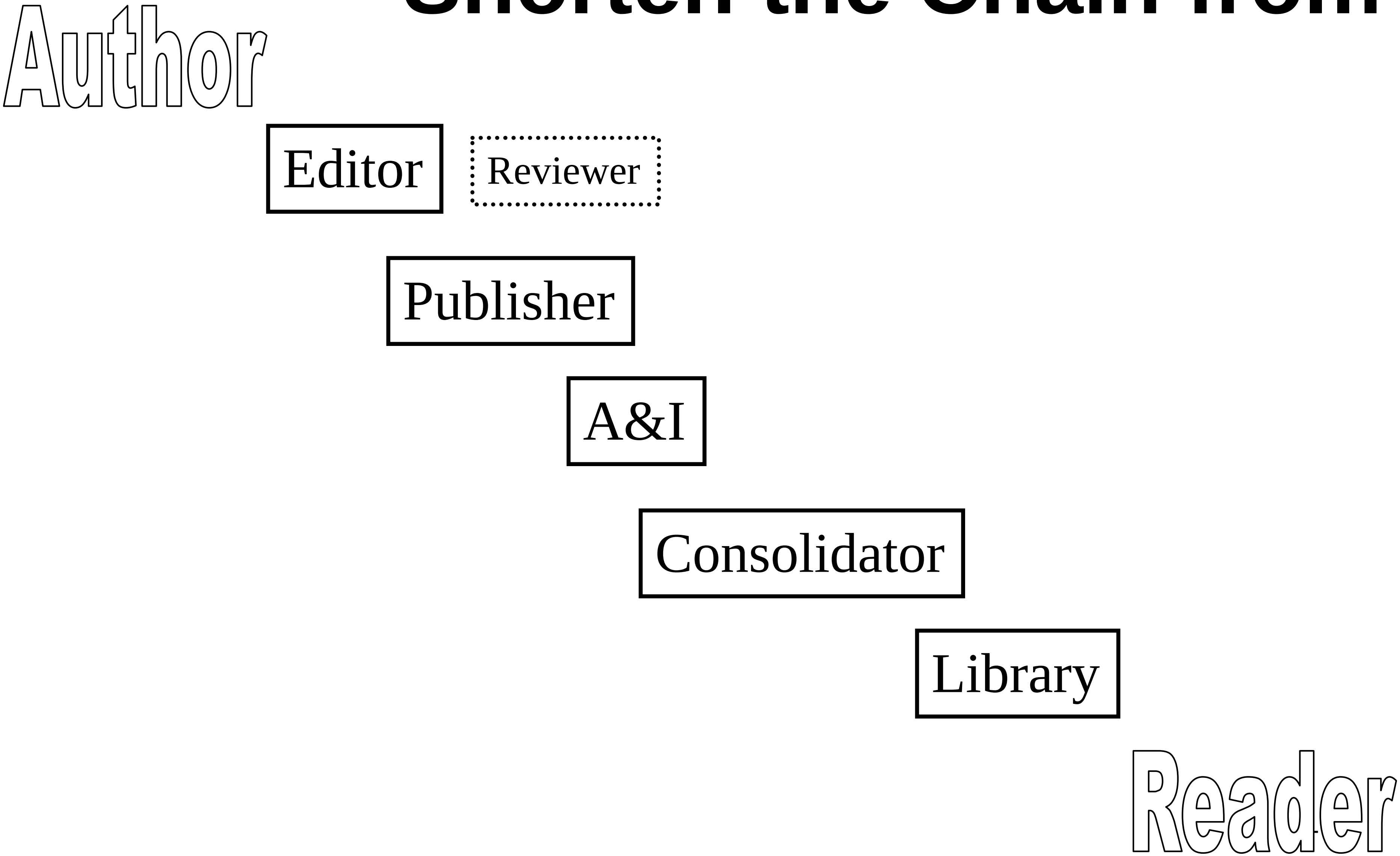
Inactive

Searching

Quality and the Information Life Cycle



Digital Libraries Shorten the Chain from



Author

Editor

Reviewer

Publisher

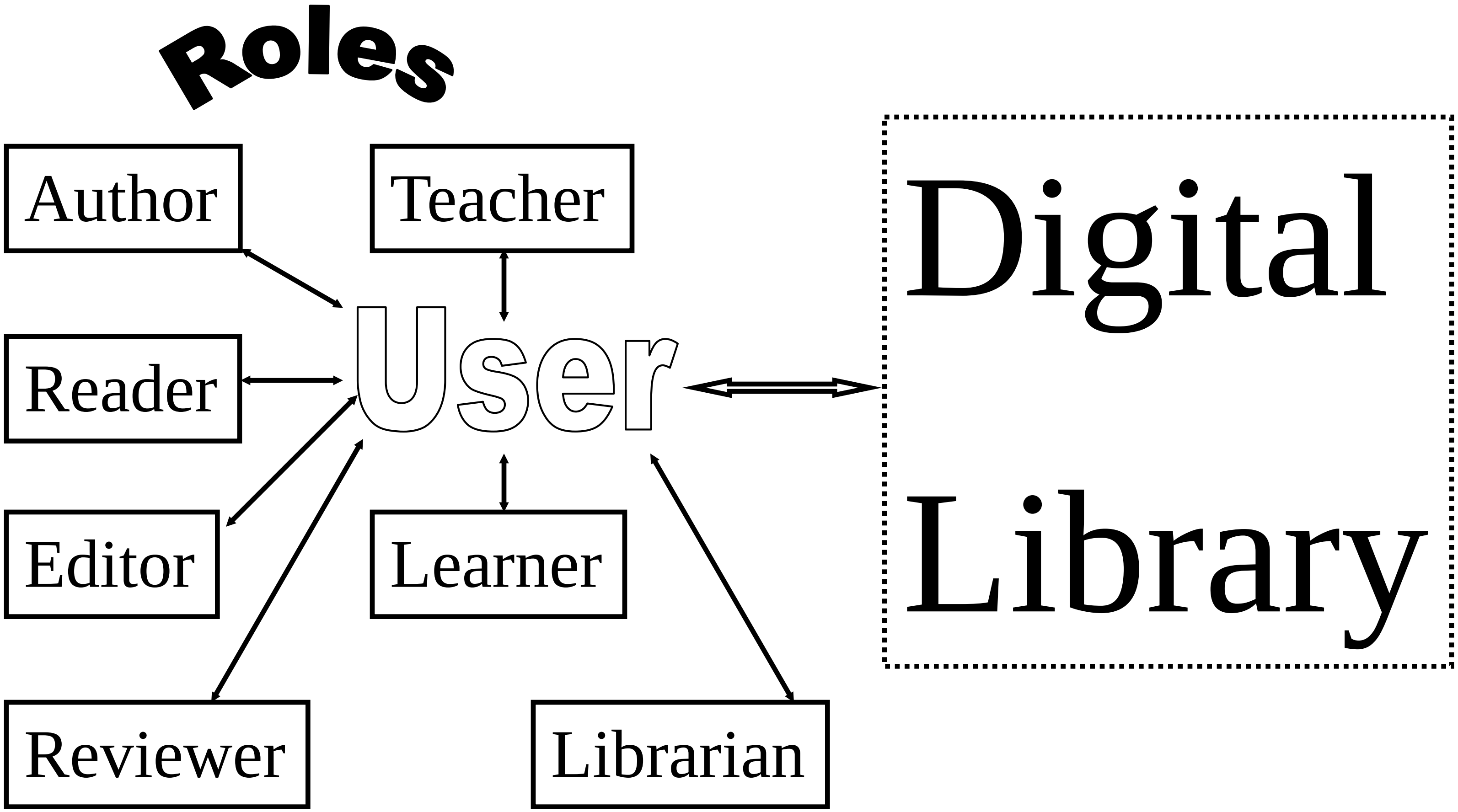
A&I

Consolidator

Library

Reader

DLs Shorten the Chain to



Roles

Author

Teacher

User

Reader

Editor

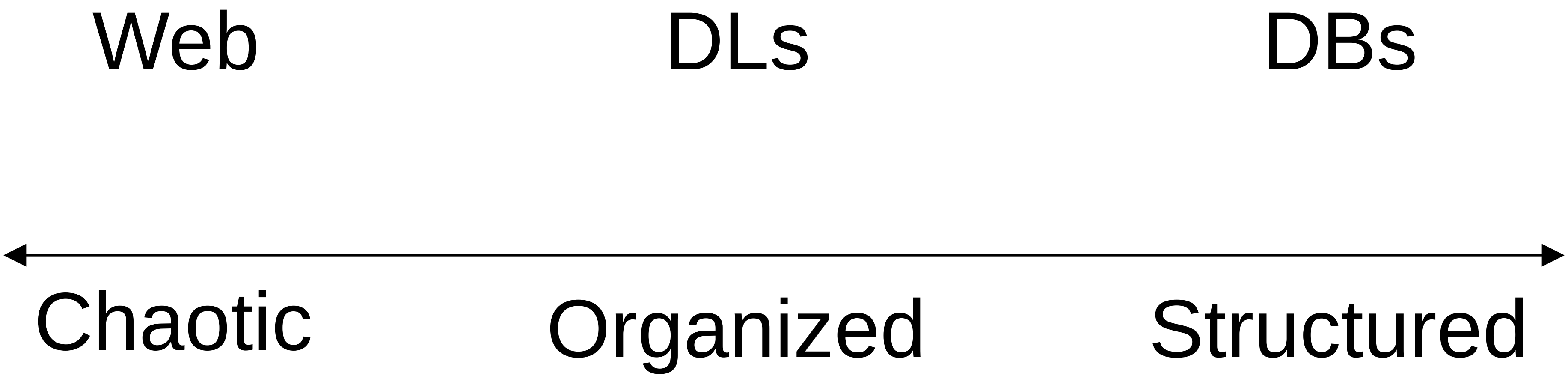
Learner

Reviewer

Librarian

Digital Library

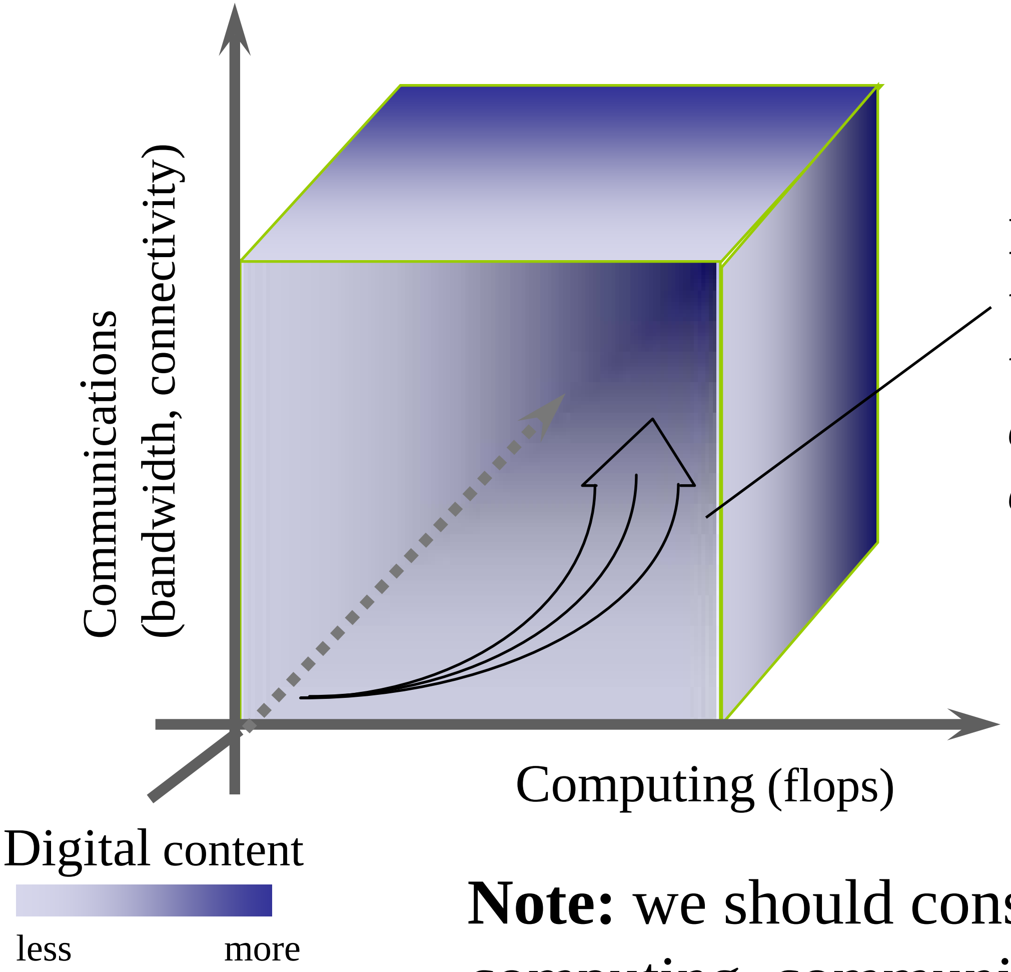
Degree of Structure



Web DLs DBs

Chaotic Organized Structured

Locating Digital Libraries in Computing and Communications Technology Space



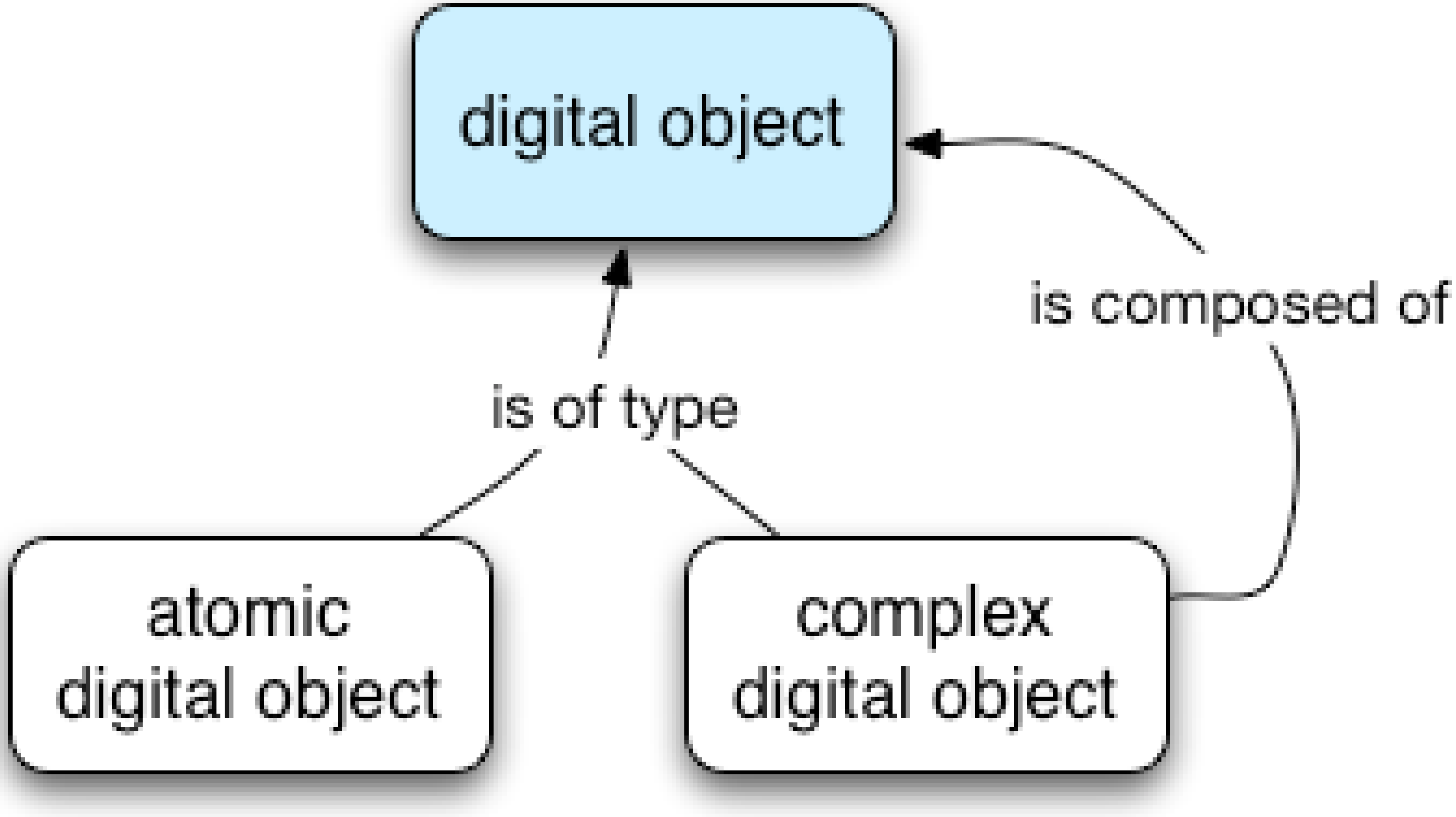
Digital Libraries technology trajectory: ***intellectual access to globally distributed information***

Note: we should consider 4 dimensions: computing, communications, content, and community (people)

Digital Objects (DOs)

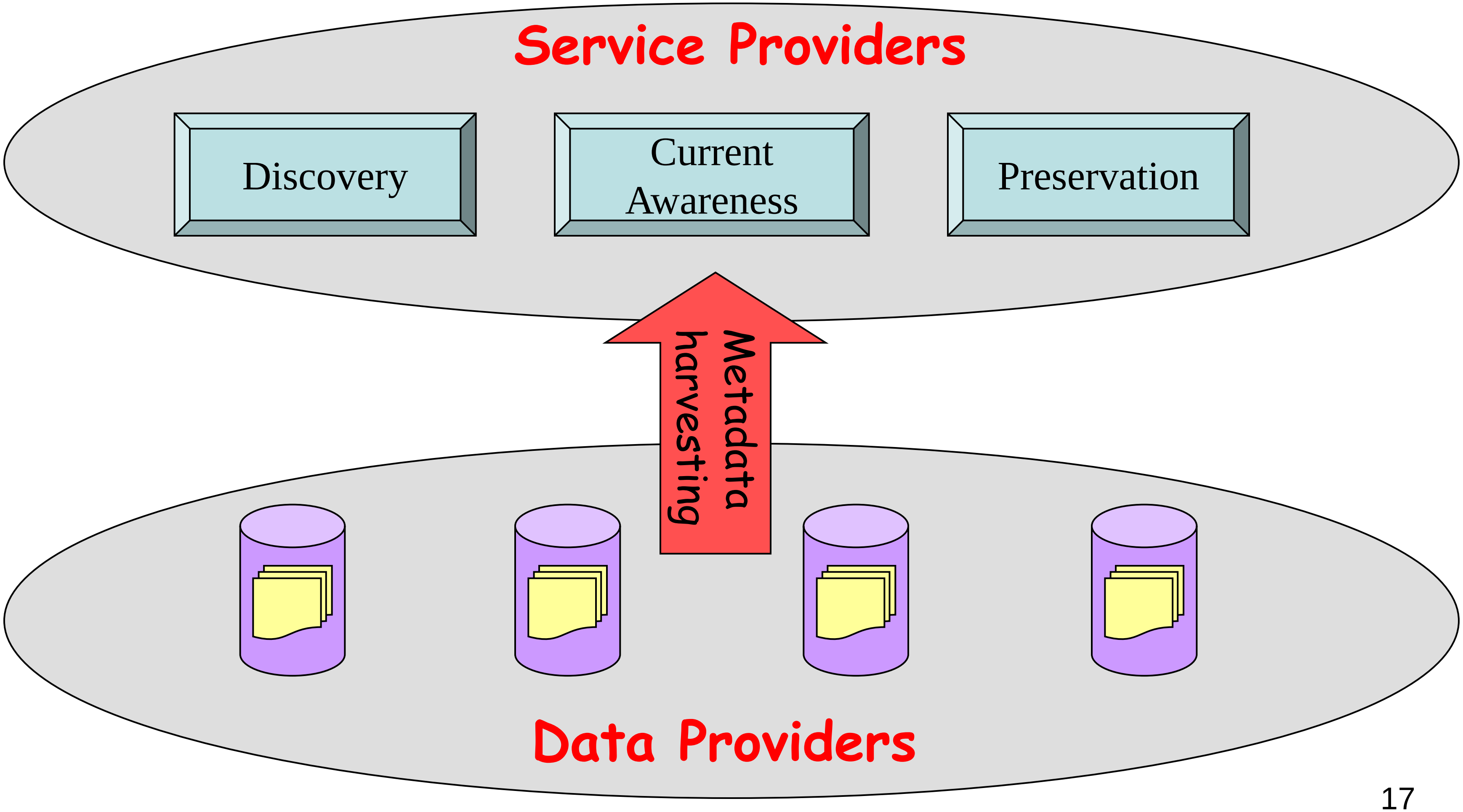
- Born digital
