5S and the Reference Model

Edward A. Fox
Virginia Tech, Dept. of CS
fox@vt.edu
http://fox.cs.vt.edu

DELOS Reference Model Workshop
Frascati-Rome, June 1-2, 2006

Acknowledgements (selected)

- **5S Helpers:** Weiguo Fan, Marcos Gonçalves, Doug Gorton, Rohit Kelapure, Neill Kipp, Uma Murthy, Ananth Raghavan, Rao Shen, Hussein Suleman, Srinivas, Vemuri, Layne Watson, …

- **Sponsors:** ACM, AOL, CAPES, DFG, IBM, Microsoft, NSF (IIS-9986089, 0086227, 0080748, 0325579, 0535057, 0535060; ITR-0325579; DUE-0121679, 0136690, 0121741, 0333601), SUN

Outline

- **Key Publications**
- 5S Approach
- Modeling and Implementation
- Integration
- Quality
- Challenges for the Reference Model

Doctoral Dissertations

- April 2006, PhD dissertation, Rao Shen, "Applying the 5S Framework To Integrating Digital Libraries", http://scholar.lib.vt.edu/theses/available/etd-04212006-135018/
- Nov. 2004, PhD dissertation, Marcos Andre Goncalves, "Streams, Structures, Spaces,Scenarios, and Societies (5S): A Formal Digital Library Framework and Its Applications", http://scholar.lib.vt.edu/theses/available/etd-12052004-135923/
- Nov. 2002, PhD dissertation, Hussein Suleman, "Open Digital Libraries", http://scholar.lib.vt.edu/theses/available/etd-11222002-155624/

See http://scholar.lib.vt.edu or www.ndltd.org

Masters Theses / Reports

- May 2005, MS thesis, Ananth Raghavan, Schema Mapper: A Visualization Tool for Incremental Semi-automatic Mapping-based Integration of Heterogeneous Collections into Archaeological Digital Libraries: The ETANA-DL Case Study
- April 2004, Unnikrishnan Ravindranathan, Prototyping Digital Libraries Handling Heterogeneous Data Sources - An ETANA-DL Case Study
- June 2003, MS thesis, Rohit Dilip Kelapure, Scenario-Based Generation of Digital Library Services
- May 2003, MS independent study report, Ganesh K. Panchanathan, "Digital library logging and analysis using XML
- Nov. 2002, MS thesis, Qinwei Zhu, "5SGraph: A Modeling Tool for Digital Libraries
- May 2002, MS thesis, Jun Wang, VIDI: A Lightweight Protocol Between Visualization Systems and Digital Libraries

Some Other Notable Papers

- M. Goncalves, E. Fox, L. Watson, N. Kipp. Streams, Structures, Spaces, Scenarios, Societies (5S): A Formal Model for Digital Libraries. ACM Trans. Information Systems, April 2004, 22(2): 270-312
- Rao Shen, Naga Srinivas Vemuri, Weiguo Fan, and Edward A. Fox. What is a Successful Digital Library? ECDL 2006, Alicante, Spain, Sept. 17-21, 2006
- Rao Shen, Naga Srinivas Vemuri, Weiguo Fan, Ricardo da S. Torres and Edward A. Fox. Exploring Digital Libraries: Integrating Browsing, Searching, and Visualization. JCDL 2006, June 11-15, 2006, Chapel Hill, NC
- Naga Srinivas Vemuri, Rao Shen, Sameer Tupe, Weiguo Fan and Edward A. Fox. ETANA-ADD: An Interactive Tool for Integrating Archaeological DL Collections. JCDL 2006, June 11-15, 2006, Chapel Hill, NC
- Jeffrey Pomerantz, Barbara M. Wildemuth, Seungwon Yang and Edward A. Fox. Curriculum Development for Digital Libraries. JCDL 2006, June 11-15, 2006, Chapel Hill, NC
- Ananth Raghavan, Naga Srinivas Vemuri, Rao Shen, Marcos Andre Goncalves, Weiguo Fan, and Edward A. Fox. Incremental, Semi-automatic, Mapping-Based Integration of Heterogeneous Collections into Archaeological Digital Libraries: Megiddo Case Study. ECDL2005, Vienna, Sept. 18-23, 139-150
- Rao Shen, Marcos Andre Goncalves, Weiguo Fan, and Edward A. Fox. Requirements Gathering and Modeling of Domain-Specific Digital Libraries with the 5S Framework: An Archaeological Case Study with ETANA. In Proceedings ECDL2005, Vienna, Sept. 18-23, 1-12
- M. A. Goncalves, L. T. Watson, and E. A. Fox. Towards a Digital Library Theory: A Formal Digital Library Ontology. In Mathematical Formal Methods workshop, SIGIR 2004, July 29, Sheffield, England

Outline

- Key Publications
- **5S Approach**
- Modeling and Implementation
- Integration
- Quality
- Challenges for the Reference Model

Start with Minimal Digital Libraries

- Key concepts, core ideas
- Minimalist perspective
- Underlying concepts: 5S (ETANA example)
- Higher DL constructs
- Bases:
  - Literature
  - Informal explanations
  - Formal definitions

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**)

Example of 5Ss: ETANA-DL

- Archaeological DL (Electronic Tools for Ancient Near Eastern Archaeology Digital Library)
- Integrated DL
  - Heterogeneous data handling
- Applies and extends the OAI-PMH
  - Open Archives Initiative Protocol for Metadata Handling
- Design considerations
  - Componentized
  - Extensible
  - Portable
  - Work based on 5S framework

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 (e.g., structured stream = DO or protocol), profiles, logs, P2P network, services
Spaces	Measure, measurable, topological, vector, probabilistic	Defines logical and presentational views of several DL components; host and user locations; GIS
Scenarios	Searching, browsing, recommending	Details the behavior of DL services, workflows, life cycle, preservation
Societies	Service managers, learners, teachers, etc.	Defines managers, responsible for running DL services; actors, that use those services; and relationships among them (including policies)

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 – cont'd

- 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

Foundations for Information Systems: Digital Libraries and the 5S Framework

- Ch. 1. Introduction (Motivation, Synopsis)
- Part 1 – The “Ss”
- Part 2 – Higher DL Constructs
- Part 3 – Advanced Topics
- Appendix

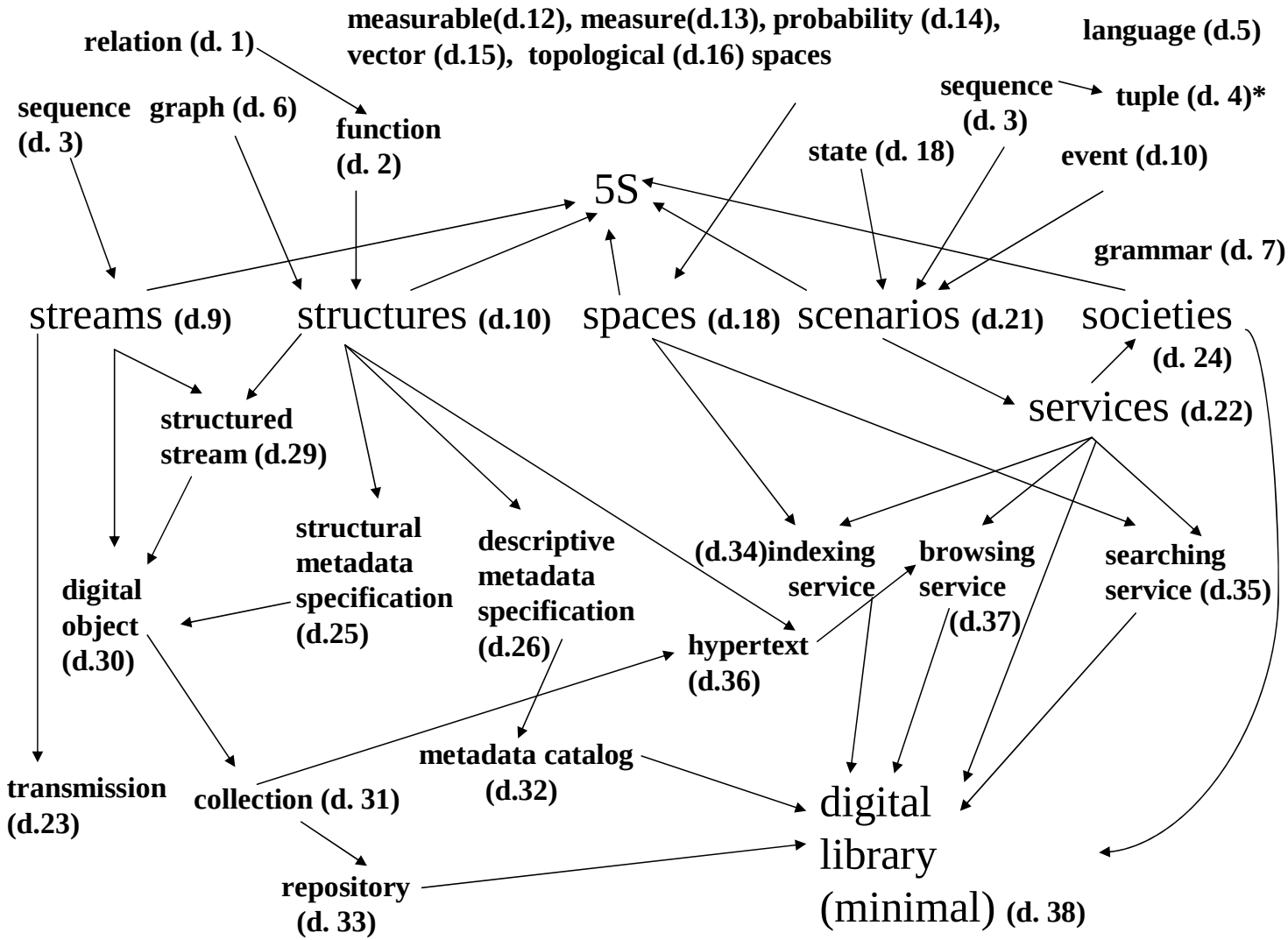
Book Parts and Chapters - 2

- Part 2 – Higher DL Constructs
  - Ch. 7: Collections
  - Ch. 8: Catalogs
  - Ch. 9: Repositories and Archives
  - Ch. 10: Services
  - Ch. 11: Systems
  - Ch. 12: Case Studies