- Digitized version of “real” object
  - Is the DO version the same, better, or worse?
  - Separation of structure, meaning, use
    - Rendered on paper, laptop, handheld – or CAVE
    - Semantic Web (human or machine processing)
- Surrogate for “real” object
  - Hybrid systems with real and digital objects
  - Data, documents, subdocuments, metadata

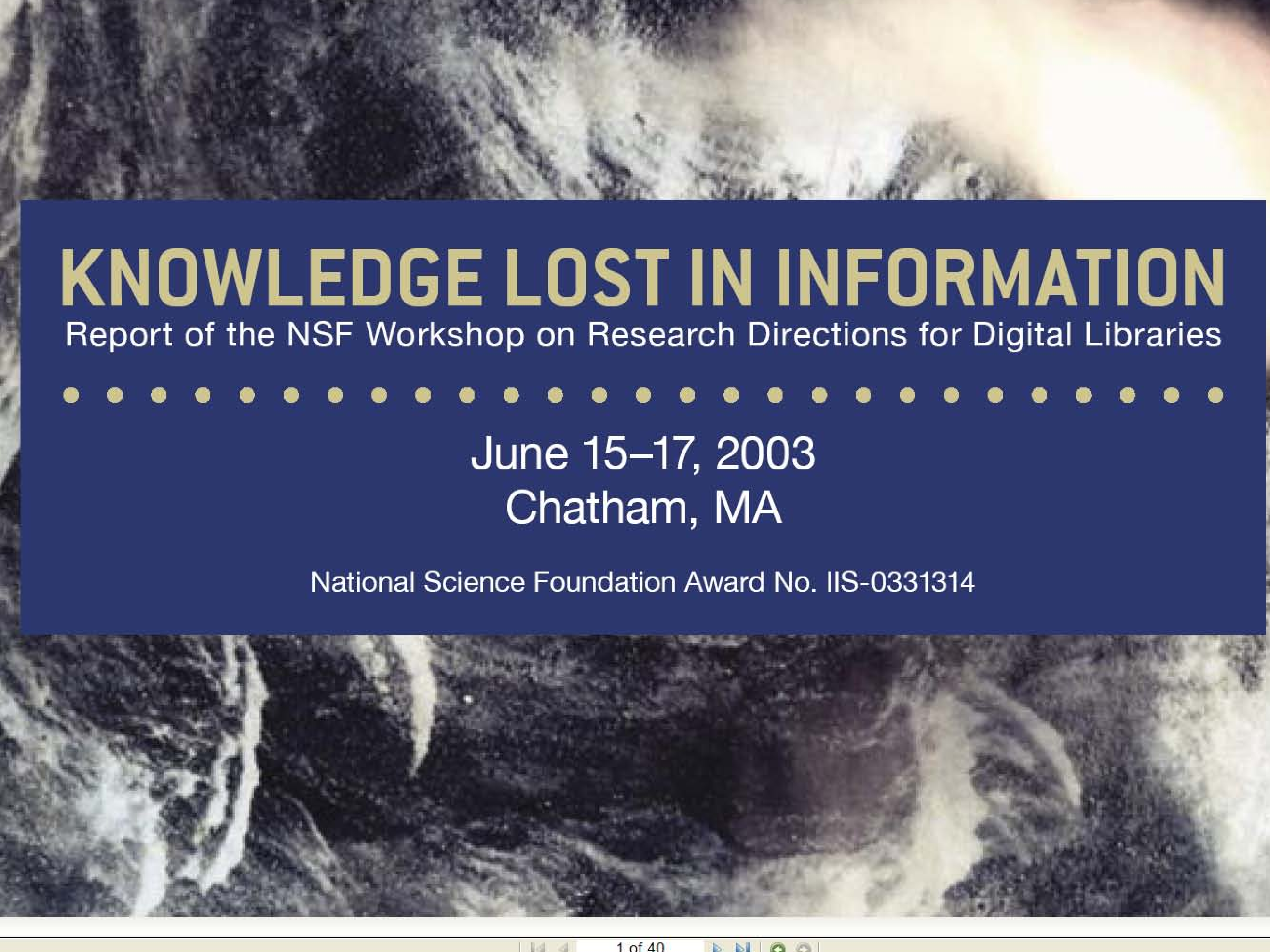
A concept map for complex object composition



A digital library = repository of collections and metadata + services

The World According to the Open Archives Initiative



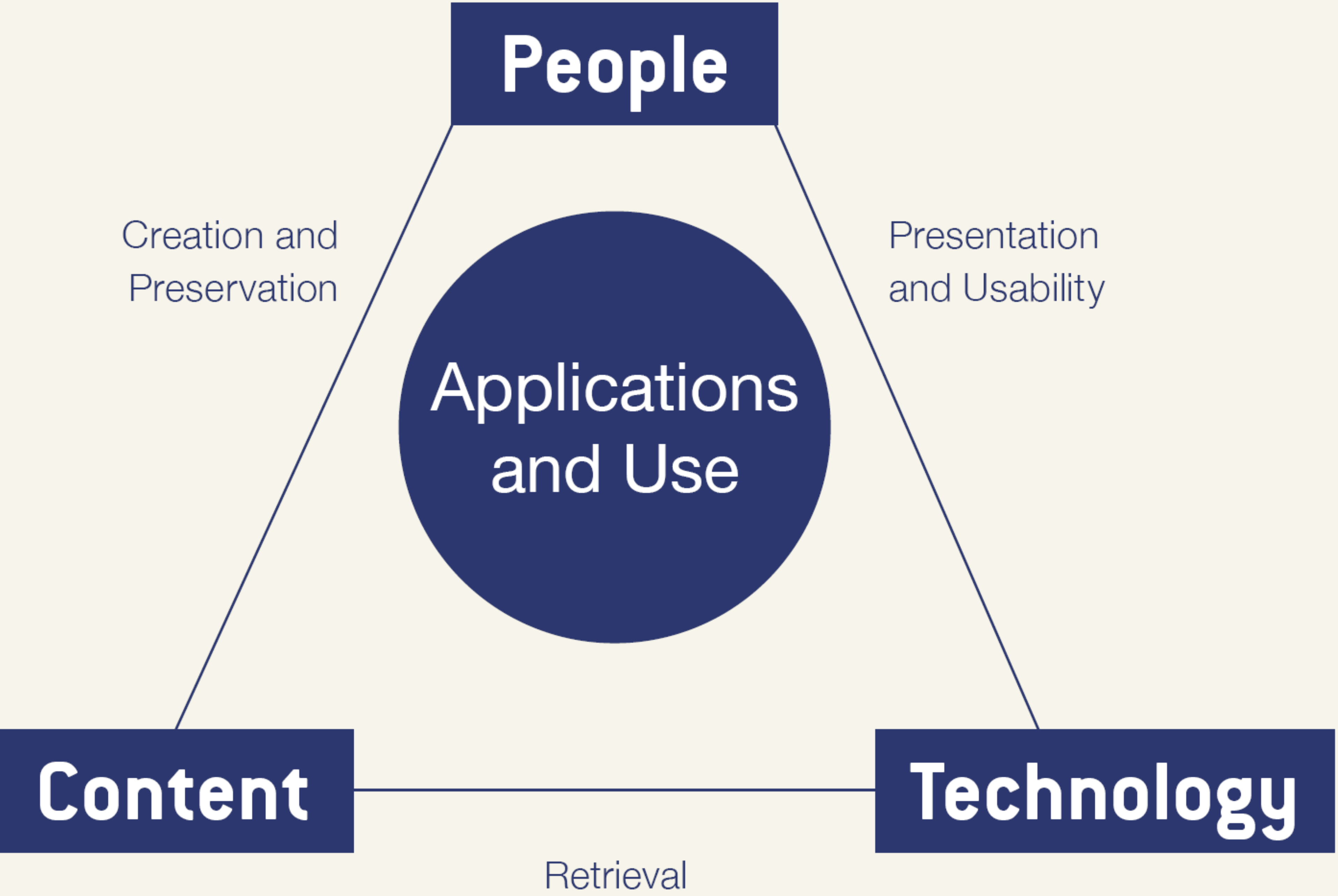


KNOWLEDGE LOST IN INFORMATION

Report of the NSF Workshop on Research Directions for Digital Libraries

June 15–17, 2003
Chatham, MA

National Science Foundation Award No. IIS-0331314



People

Creation and Preservation

Presentation and Usability

Applications and Use

Content

Technology

Retrieval

Educational Repositories Connect:

Users: students, educators, life-long learners

Content: structured learning materials; large real-time or archived datasets; audio, images, animations; primary sources; digital learning objects (e.g. applets); interactive (virtual, remote) laboratories; ...

Tools: search; refer; validate; integrate; create; customize; publish; share; notify; collaborate; ...

Collections

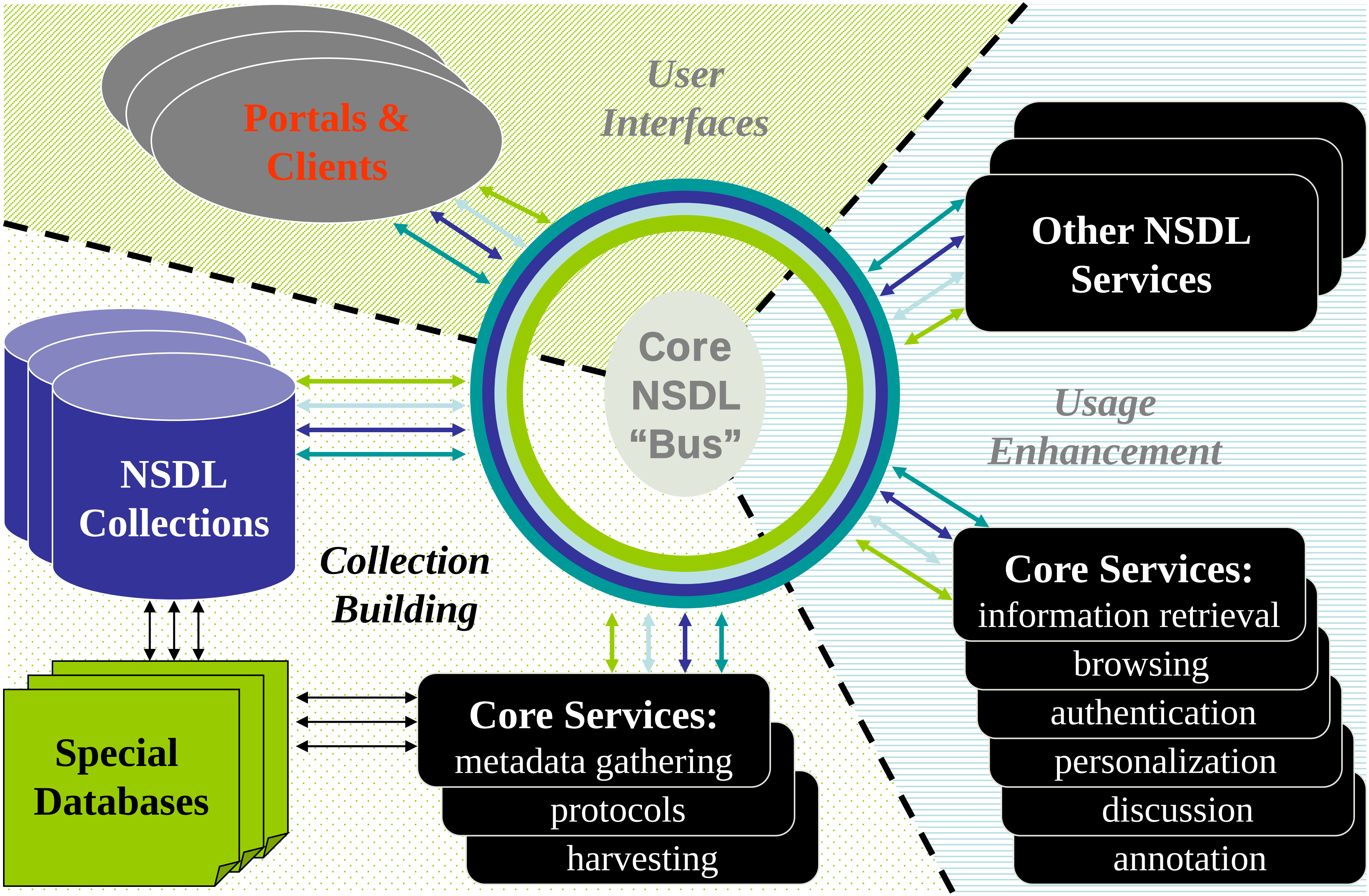
- Discovery of content
- Classification and cataloguing
- Acquisition and/or linking; referencing
- Disciplinary-based themes define a natural body of content, but other possibilities are also encouraged
- Access to massive real-time or archived datasets
- Software tool suites for analysis, modeling, simulation, or visualization
- Reviewed commentary on learning materials and pedagogy

Services

- Help services, frequently asked questions, etc.
- Synchronous/asynchronous collaborative learning environments using shared resources
- Mechanisms for building personal annotated digital information spaces
- Reliability testing for applets or other digital learning objects
- Audio, image, and video search capability
- Metadata system translation
- Community feedback mechanisms

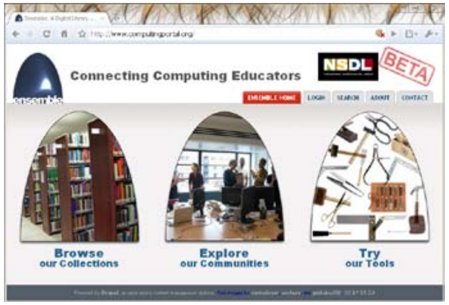
NSDL Information Architecture

Essentially as developed by the Technical Infrastructure Workgroup

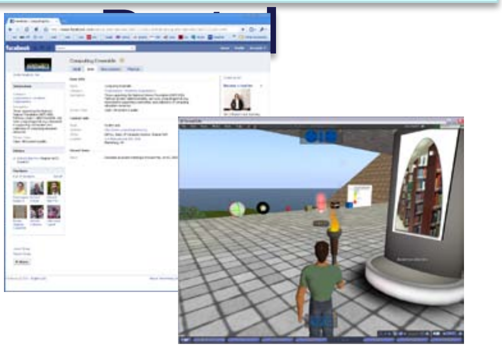


The Ensemble Computing Portal

Many-to-Many Information Connections in a Distributed Digital Library Portal



Distributed DL Portal



Collections

AlgoViz

CITIDEL

paws

Computer Science Syllabus

csta Computer Science Teachers Association

Services

Search

Forum

Group

Blog

Browse

Notification

Tools

Communities

ACM

CS1

FOCES

TECH Developers

Ensemble

LIKES

AP CS

A collaborative research project to build a distributed portal with up-to-date contents for all computing communities.