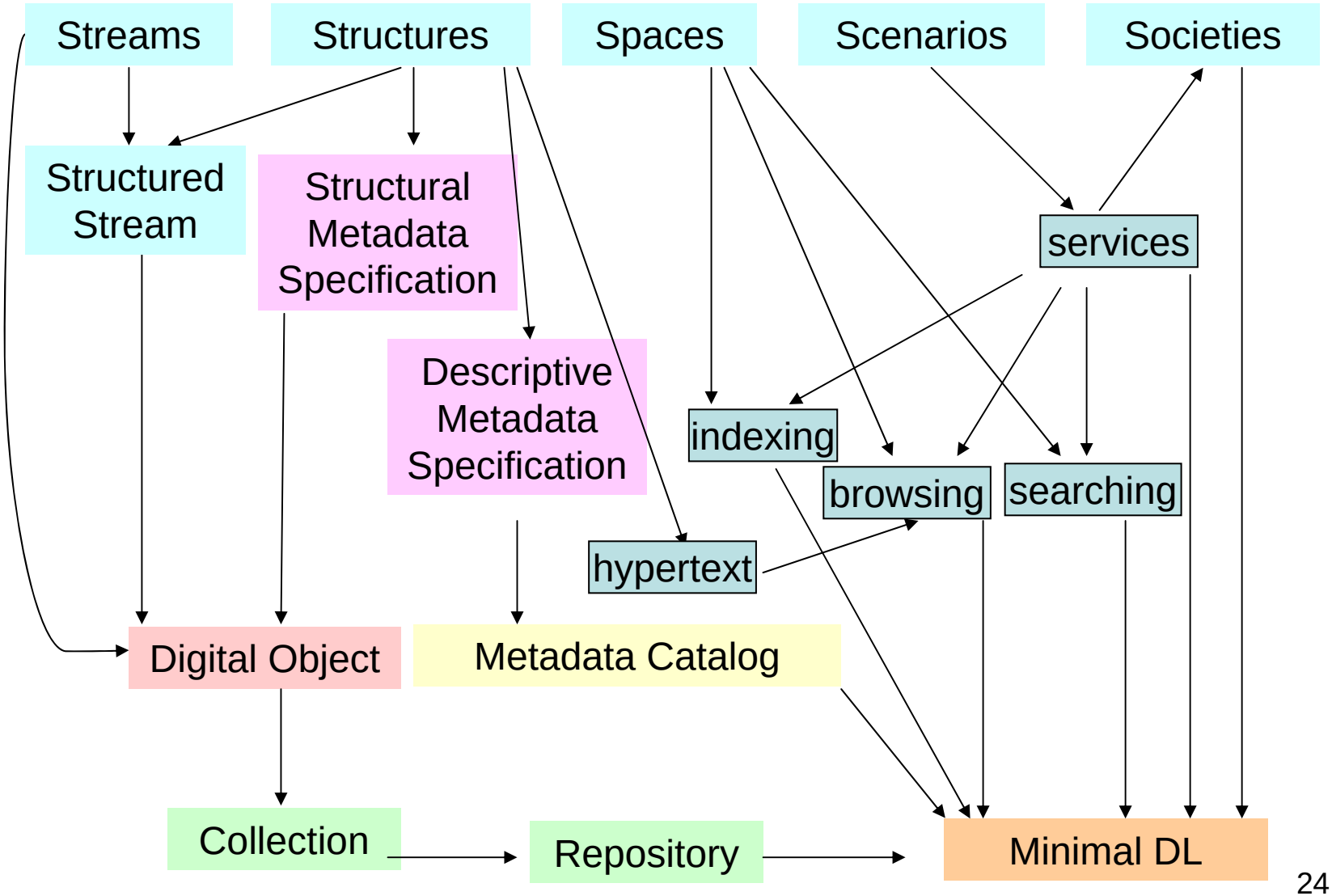
Book Parts and Chapters - 3

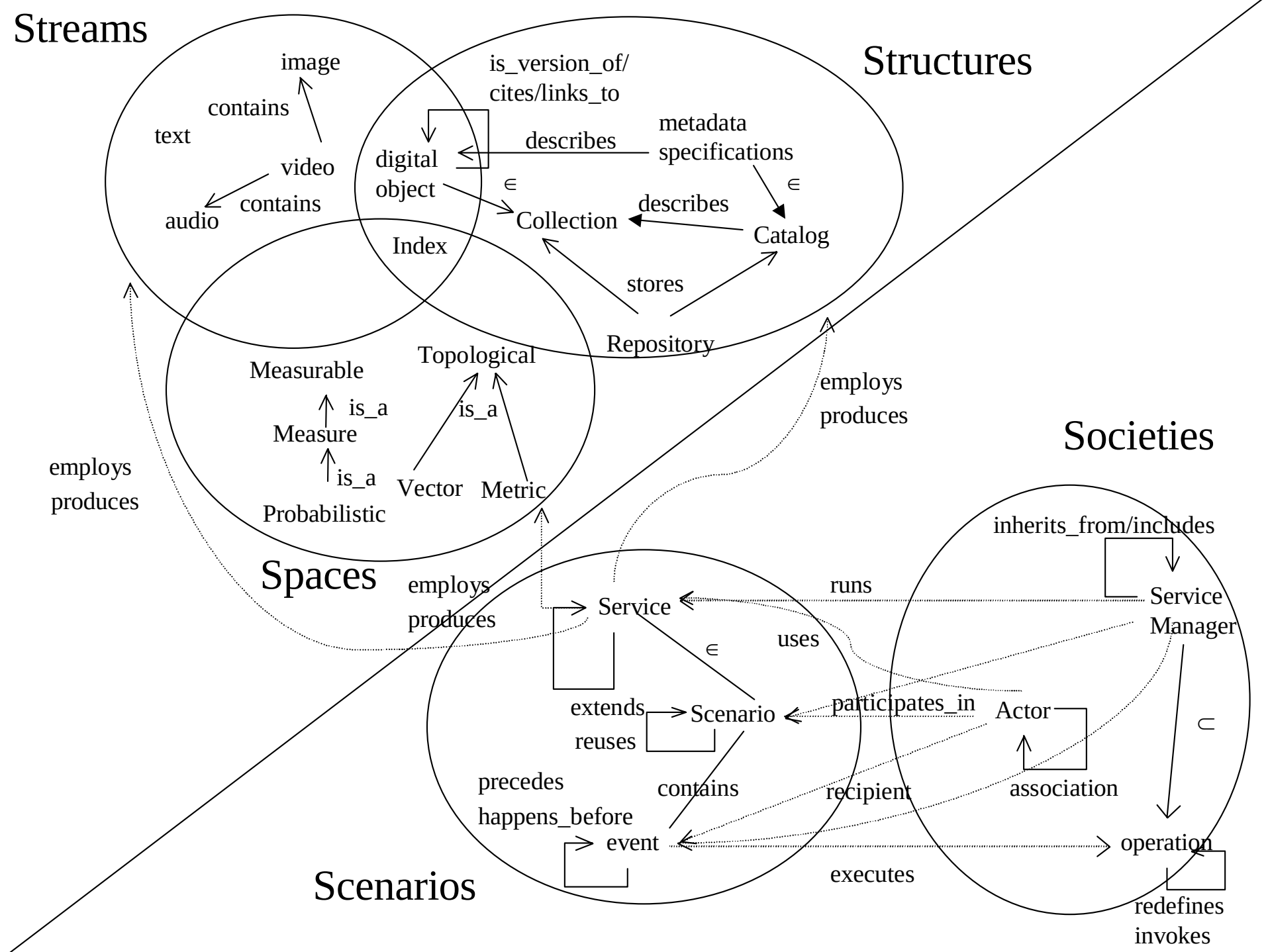
- Part 3 – Advanced Topics
  - Ch. 13: Quality
  - Ch. 14: Integration
  - Ch. 15: How to build a digital library
  - Ch. 16: Research Challenges, Future Perspectives
- Appendix
  - A: Mathematical preliminaries
  - B: Formal Definitions: Ss
  - C: Formal Definitions: DL terms, Minimal DL
  - D: Formal Definitions: Archeological DL
  - E: Glossary of terms, mappings

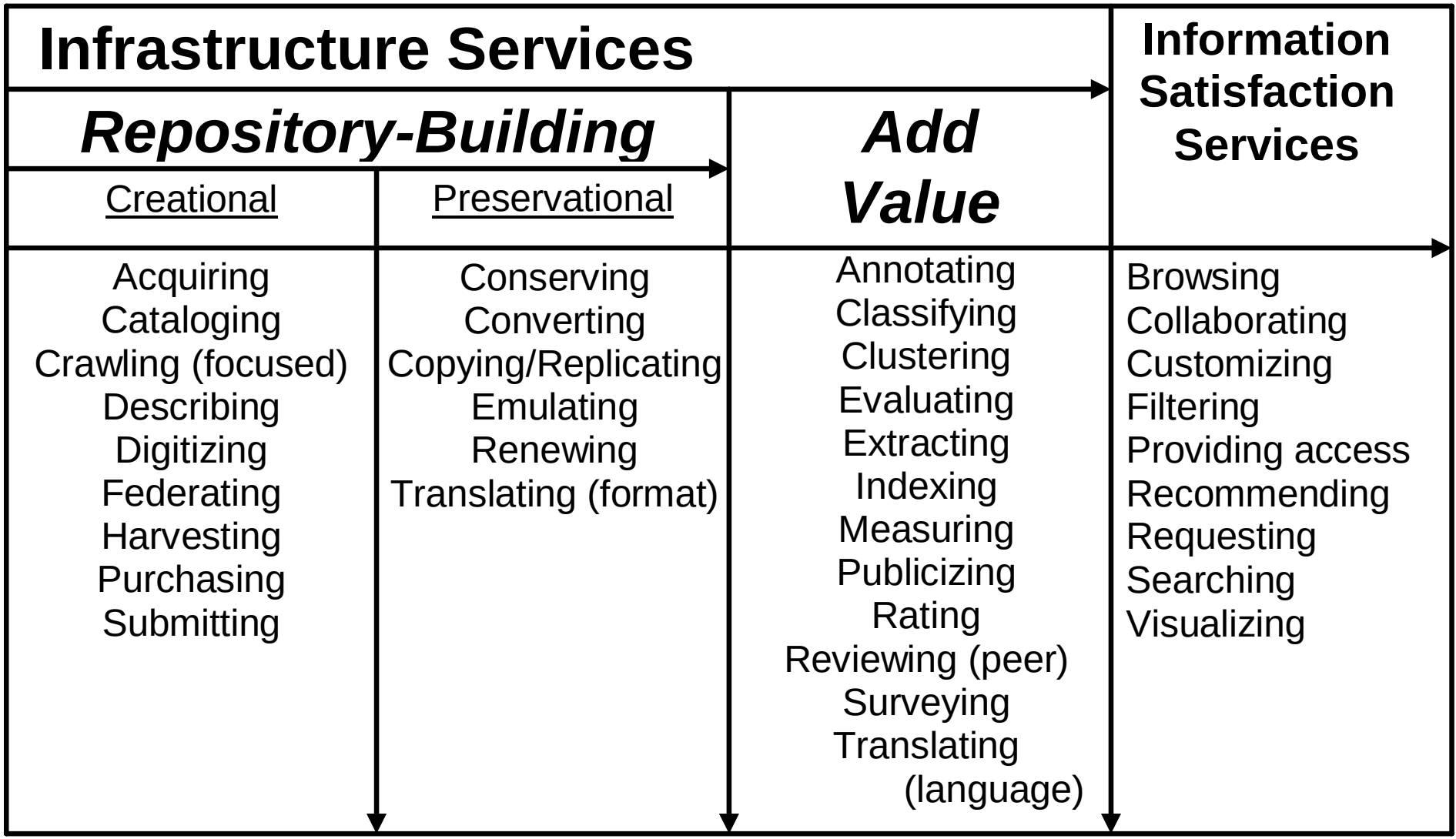
5S and DL formal definitions and compositions (April 2004 TOIS)



A Minimal DL in the 5S Framework




Streams
image
contains
text
video
audio
contains
Structures
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
Probabilistic
Spaces
employs
produces
employs
produces
employs
produces
Societies
inherits_from/includes
runs
Service
Service
Manager
uses
∈
extends
reuses
Scenario
participates_in
Actor
association
⊂
precedes
happens_before
contains
recipient
event
operation
executes
redefines
invokes
Scenarios


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

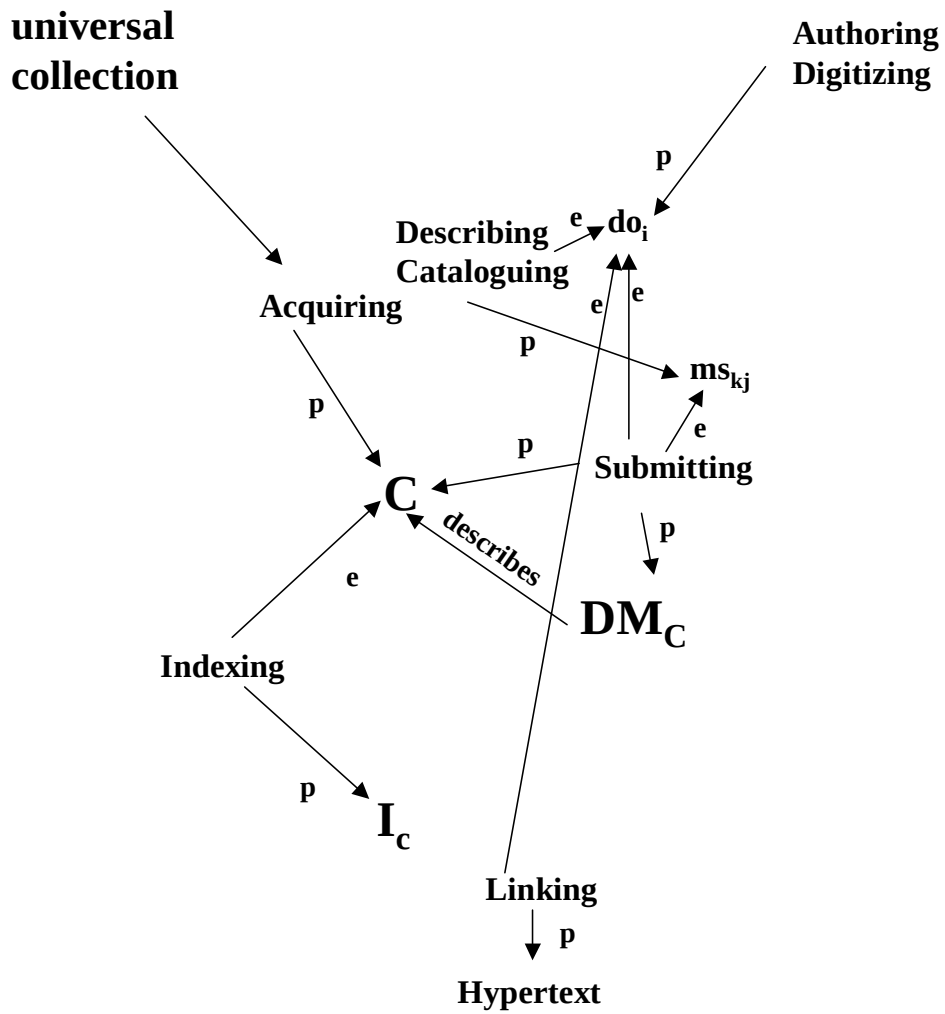
Ontology: Applications

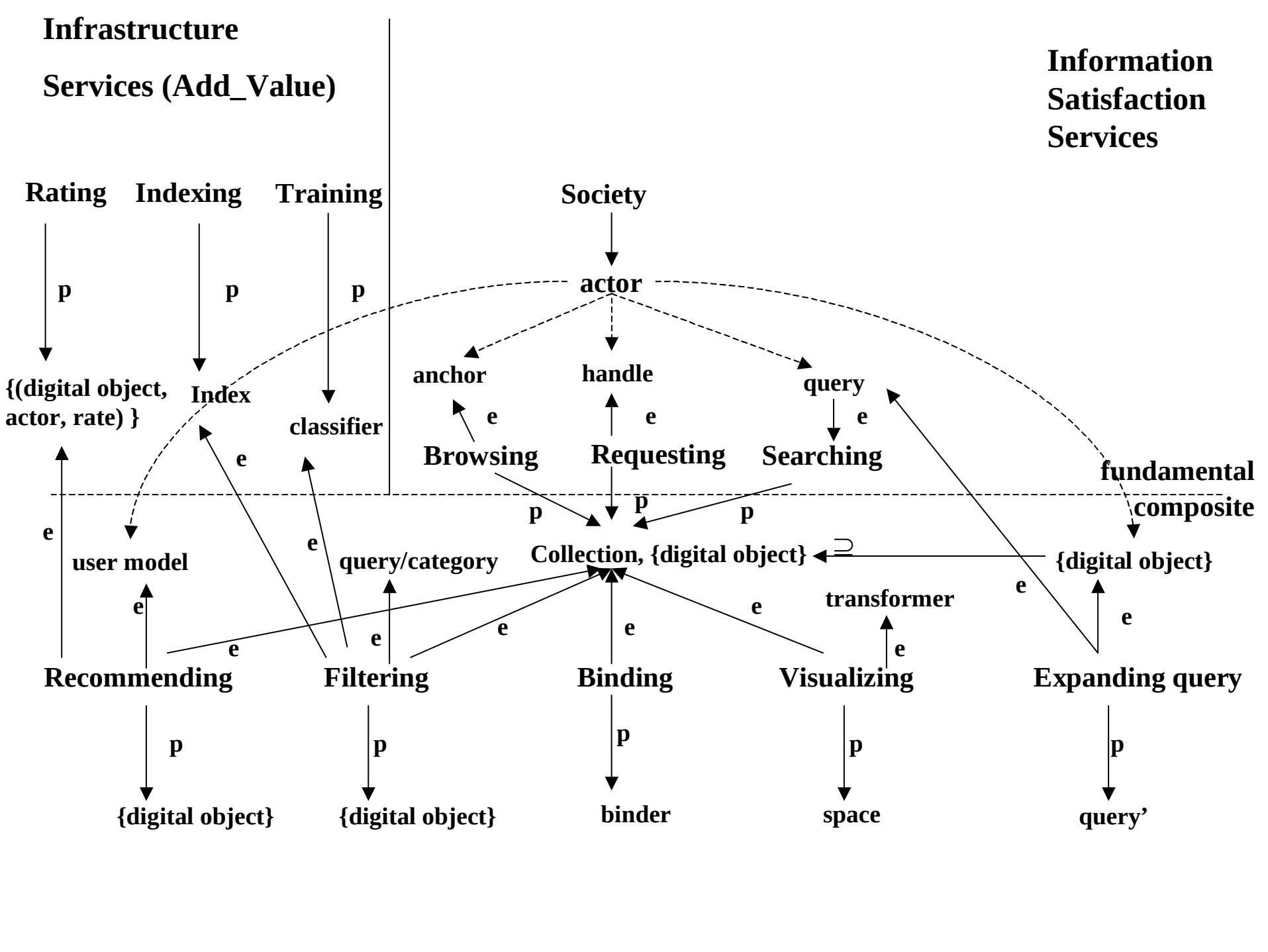
Service	User input	Other Service Input	Output
Acquiring	$\{do_i : i \in I\}$	none	C_j
Annotating	do_i, ann_{ik}	(h_i, ans_{ip})	(h_i, ans_{iq})
Authoring	none[a]	none	do_i
Binding	$\{do_i : i \in I\}, bi_{um}$	$\{do_j : j \in J\}$	bi_{un}
Browsing	anchor	$Hyptxt_j$	$\{do_i : i \in I\}$
Cataloging	h_i, ms_{ik}	(h_i, mss_{ip})	(h_i, mss_{iq})
Classifying	do_i	$class_{Ct}$	$(do_i, \{c_x, \ldots, c_y\})$
Clustering	$\{do_i : i \in I\}$	none	$\{clu_k : k \in K\}$
Conserving	C_i	none	C_k
Converting	do_i	none	do_j
Copying/ Replicating	do_i	none	do_j
Crawling (focused)	C_i	none	C_k
Customizing (interface)	ac_i , trf_k	sp_q	sp_j
Describing	none	do_i	ms_{ik}
Digitizing	none [b]	none	do_i
Disseminating	$\{h_x, \ldots, h_y\}$	none	$\{do_x, ..., do_y\}$
Evaluating	do_i	none	(do_i, w_i)
Expanding (query)	$\{do_i : i \in I\}$	I_C, $q_i, \{do_j : j \in J\}$	q_k