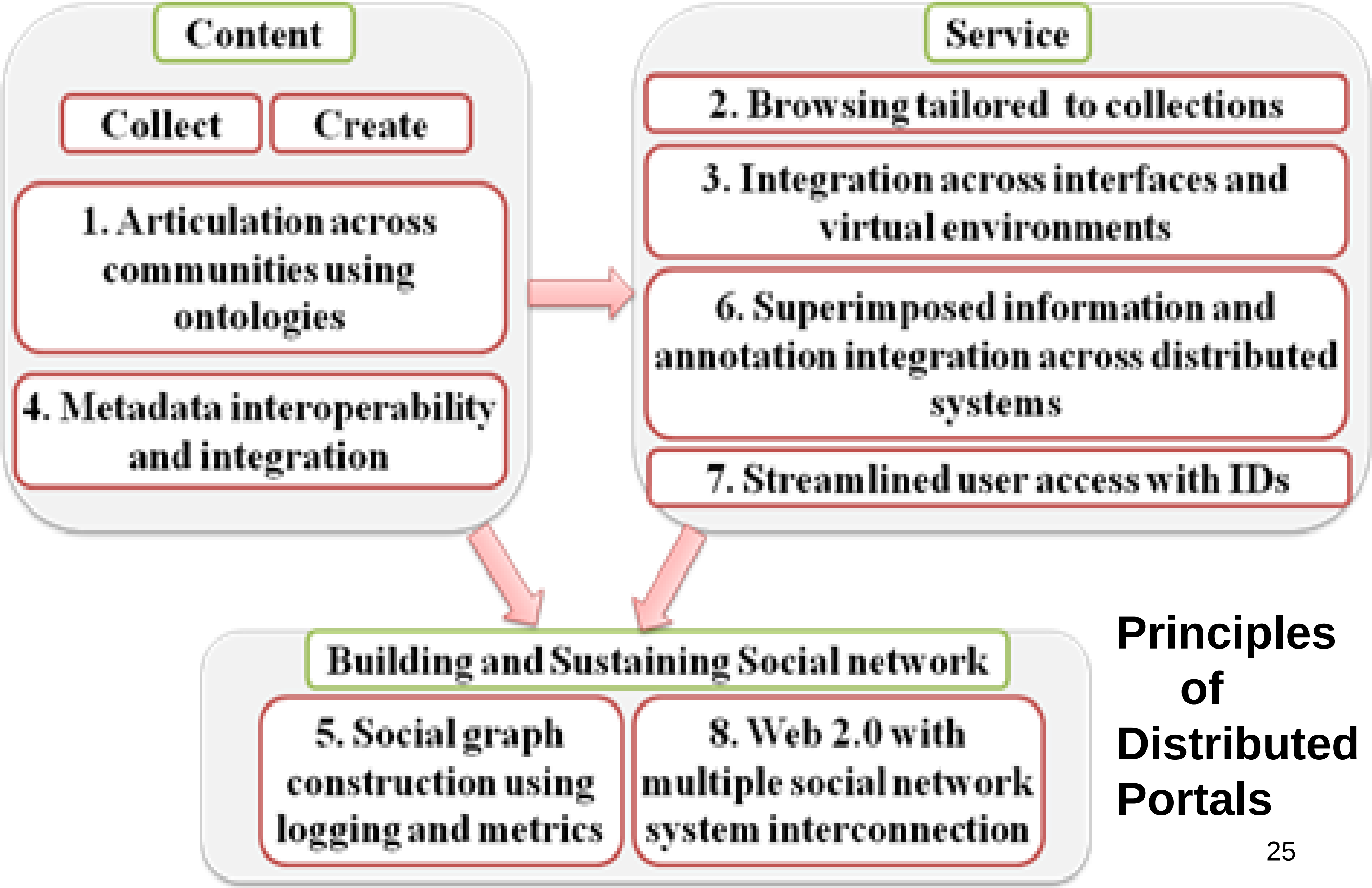
http://www.computingportal.org/

ensemble

NSF

NSDL THE NATIONAL SCIENCE DIGITAL LIBRARY

Ensemble: PDP-8 Overview



Content

Collect

Create

1. Articulation across communities using ontologies

4. Metadata interoperability and integration

Service

2. Browsing tailored to collections

3. Integration across interfaces and virtual environments

6. Superimposed information and annotation integration across distributed systems

7. Streamlined user access with IDs

Building and Sustaining Social network

5. Social graph construction using logging and metrics

8. Web 2.0 with multiple social network system interconnection

Principles of Distributed Portals

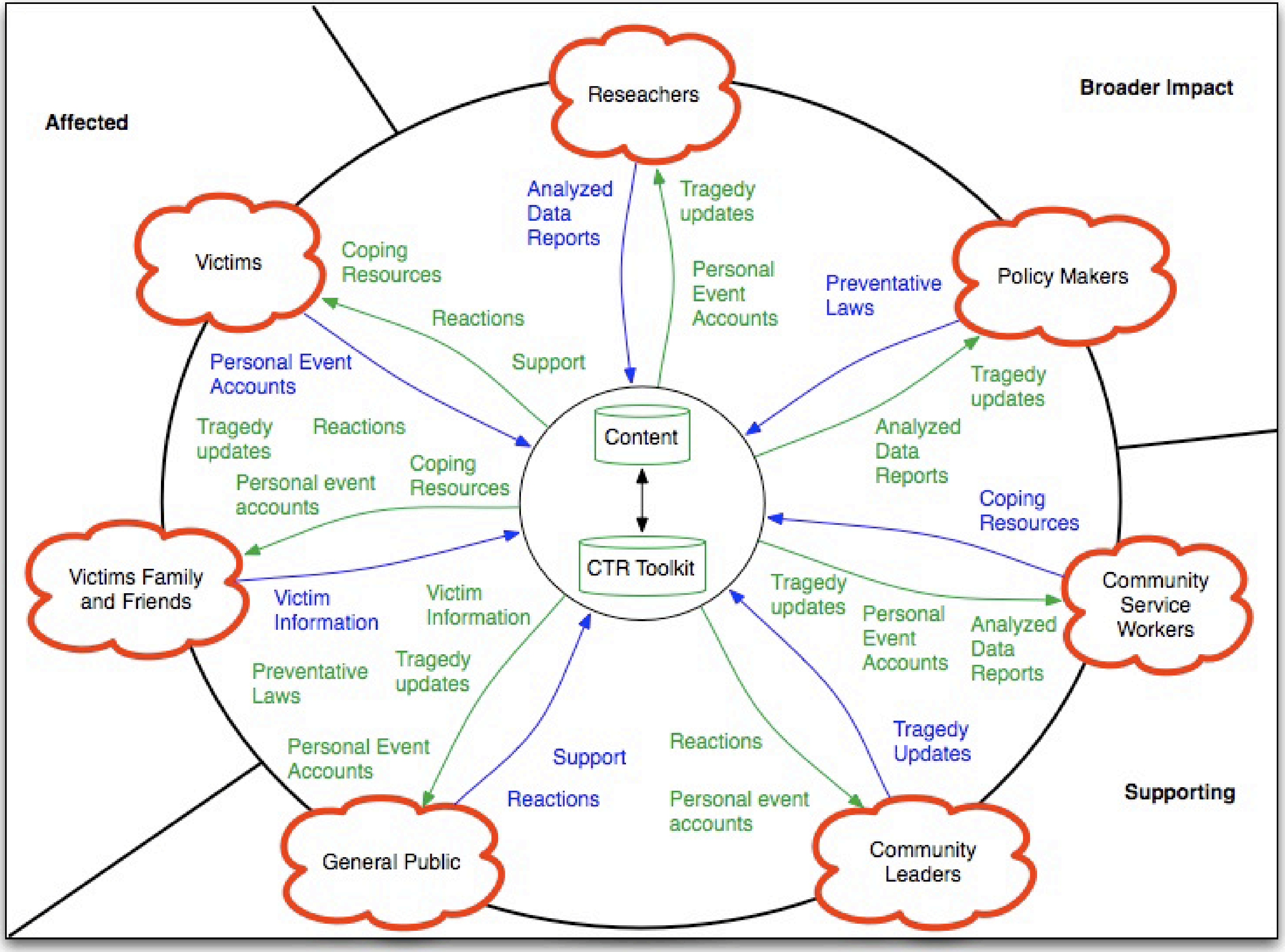
Networked Digital Library of Theses and Dissertations: www.ndltd.org

- N D Ltd or Noodle TD
- Vision: Every thesis and dissertation in the world is:
  - Devised to take advantage of the most helpful electronic publishing methods
  - Shared globally and easily found
  - Supported by a suite of digital library services to aid authors, researchers, learners, universities
  - Preserved and migrated permanently

Crisis, Tragedy & Recovery net

www.ctrnet.net

CTR stakeholders



CC2001 Computer Science volume

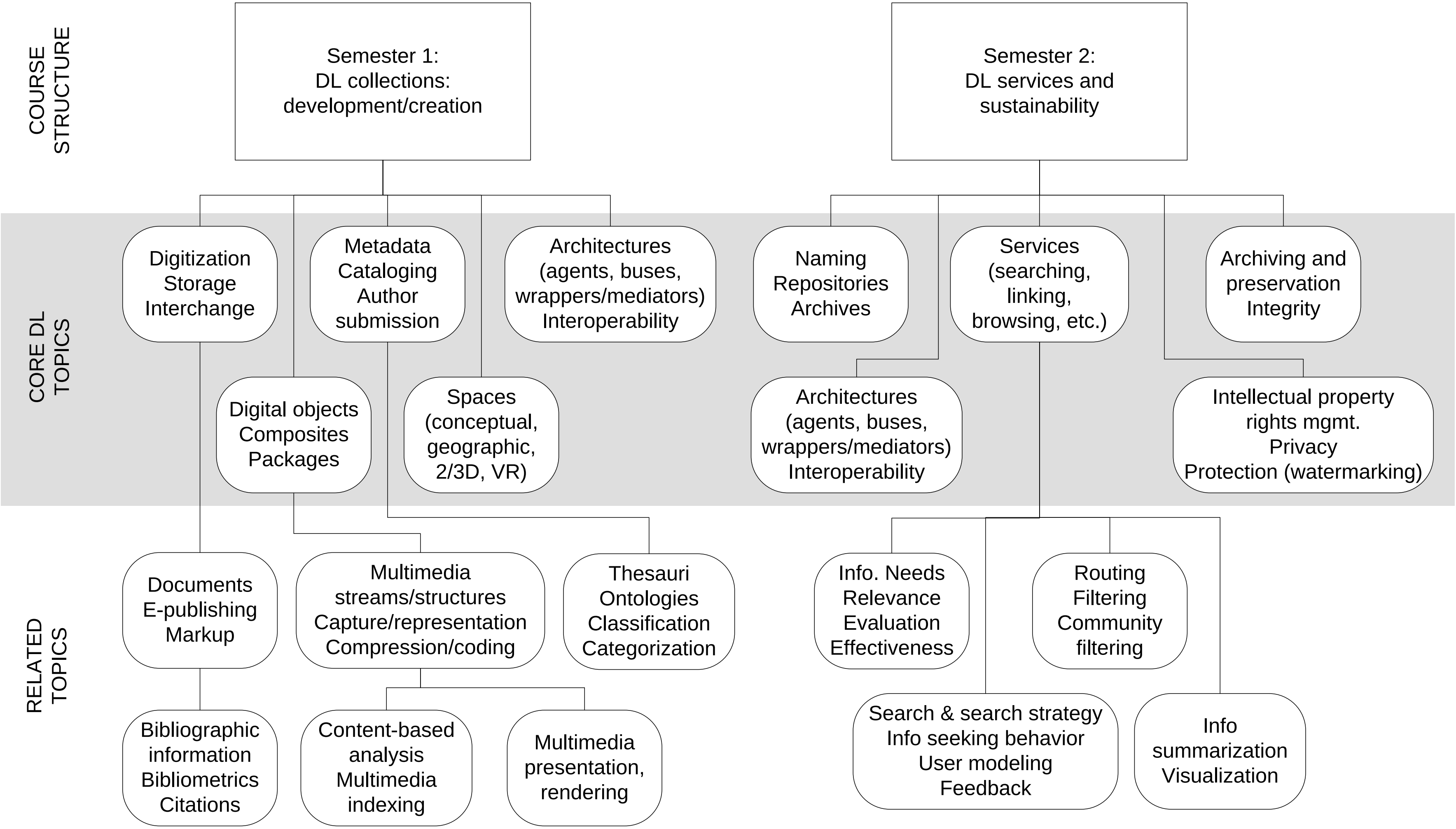
- **DS. Discrete Structures**
- **PF. Programming Fundamentals**
- **AL. Algorithms and Complexity**
- **AR. Architecture and Organization**
- **OS. Operating Systems**
- **NC. Net-Centric Computing**
- **PL. Programming Languages**
- **HC. Human-Computer Interaction**
- **GV. Graphics and Visual Computing**
- **IS. Intelligent Systems**
- **IM. Information Management**
- **SP. Social and Professional Issues**
- **SE. Software Engineering**
- **CN. Computational Science**

CC2001 Information Management Areas

IM1. Information models and systems*	IM8. Distributed DBs
IM2. Database systems*	IM9. Physical DB design
IM3. Data modeling*	IM10. Data mining
IM4. Relational DBs	IM11. Information storage and retrieval
IM5. Database query languages	IM12. Hypertext and hypermedia
IM6. Relational DB design	IM13. Multimedia information & systems
IM7. Transaction processing	IM14. Digital libraries

* Core components

DL Curriculum Framework


COURSE STRUCTURE
Semester 1: DL collections: development/creation
Semester 2: DL services and sustainability
CORE DL TOPICS
Digitization Storage Interchange
Metadata Cataloging Author submission
Architectures (agents, buses, wrappers/mediators) Interoperability
Naming Repositories Archives
Services (searching, linking, browsing, etc.)
Archiving and preservation Integrity
Digital objects Composites Packages
Spaces (conceptual, geographic, 2/3D, VR)
Architectures (agents, buses, wrappers/mediators) Interoperability
Intellectual property rights mgmt. Privacy Protection (watermarking)
RELATED TOPICS
Documents E-publishing Markup
Multimedia streams/structures Capture/representation Compression/coding
Thesauri Ontologies Classification Categorization
Info. Needs Relevance Evaluation Effectiveness
Routing Filtering Community filtering
Bibliographic information Bibliometrics Citations
Content-based analysis Multimedia indexing
Multimedia presentation, rendering
Search & search strategy Info seeking behavior User modeling Feedback
Info summarization Visualization

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

LIKES and 4 Needs of Others (www.LivingKnowledgeSociety.org)

1. Processes
   - Programs, algorithms, workflows, business processes, packages/toolkits, problem solving
2. Modeling, simulation
   - Analyze, abstract, connect, validate, predict, refine
3. Managing information
   - Data, information, and knowledge
   - PIM, create/represent/search/retrieve/reuse/…
4. Sensory connection, interaction
   - HCI, games, visualization, collaboration

AP CS Principles: Big Ideas

1. **Computing is a creative human activity that engenders innovation and promotes exploration.**
2. **Abstraction reduces information and detail to focus on concepts relevant to understanding and solving problems.**
3. **Data and information facilitate the creation of knowledge.**
4. **Algorithms are tools for developing and expressing solutions to computational problems.**

AP CS Principles: Big Ideas (2)

5. **Programming is a creative process that produces computational artifacts.**
6. **Digital devices, systems, and the networks that interconnect them enable and foster computational approaches to solving problems.**
7. **Computing enables innovation in other fields including science, social science, humanities, arts, medicine, engineering, and business.**

Outline

- An informal overview of digital libraries
- **An informal view of 5S**
- A formal perspective of 5S
- What has been done with 5S
- Future plans

5S Layers

Societies

Scenarios

Spaces

Structures

Streams

5S Contextualized

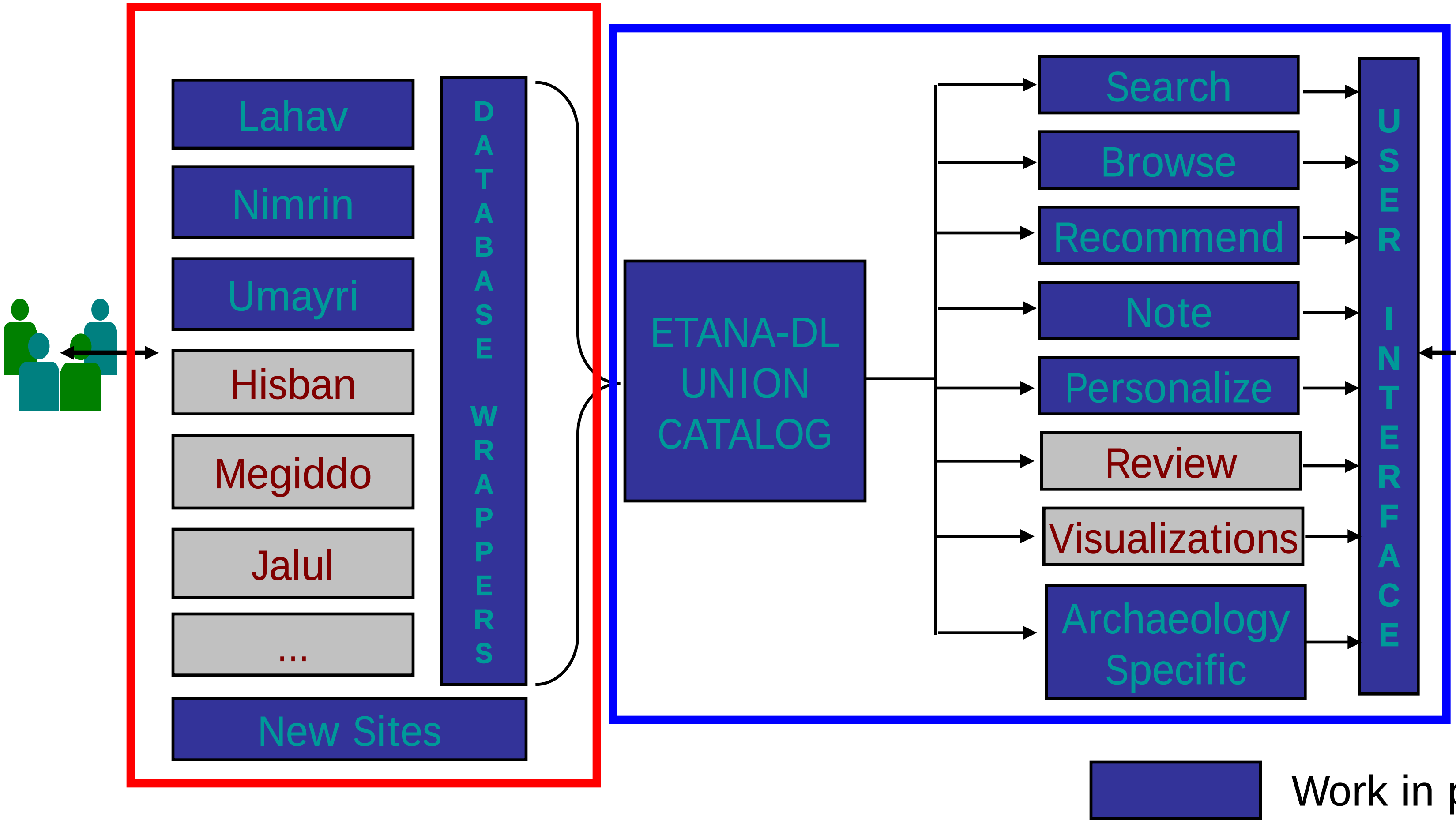
- Societies/communities/users served
- Scenarios/services supported
- Management of physical/conceptual/ feature spaces
- Use of structures/organizational devices
- Streams of content and communication

Informal 5S & DL Definitions

DLs are complex systems that

- help satisfy info needs of users (**societies**)
- provide info services (**scenarios**)
- organize info in usable ways (**structures**)
- present info in usable ways (**spaces**)
- communicate info with users (**streams**)

ETANA-DL Architecture
DigBase and DigKit



Lahav

Nimrin

Umayri

Hisban

Megiddo

Jalul

...