Ontology: Applications

- Expand definition of minimal DL by characterizing
  - typical DL services
  - in the context of “employs” and “produces” relationships
- Use characterization to:
  - Reason about how DL services can be built from other DL components
  - As well as be composed with other services through extension or reuse

Composition of key fundamental / infrastructure services

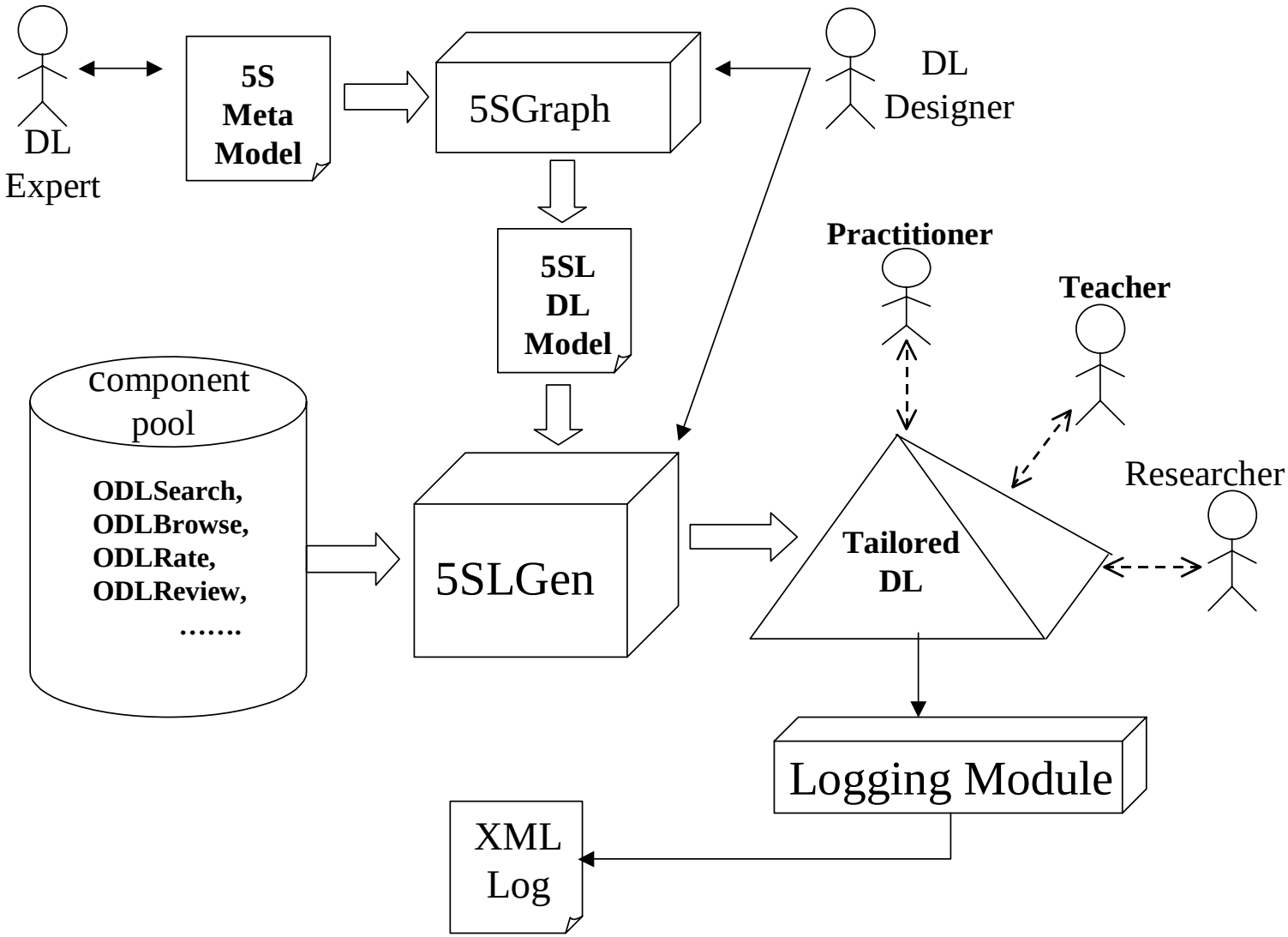



Infrastructure
Services (Add_Value)
Information
Satisfaction
Services
Rating
Indexing
Training
Society
p
p
p
actor
{(digital object,
actor, rate) }
Index
classifier
anchor
handle
query
e
e
e
Browsing
Requesting
Searching
e
fundamental
composite
p
p
p
e
user model
e
query/category
Collection, {digital object}
{digital object}
e
e
transformer
e
e
e
e
e
e
e
Recommending
Filtering
Binding
Visualizing
Expanding query
p
p
p
p
p
{digital object}
{digital object}
binder
space
query'

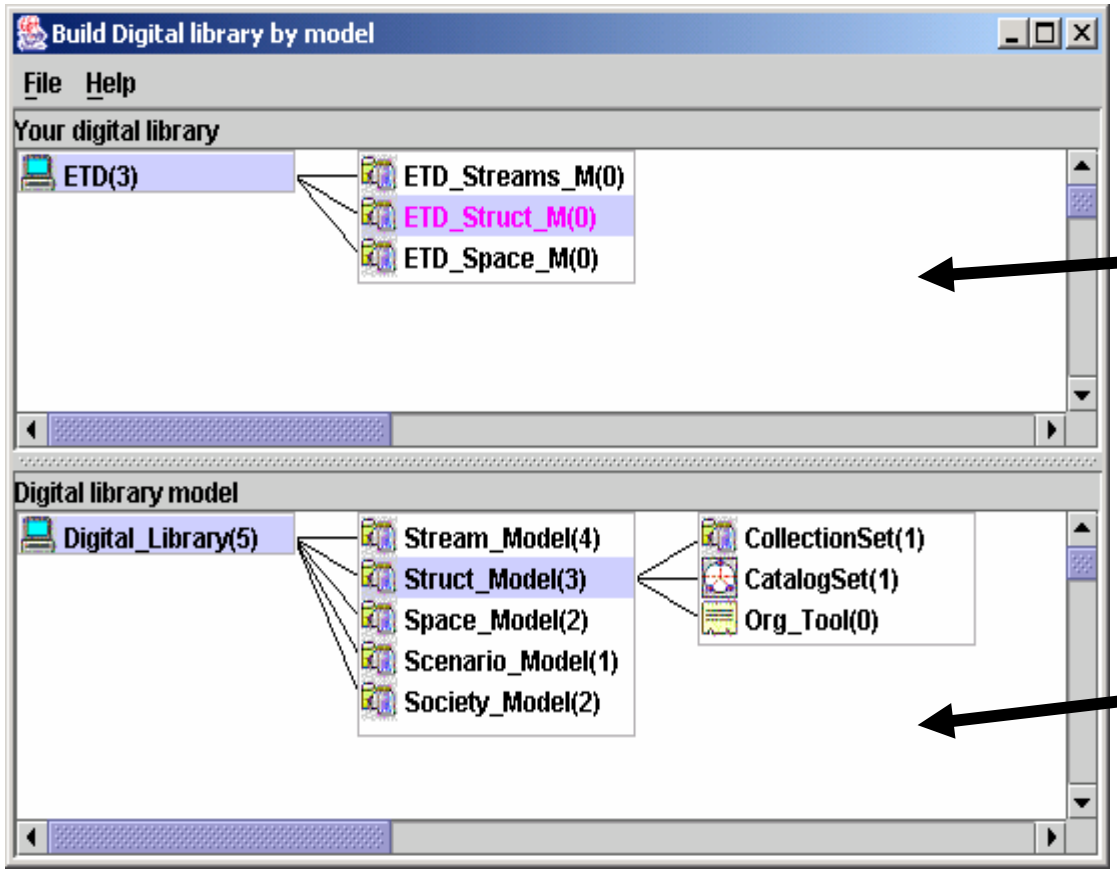
Outline

- Key Publications
- 5S Approach
- **Modeling and Implementation**
- Integration
- Quality
- Challenges for the Reference Model

Tools/Applications



Overview of 5SGraph



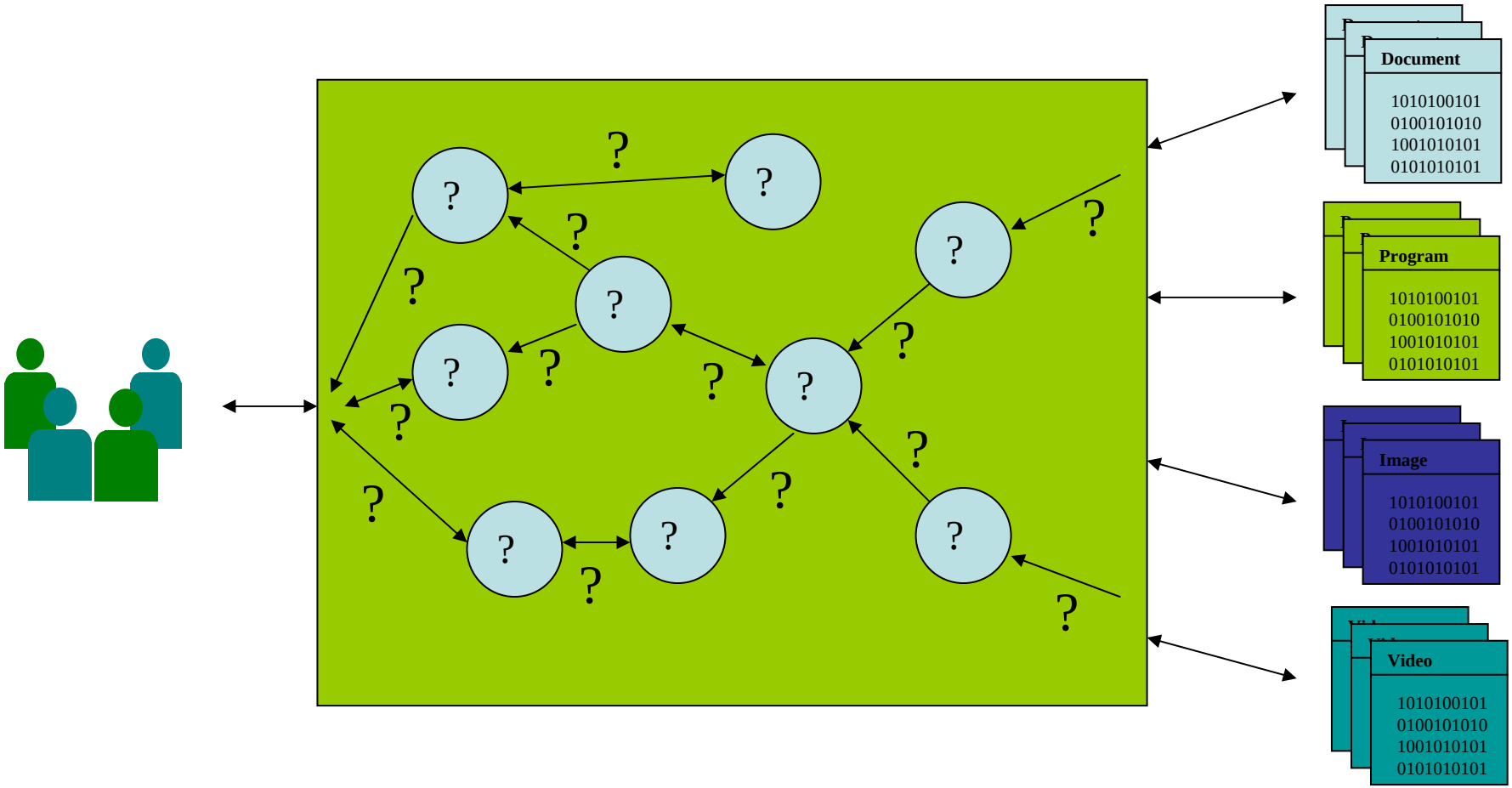
Workspace

(instance model)