New Sites

DATABASE WRAPPERS

ETANA-DL UNION CATALOG

Search

Browse

Recommend

Note

Personalize

Review

Visualizations

Archaeology Specific

USER INTERFACE

🟦 Work in progress

ETANA Societies

1. Historic and pre-historic societies (being studied)
2. Archaeologists (in academic institutes, fieldwork settings, or local and national governmental bodies)
3. Project directors
4. Technical staff (consisting of photographers, technical illustrators, and their assistants)
5. Field staff (responsible for the actual work of excavation)
6. Camp staff (e.g., camp managers, registrars, tool stewards)
7. General public (e.g., educators, learners, citizens)

ETANA Societies

- Social issues

1. Who owns the finds?
2. Where should they be preserved?
3. What nationality and ethnicity do they represent?
4. Who has publication rights?
5. What interactions took place between those at the site studied, and others? What theories are proposed by whom about this?

ETANA Scenarios

1. Life in the site in former times
2. Digital recording: the planning stage and the excavation stage
3. Planning stage: remote sensing, fieldwalking, field surveys, building surveys, consulting historical and other documentary sources, and managing the sites and monuments
4. Excavation
    1. Detailed information is recorded, including for each layer of soil, and for features such as pole holes, pits, and ditches.
    2. Data about each artifact is recorded together with information about its exact find spot.
    3. Numerous environmental and other samples are taken for laboratory analysis, and the location and purpose of each is carefully recorded.
    4. Large numbers of photographs are taken, both general views of the progress of excavation and detailed shots showing the contexts of finds.
5. Organization and storage of material
6. Analysis and hypotheses generation and testing
7. Publications, museum displays
8. Information services for the general public

ETANA Spaces

1. Geographic distribution of found artifacts
2. Temporal dimension (as inferred by archaeologists)
3. Metric or vector spaces
    1. used to support retrieval operations, and to calculate distance (and similarity)
    2. used to browse / constrain searches spatially
4. 3D models of the past, used to reconstruct and visualize archaeological ruins
5. 2D interfaces for human-computer interaction

ETANA Structures

1. Site Organization
   1. Region, site, partition, sub-partition, locus, …
2. Temporal orderings (ages, periods)
3. Taxonomies
   1. for bones, seeds, building materials, …
4. Stratigraphic relationships
   1. above, beneath, coexistent

ETANA Streams

1. successive photos and drawings of excavation sites, loci, unearthed artifacts
2. audio and video recordings of excavation activities and discussions
3. textual reports
4. 3D models used to reconstruct and visualize archaeological ruins.

Ensemble in 5 S - Societies

- What *Societies* must Ensemble serve?
  - Teachers
  - Students, perhaps
  - Groups with computing education tasks
  - The NSDL
  - The NSF
  - Partner sites (providers and harvesters)
  - The developers
  - Related hardware / software components

Ensemble in 5S - Scenarios

- What *Scenarios* must be addressed? (a sample)
  - Search, Browse
  - User registration, login
  - Commenting, rating, tagging
  - Acquisition/de-acquisition/user contributing
  - Share resources in, and collect data from, other places (CiteULike, Facebook)
  - Acknowledge contributions
  - Harvest and be harvested
  - Join groups, participate in discussions
  - Recover from failures
    - Computer systems, storage

Ensemble in 5S - Spaces

- What *Spaces* will matter in Ensemble?
  - User interface (2D generally, 3D in Second Life)
  - Education level
  - Curriculum standards or recommendations
  - Topic spaces
  - Vector and feature spaces to support indexing, searching, and classifying

Ensemble in 5S - Structures

- What *Structures* will we hold?
  - Metadata
  - Computing Ontology
  - Database schema and tables
  - Taxonomies
    - Educational schema
    - Computing topics (Knowledge units)
    - Rating schemes

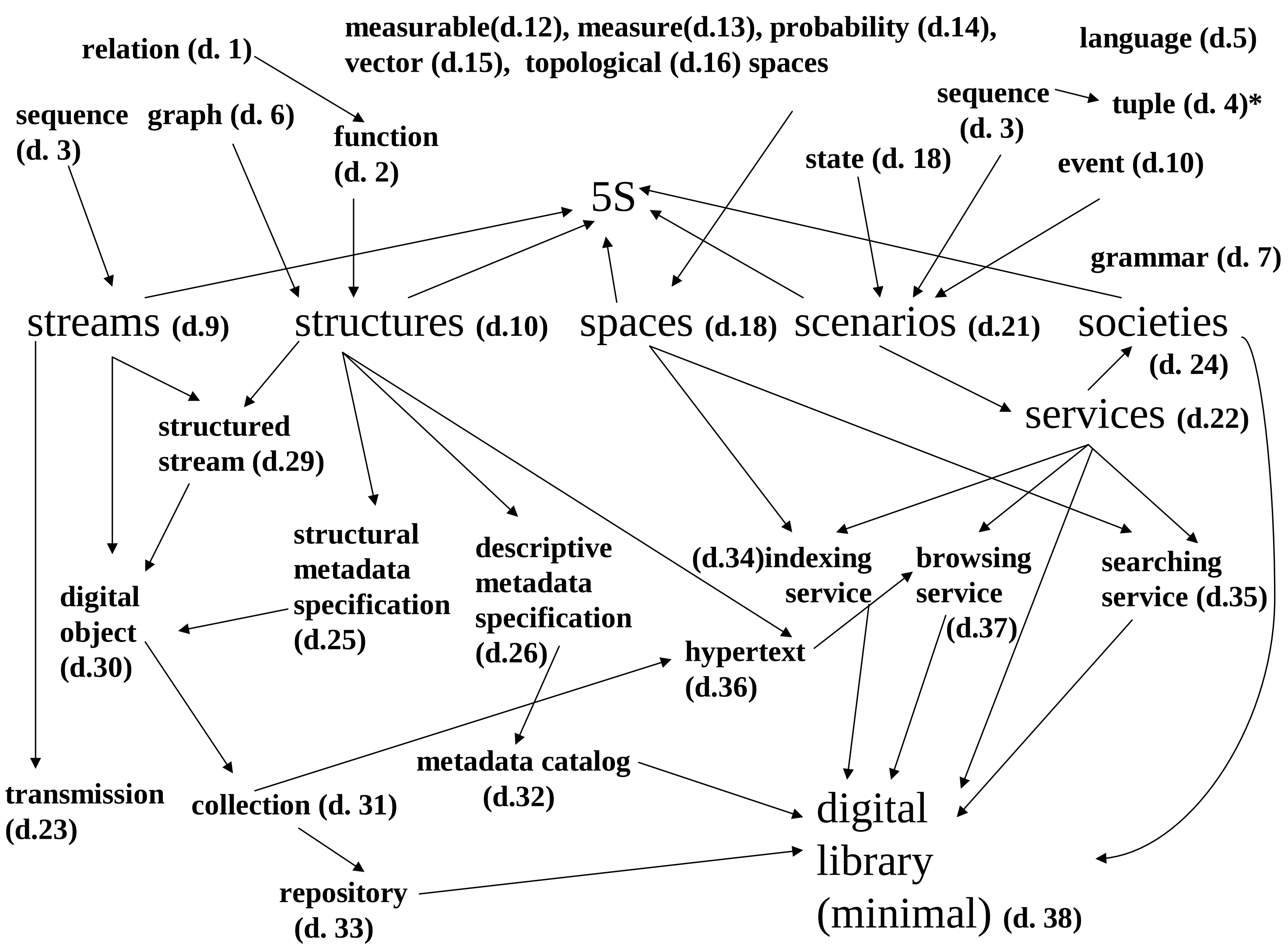
Ensemble in 5 S - Streams

- What *Streams* of data will we see?
  - All the document types we can imagine: text, word processor, PDF, spreadsheets, presentations, HTML, XML, …
  - All the image types, all the video types
    - Images (jpg, tiff, …)
    - Video (avi, mov, …)
  - Program code, both source code and object code
  - Comments, ratings, tags
  - Group membership profiles
  - E-mail addresses
  - User information (preferences, …)

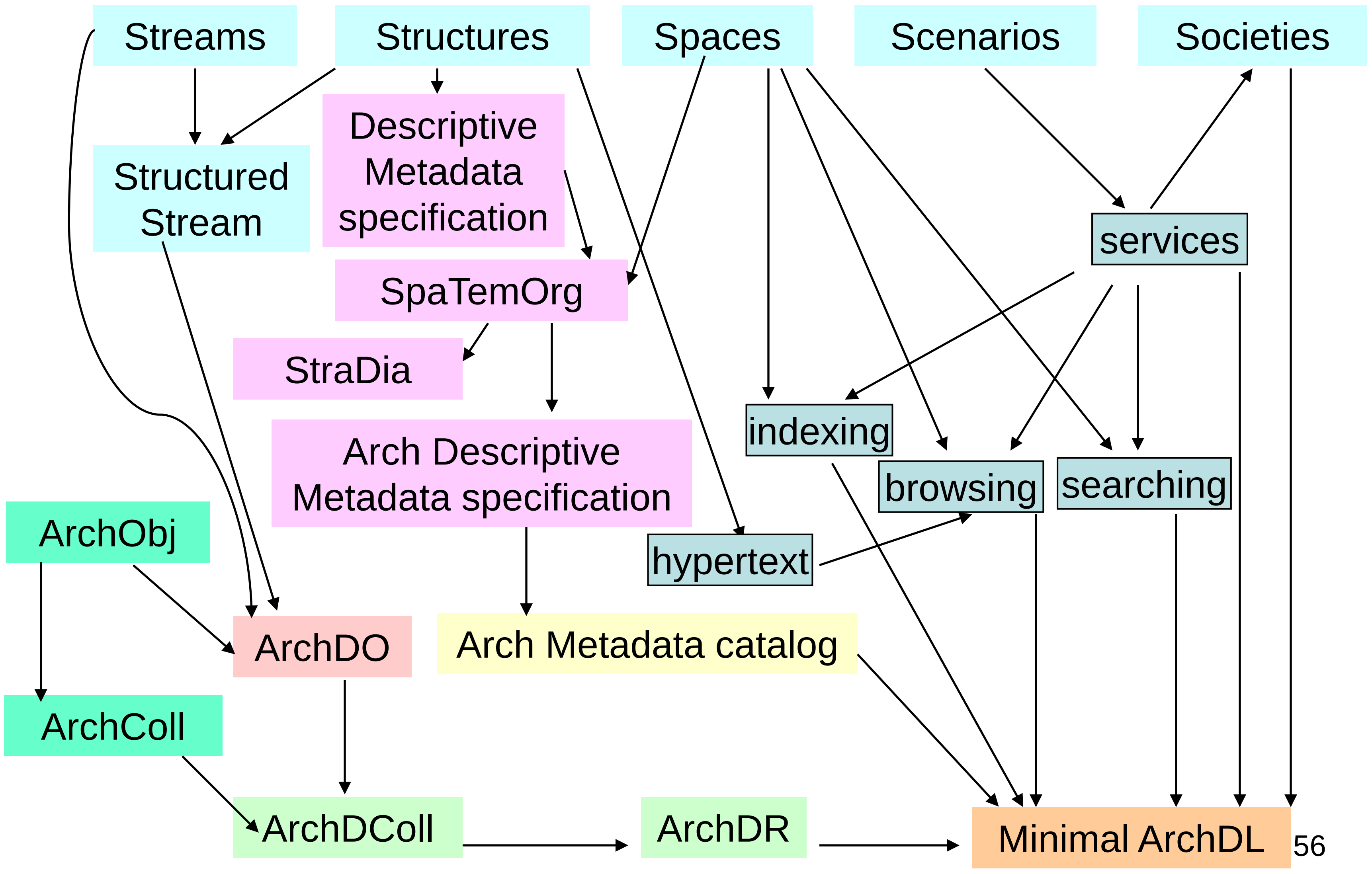
Outline

- An informal overview of digital libraries
- An informal view of 5S
- **A formal perspective of 5S**
- What has been done with 5S
- Future plans