Structured

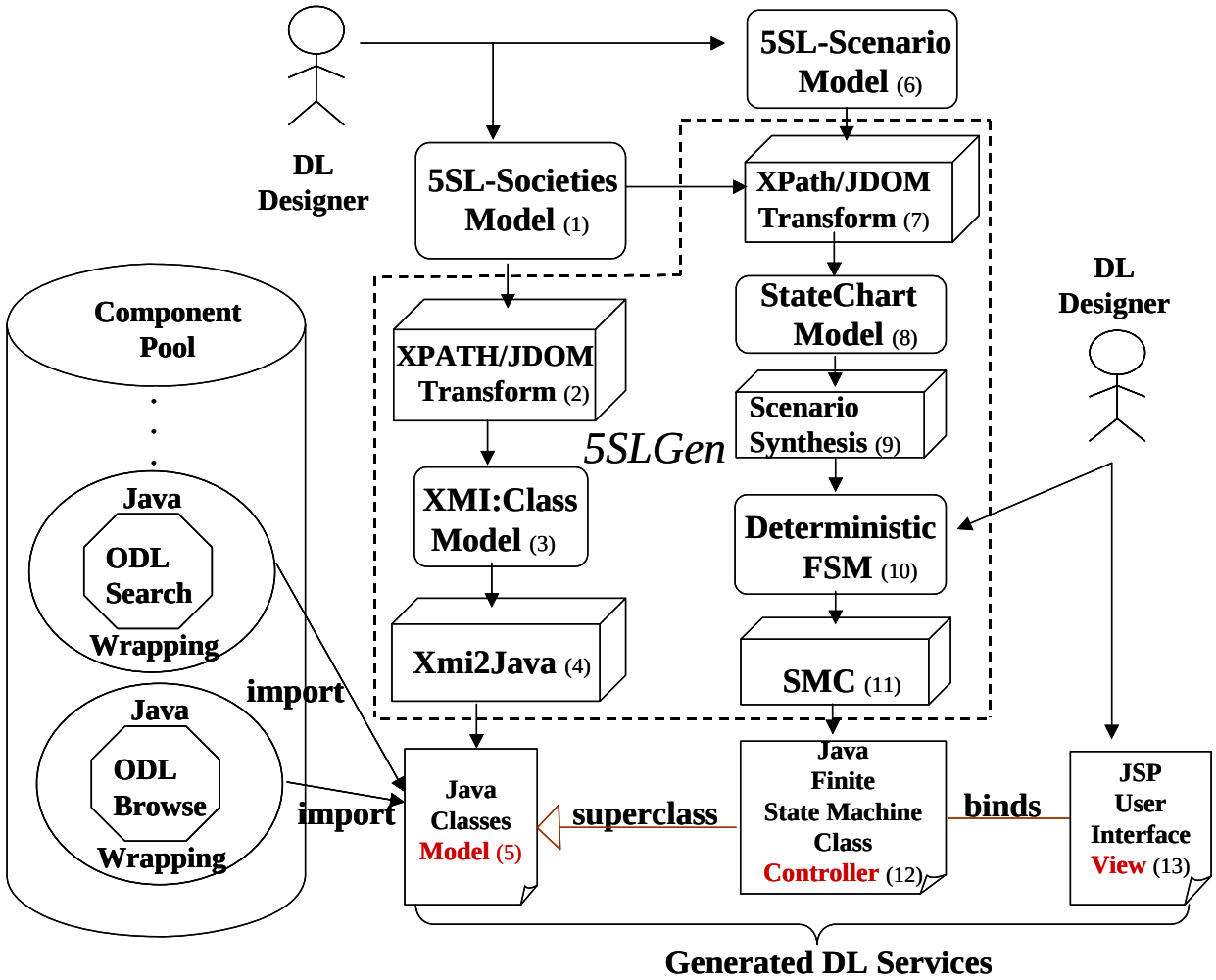
toolbox

(metamodel)



componentized digital library

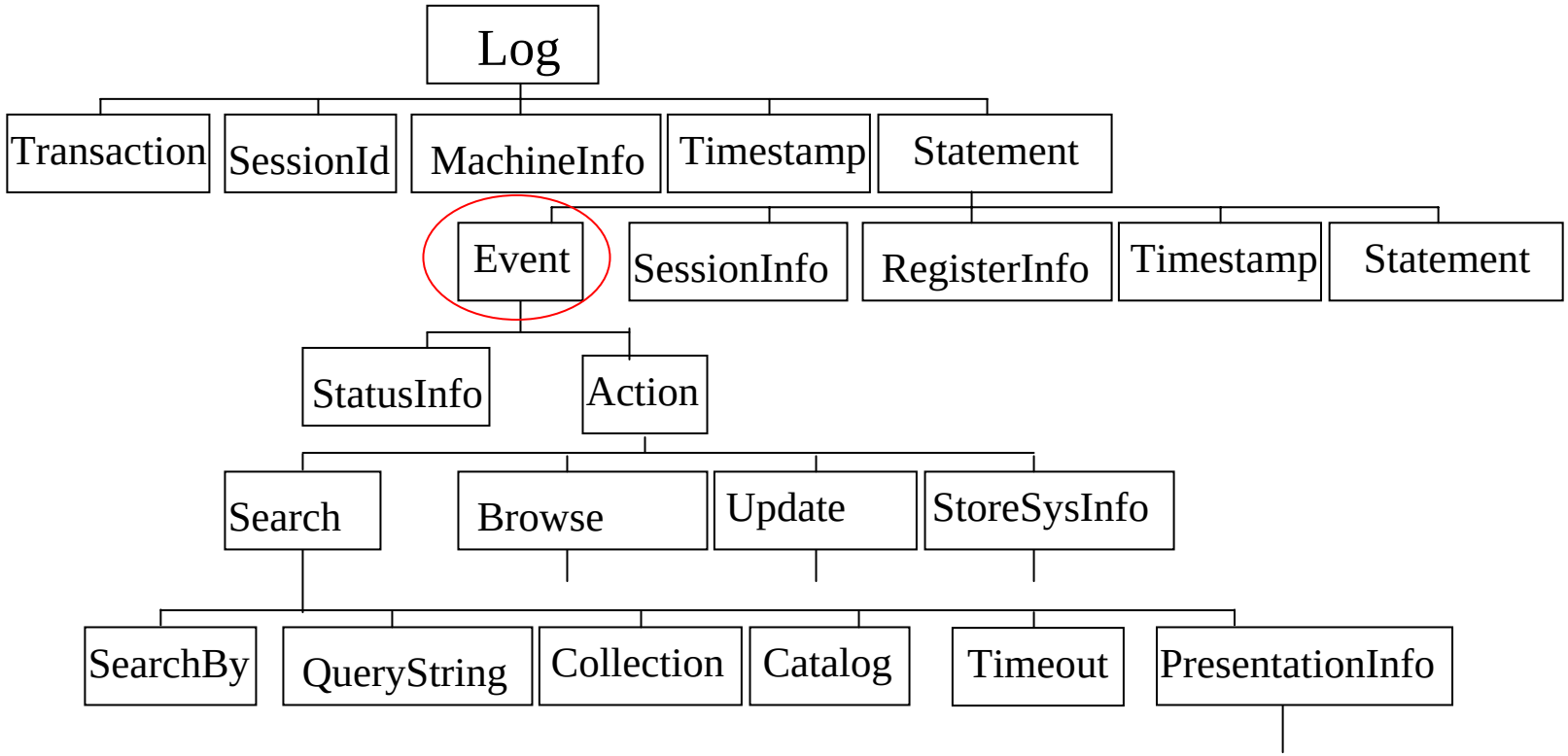
5SGen – Version 2: ODL, Services, Scenarios



XML-based DL Log Standard

- Log analysis
  - is a source of information on:
    - How patrons really use DL services
    - How systems behave while supporting user information seeking activities
- Used to:
  - Evaluate and enhance services
  - Guide allocation of resources
- Common practice in the web setting
  - Supported by web servers, proxy caches
- DL Logging can be more detailed