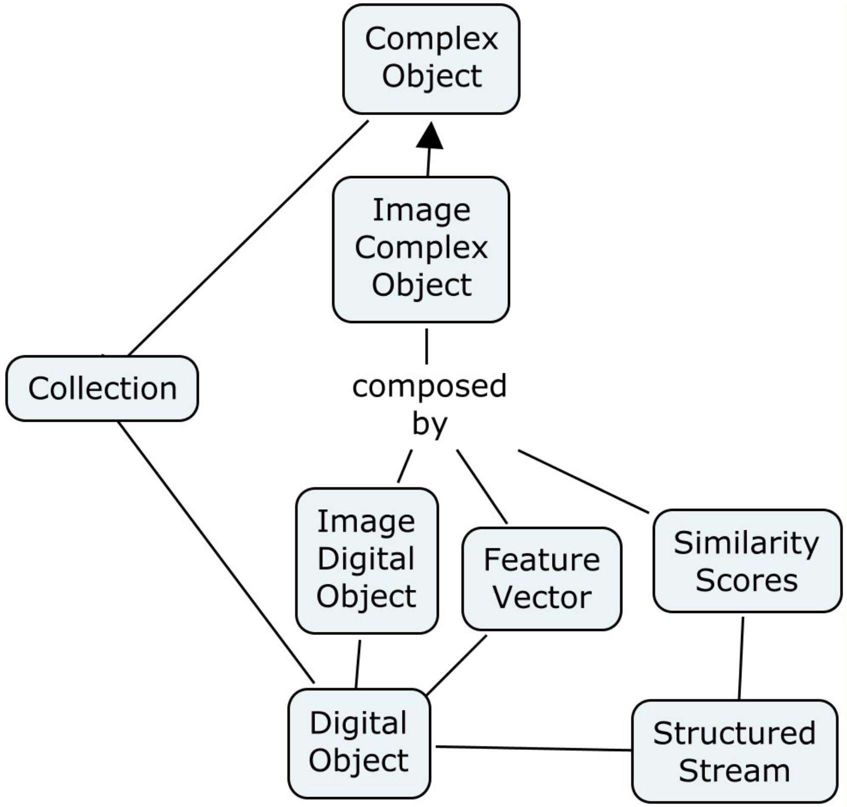
5S and DL formal definitions and compositions (April 2004 TOIS)



A Minimal ArchDL in the 5S Framework



5S Extensions to Complex Object



From [Murthy et al. 2010], a Complex object is a tuple cdo=(h, SCDO=DO *U* SM, S), where:

- h є H, where H is a set of universally unique handles (lables);
- DO={do^1,do^2,...,don}, where doi is a digital object;
- SM={sm^1,sm^2,...smn} is a set of streams;
- S is a structure that composes the cdo into its parts in SCDO.

5S Extensions to CBIR

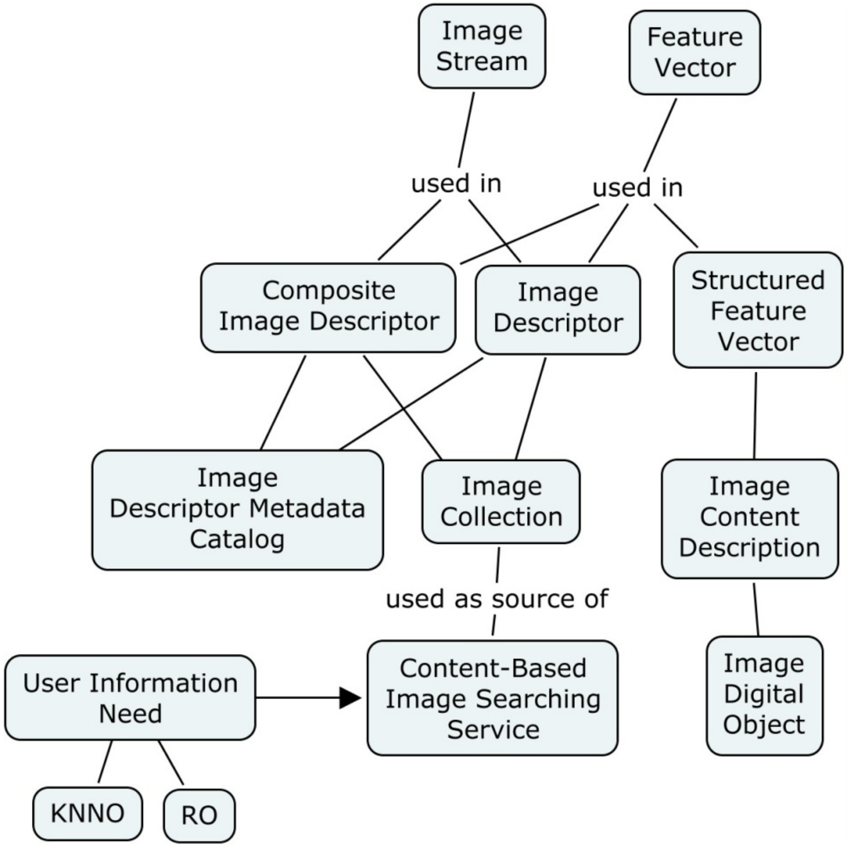


Image Stream
Feature Vector
used in
used in
Composite Image Descriptor
Image Descriptor
Structured Feature Vector
Image Descriptor Metadata Catalog
Image Collection
Image Content Description
used as source of
User Information Need
Content-Based Image Searching Service
Image Digital Object
KNNO
RO

Uma Murthy's Dissertation, esp. Ch. 8

- SIMPEL
- Enhanced CMapTools

- Case studies of SI use (music & fisheries)
- Pilot user

SI-DI metamodel - v1

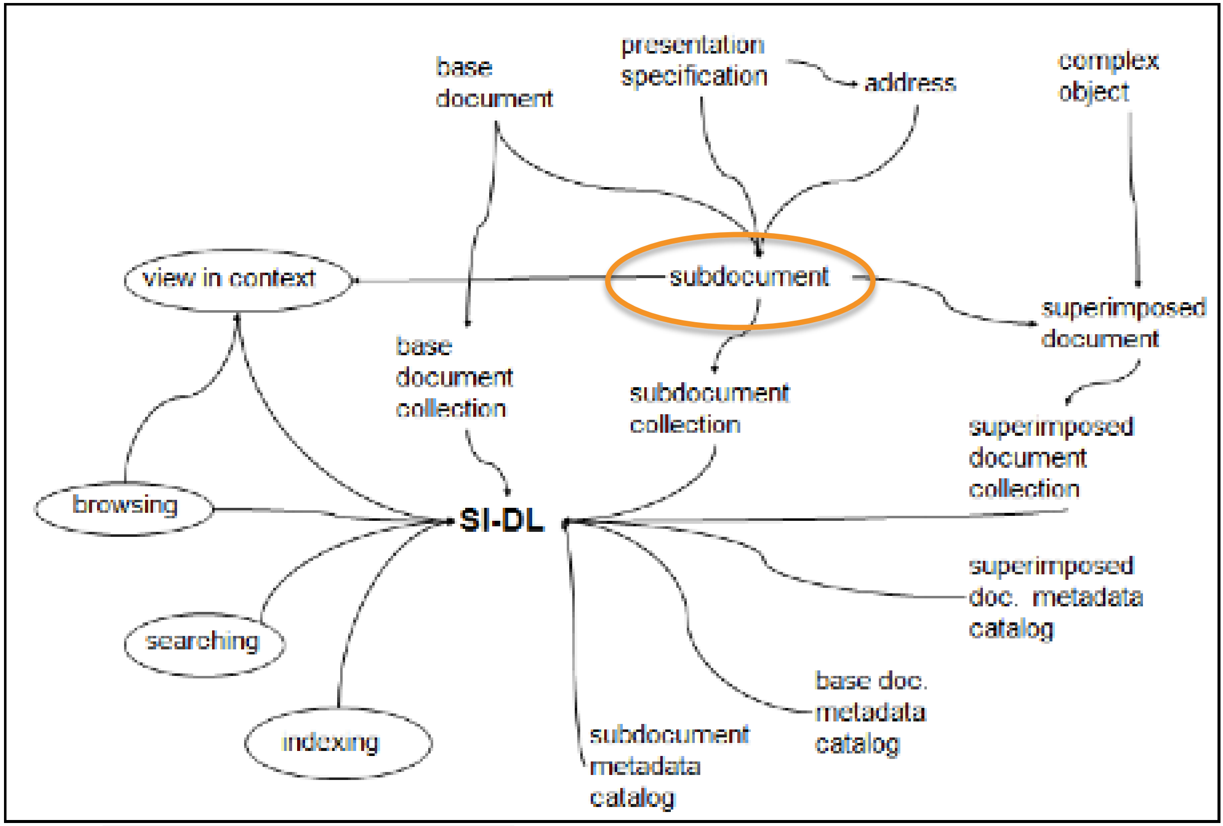
- 5S analysis of SI
- Initial metamodel

SuperIDR-classroom study

- Fish id. – learning and identification
- SuperIDR improves on traditional methods
- Insufficient data on *how* SuperIDR was used

SI-DI metamodel – v2

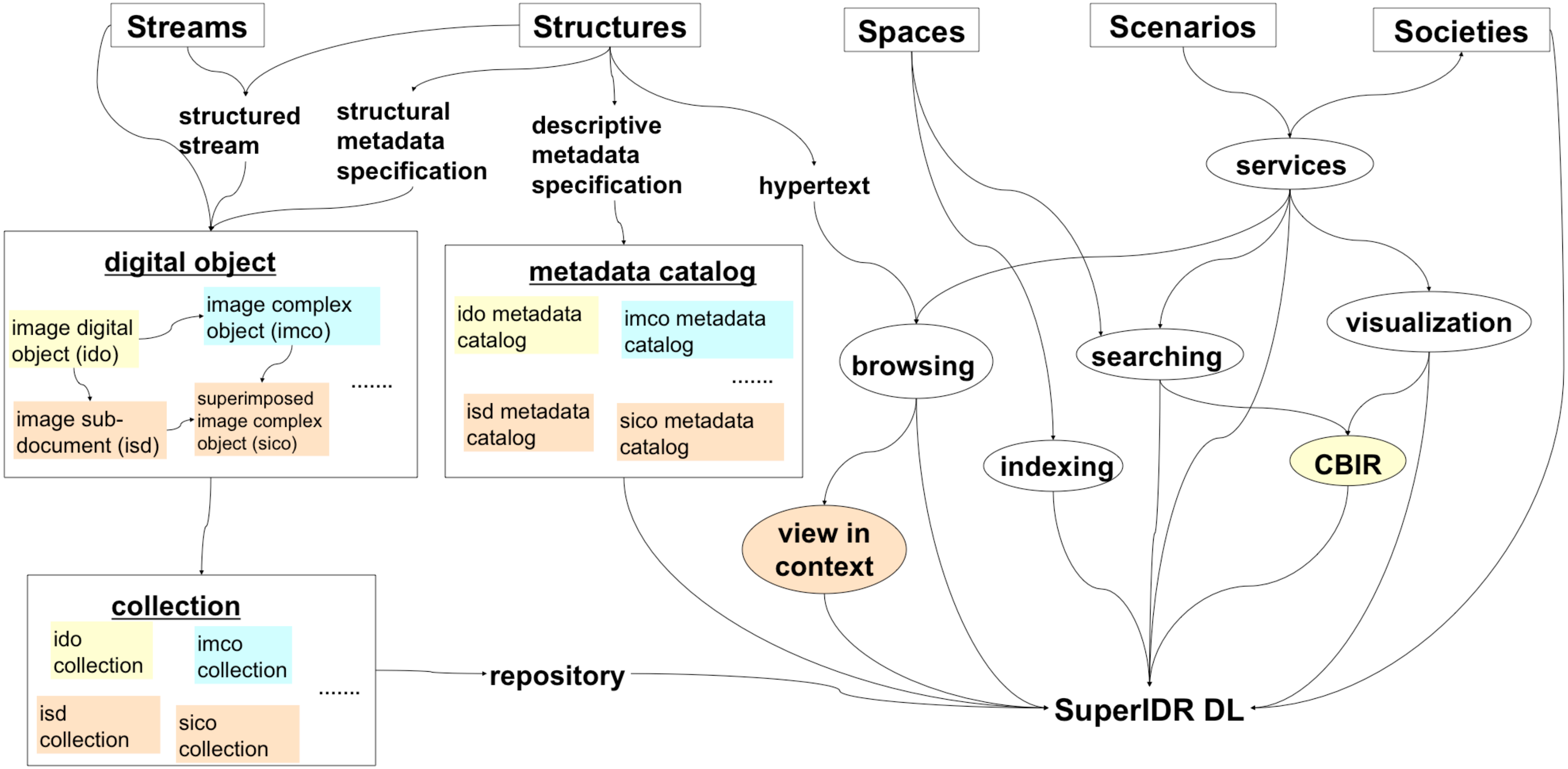
- Improved metamodel
- Case study

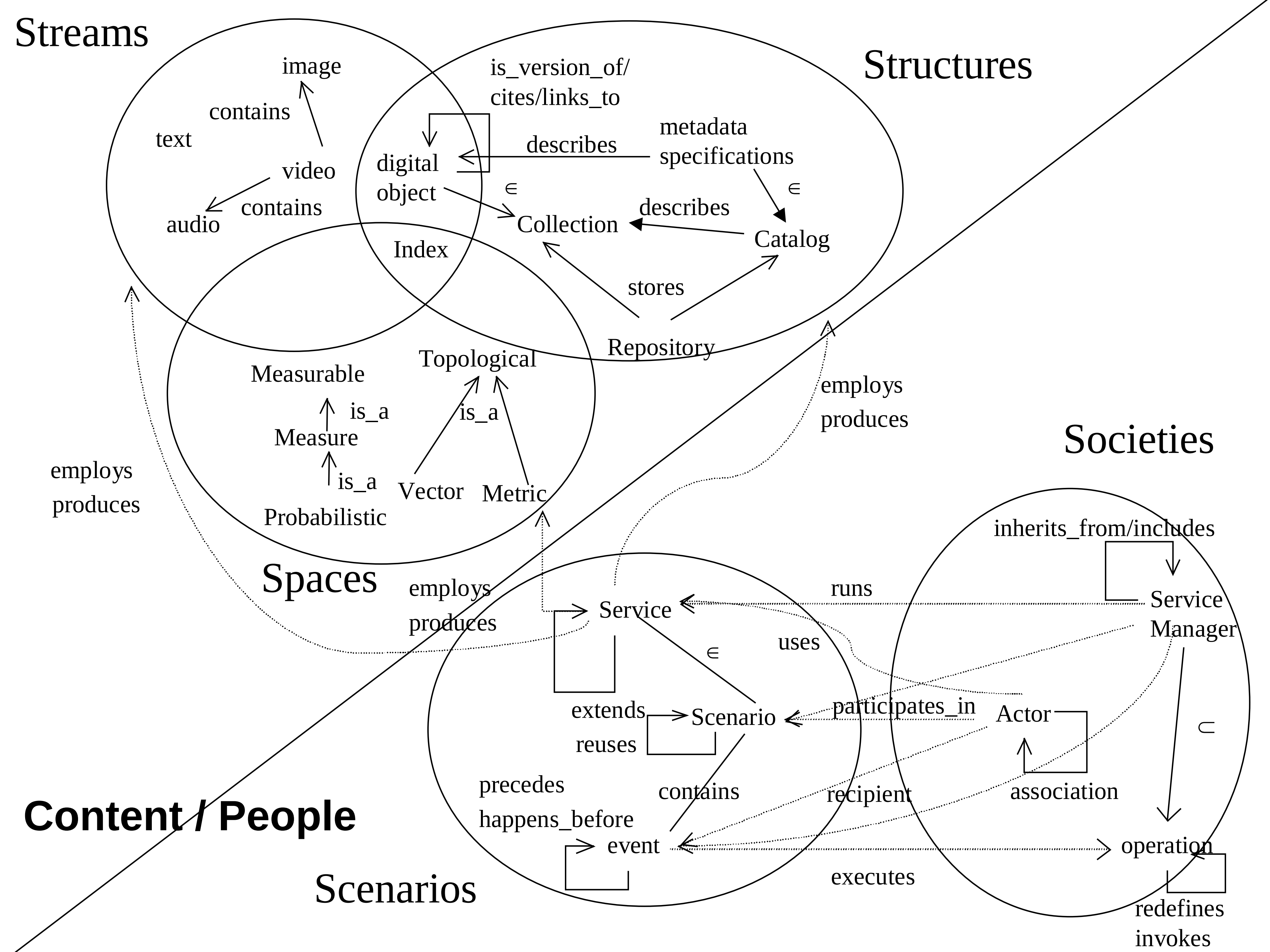


es

improvements

Definitional dependencies among concepts in the SuperIDR SI-DL





Streams

image

contains

text

video

audio

contains

Structures

is_version_of/
cites/links_to

metadata
specifications

describes

digital
object

∈

Collection

describes

∈

Catalog

Index

stores

Repository

Topological

Measurable

is_a

is_a

Measure

is_a

Vector

Metric

Probabilistic

Spaces

employs
produces

employs
produces

employs
produces

Societies

inherits_from/includes

Service
Manager

runs

Service

uses

∈

extends

Scenario

reuses

participates_in

Actor

association

⊂

precedes
happens_before

contains

recipient

event

operation

executes

redefines
invokes

Content / People

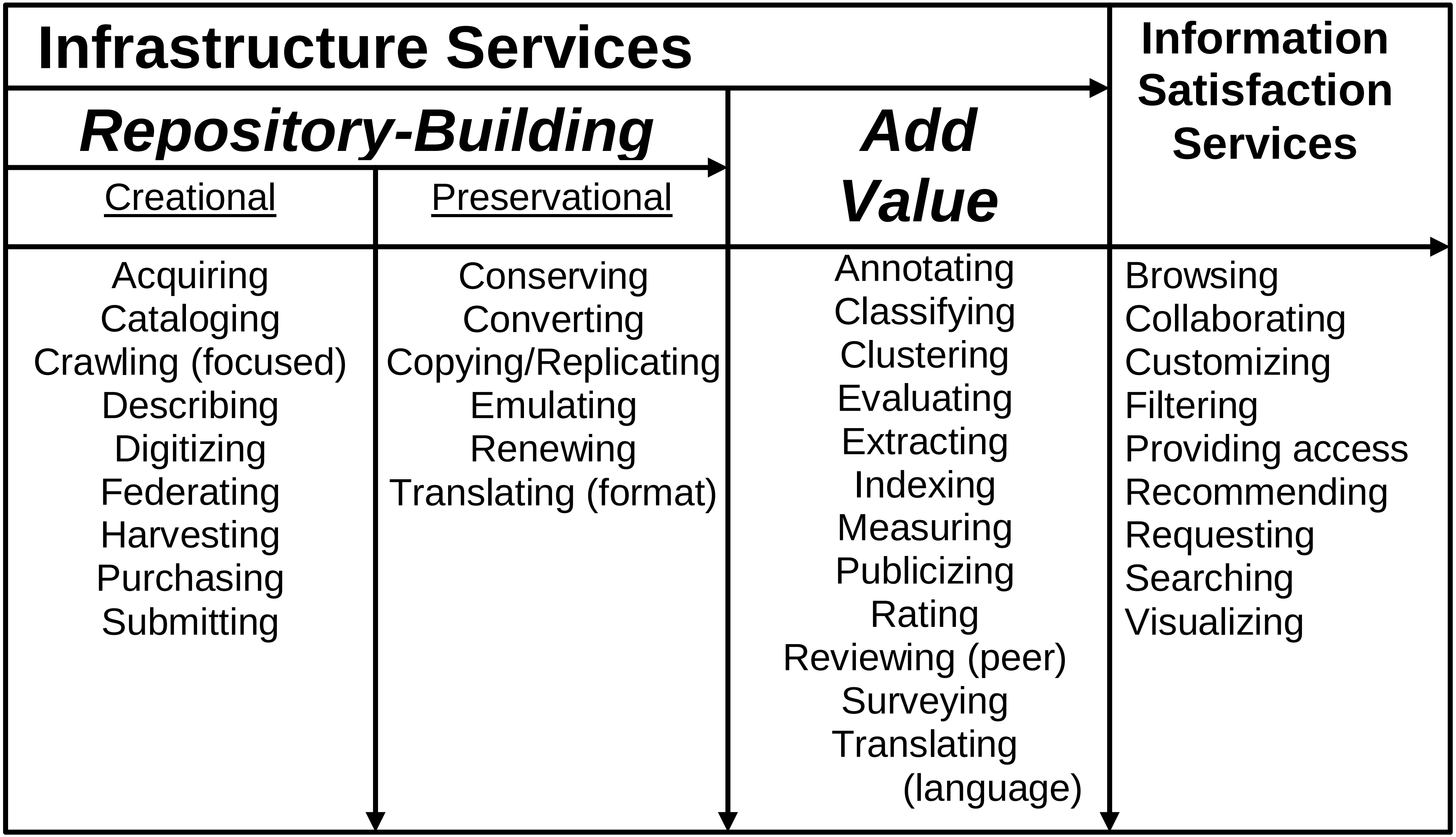
Scenarios

Extending 5S

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

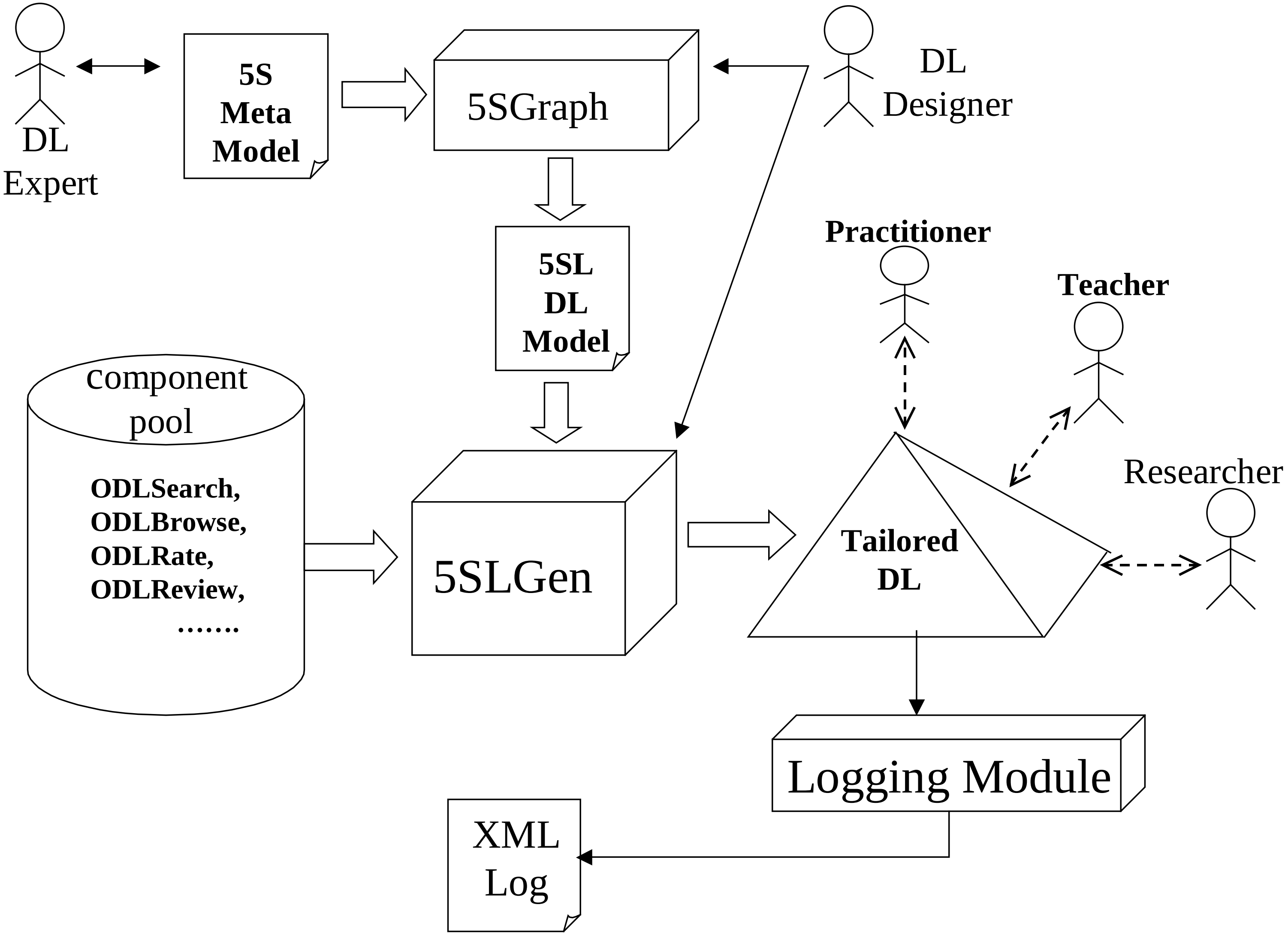
Outline

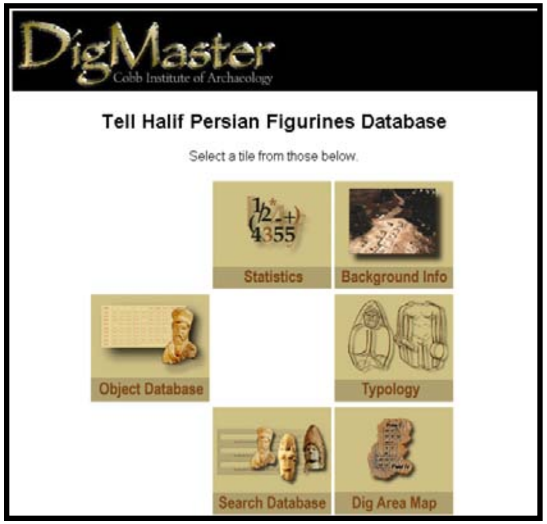
- An informal overview of digital libraries
- An informal view of 5S
- A formal perspective of 5S
- **What has been done with 5S**
- Future plans



Infrastructure Services			**Information Satisfaction Services**
Repository-Building		***Add Value***	
Creational	Preservational		
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

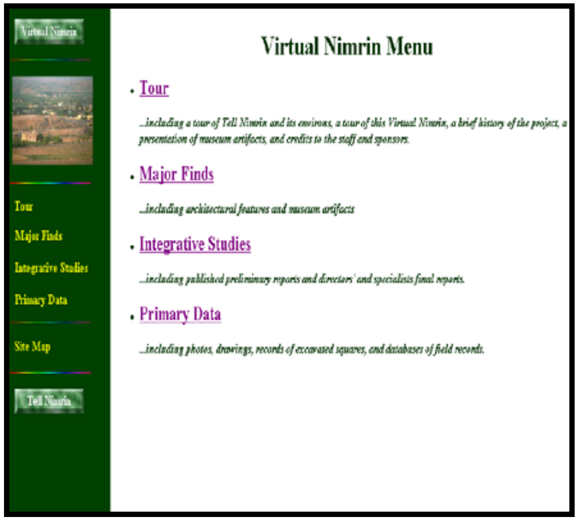
Tools/Applications





Tell Halif Persian Figurines Database

Select a tile from those below.



Virtual Nimrin Menu

- Tour

  ...including a tour of Tell Nimrin and its environs, a tour of this Virtual Nimrin, a brief history of the project, a presentation of museum artifacts, and credits to the staff and sponsors.

- Major Finds

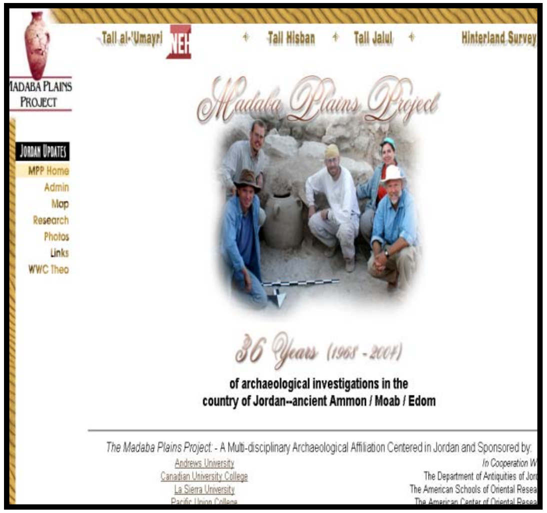
  ...including architectural features and museum artifacts

- Integrative Studies

  ...including published preliminary reports and directors' and specialists final reports.

- Primary Data

  ...including photos, drawings, records of excavated squares, and databases of field records.



36 Years (1968 - 2004)

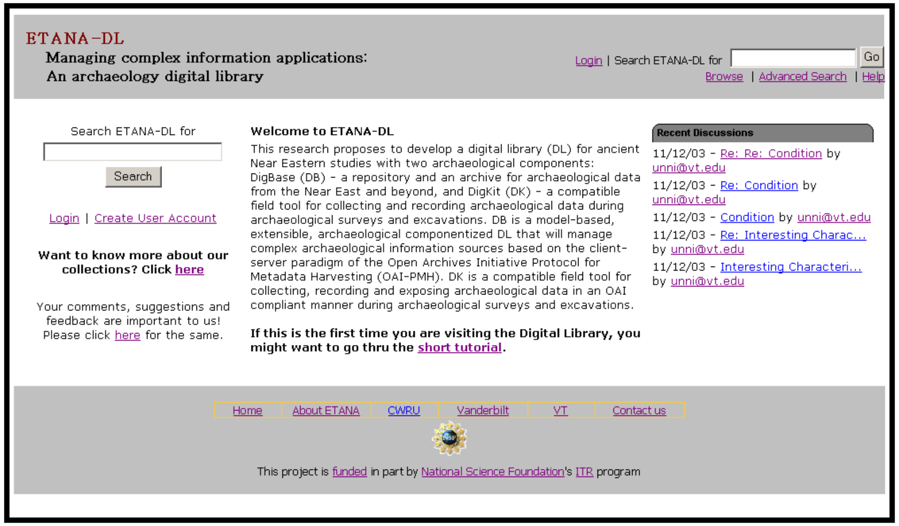
of archaeological investigations in the
country of Jordan--ancient Ammon / Moab / Edom

The Madaba Plains Project: - A Multi-disciplinary Archaeological Affiliation Centered in Jordan and Sponsored by:

Andrews University
Canadian University College
La Sierra University
Pacific Union College

In Cooperation W
The Department of Antiquities of Jord
The American Schools of Oriental Resea
The American Center of Oriental Resea

ETANA.org


ETANA-DL

Managing complex information applications:
An archaeology digital library

Login | Search ETANA-DL for
Go
Browse | Advanced Search | Help

Search ETANA-DL for

Search
Login | Create User Account

Want to know more about our collections? Click here

Your comments, suggestions and feedback are important to us! Please click here for the same.

Welcome to ETANA-DL

This research proposes to develop a digital library (DL) for ancient Near Eastern studies with two archaeological components: DigBase (DB) - a repository and an archive for archaeological data from the Near East and beyond, and DigKit (DK) - a compatible field tool for collecting and recording archaeological data during archaeological surveys and excavations. DB is a model-based, extensible, archaeological componentized DL that will manage complex archaeological information sources based on the client-server paradigm of the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH). DK is a compatible field tool for collecting, recording and exposing archaeological data in an OAI compliant manner during archaeological surveys and excavations.

If this is the first time you are visiting the Digital Library, you might want to go thru the short tutorial.

Recent Discussions

11/12/03 - Re: Re: Condition by unni@vt.edu

11/12/03 - Re: Condition by unni@vt.edu

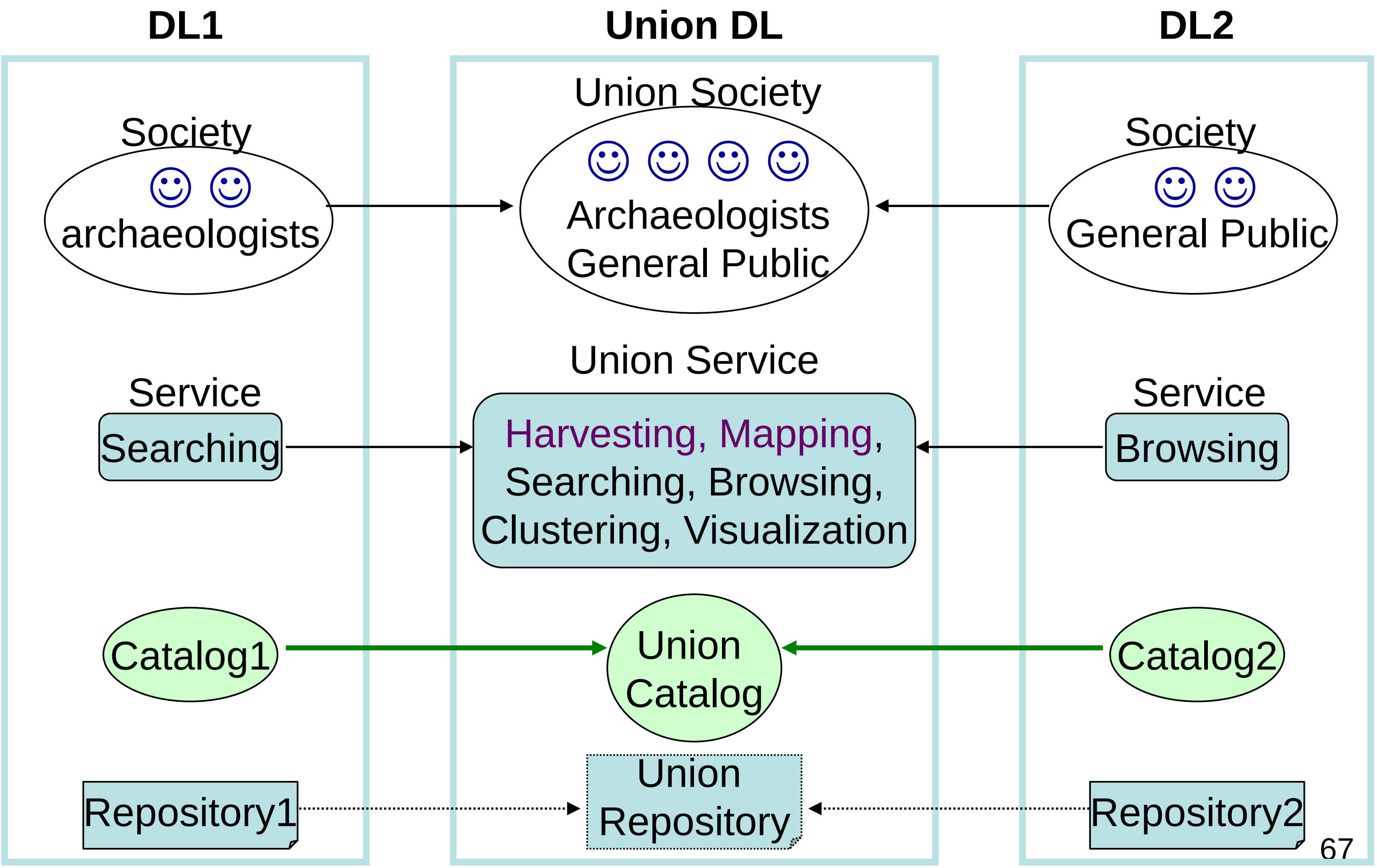
11/12/03 - Condition by unni@vt.edu

11/12/03 - Re: Interesting Charac... by unni@vt.edu

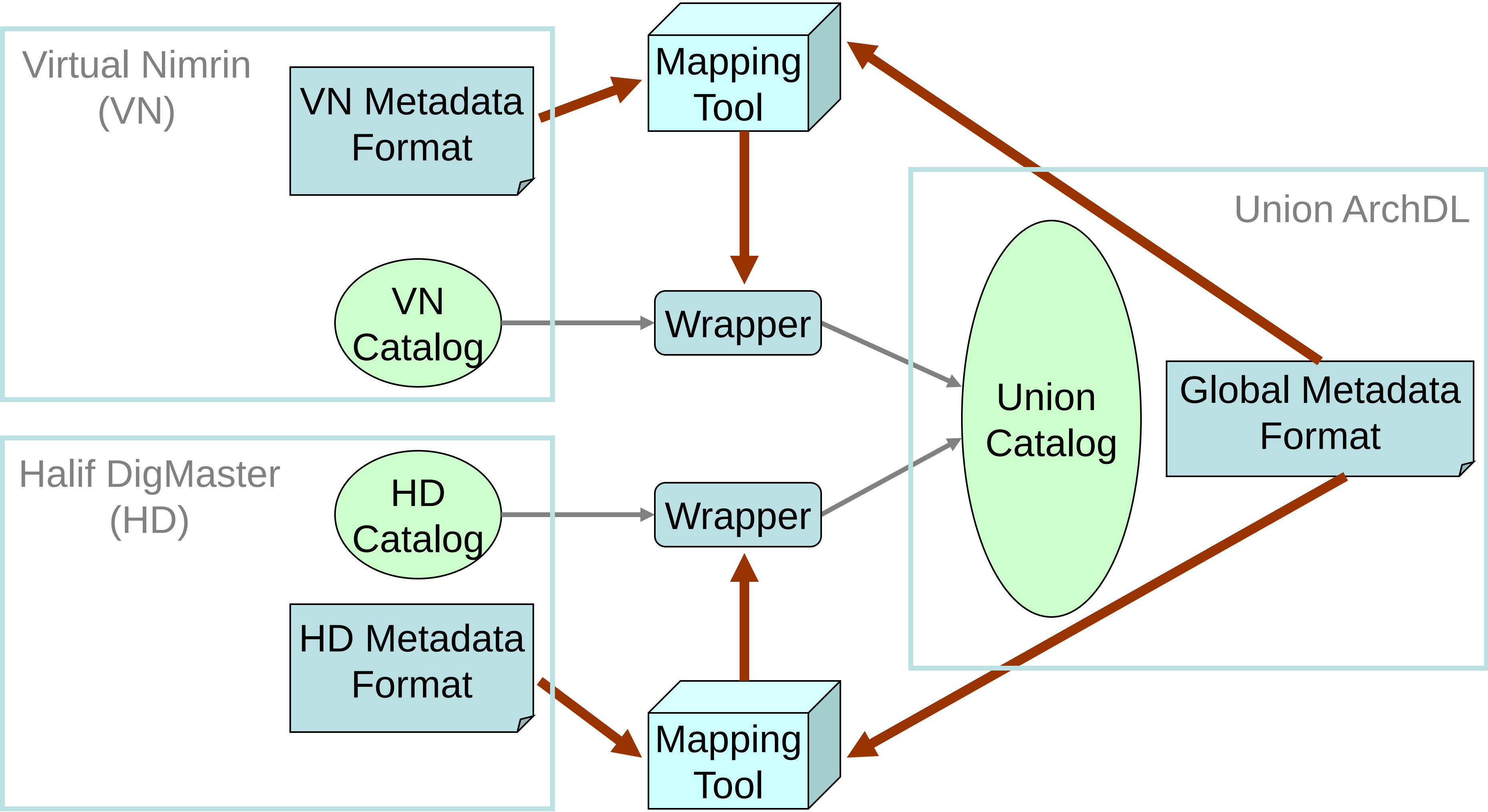
11/12/03 - Interesting Characteri... by unni@vt.edu

This project is funded in part by National Science Foundation's ITR program

Architecture of a Union DL (ETANA.org)



Union Catalog Integration



Outline

- An informal overview of digital libraries
- An informal view of 5S
- A formal perspective of 5S
- What has been done with 5S
- **Future plans**

Digital Libraries --- Objectives

- World Lit.: 24hr / 7day / from desktop
- Integrated “super” information systems: 5S: Table of related areas and their coverage
- Ubiquitous, Higher Quality, Lower Cost
- Education, Knowledge Sharing, Discovery
- Disintermediation -> Collaboration
- Universities Reclaim Property
- Interactive Courseware, Student Works
- Scalable, Sustainable, Usable, Useful

DL Overview
Why of Global Interest?

- **National projects** can preserve antiquities and heritage: cultural, historical, linguistic, scholarly
- Knowledge and information are essential to economic and technological **growth, education**
- DL - a **domain for international collaboration**
  - wherein all can **contribute** and **benefit**
  - which leverages investment in **networking**
  - which provides useful **content** on Internet & WWW
  - which will **tie nations and peoples together** more strongly and through **deeper understanding**

As data, information, and knowledge play increasingly central roles … digital library research should focus on:

- Increasing the scope and scale of information resources and services;
- Employing context at the individual, community, and societal levels to improve performance;
- Developing algorithms and strategies for transforming data into actionable information;
- Demonstrating the integration of information spaces into everyday life; and
- Improving availability, accessibility, and, thereby, productivity. (Chatham Workshop)

Questions?

Ask: fox@vt.edu. Feel free to visit Blacksburg, VA