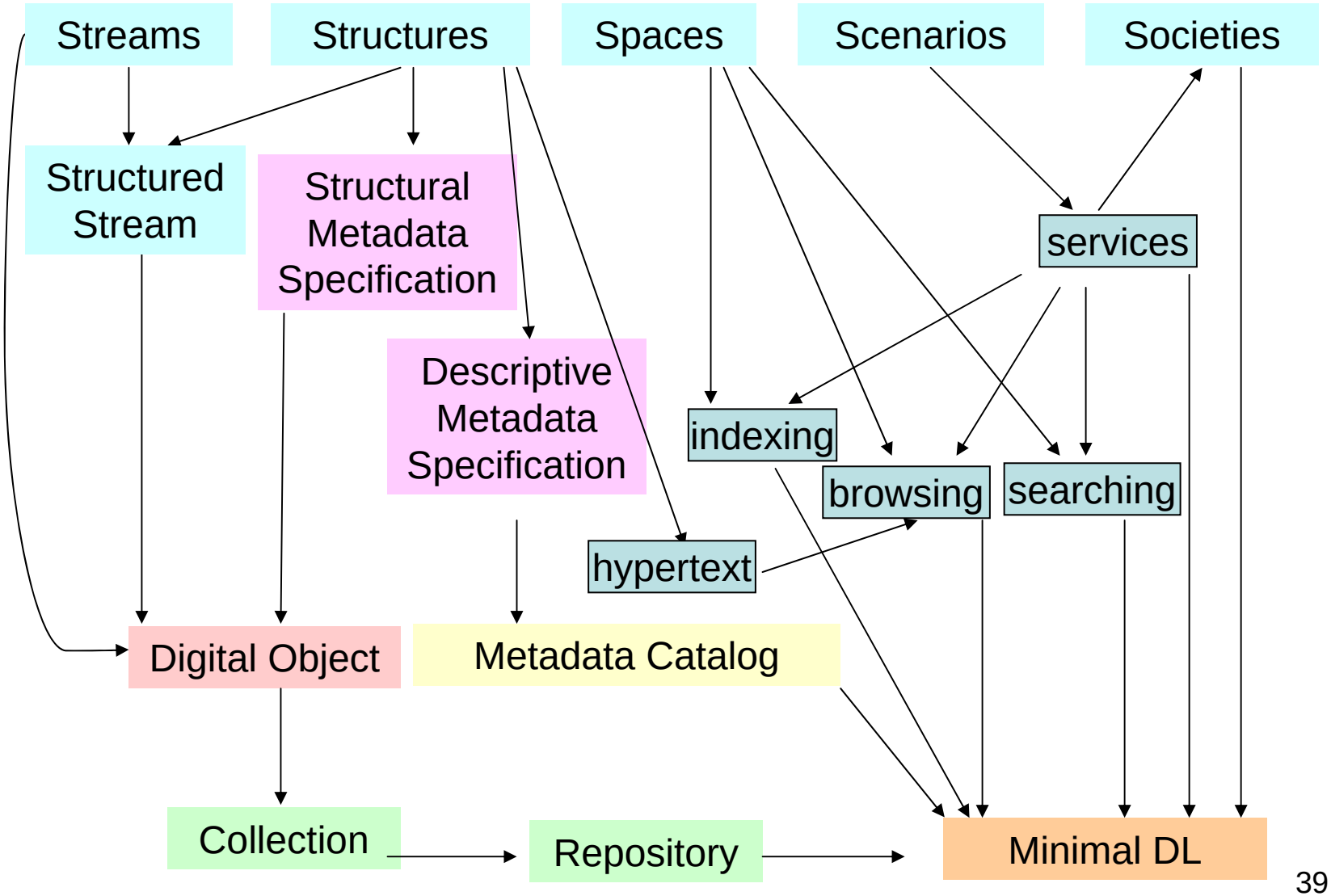
The XML Log Format



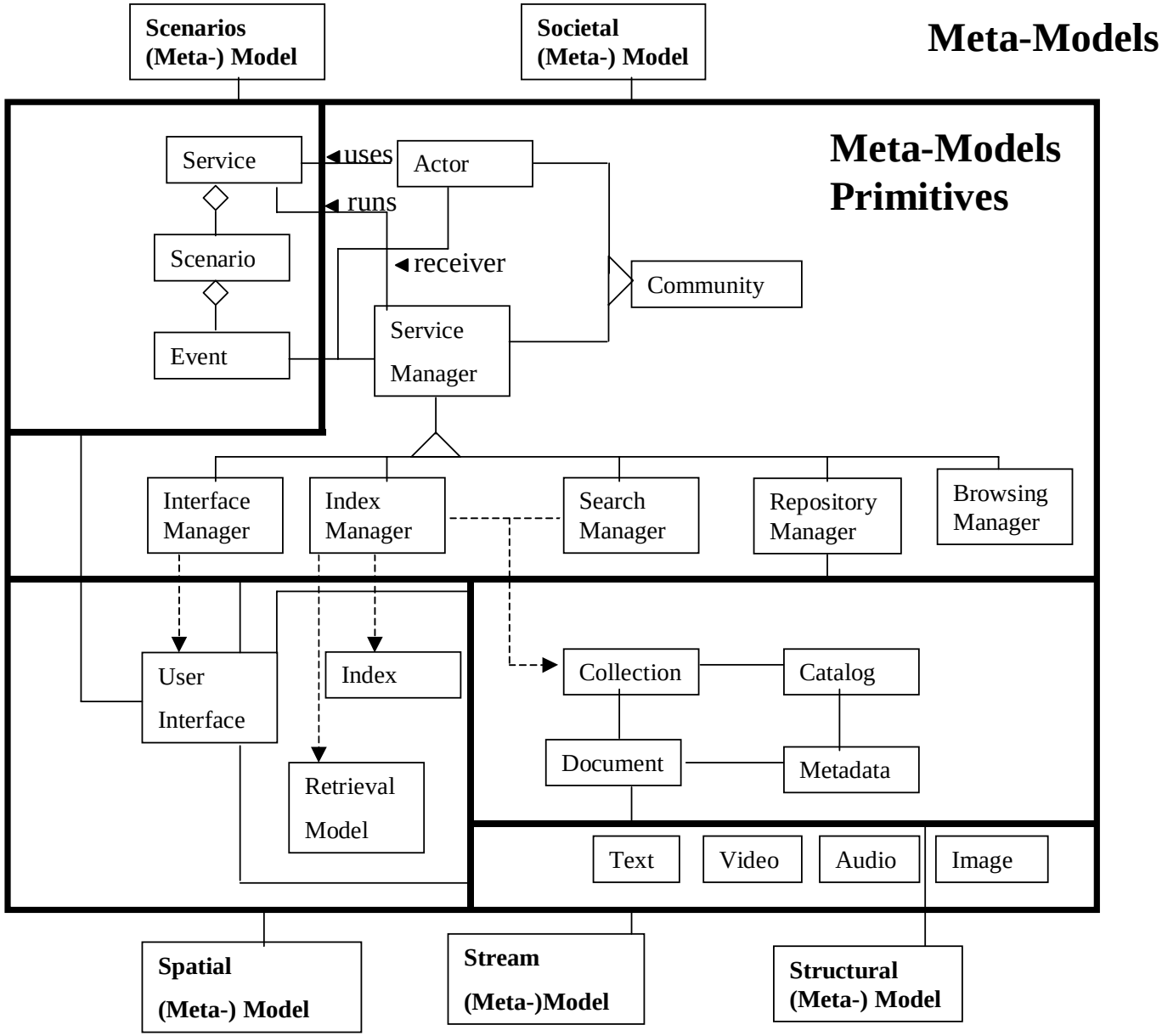
Metamodels

- Completed
  - Minimal
  - Archaeological
- Planned
  - Practical
  - System oriented
    - Doug Gorton’s thesis, so people can build models for their systems, and have them generated to work with a particular DL system

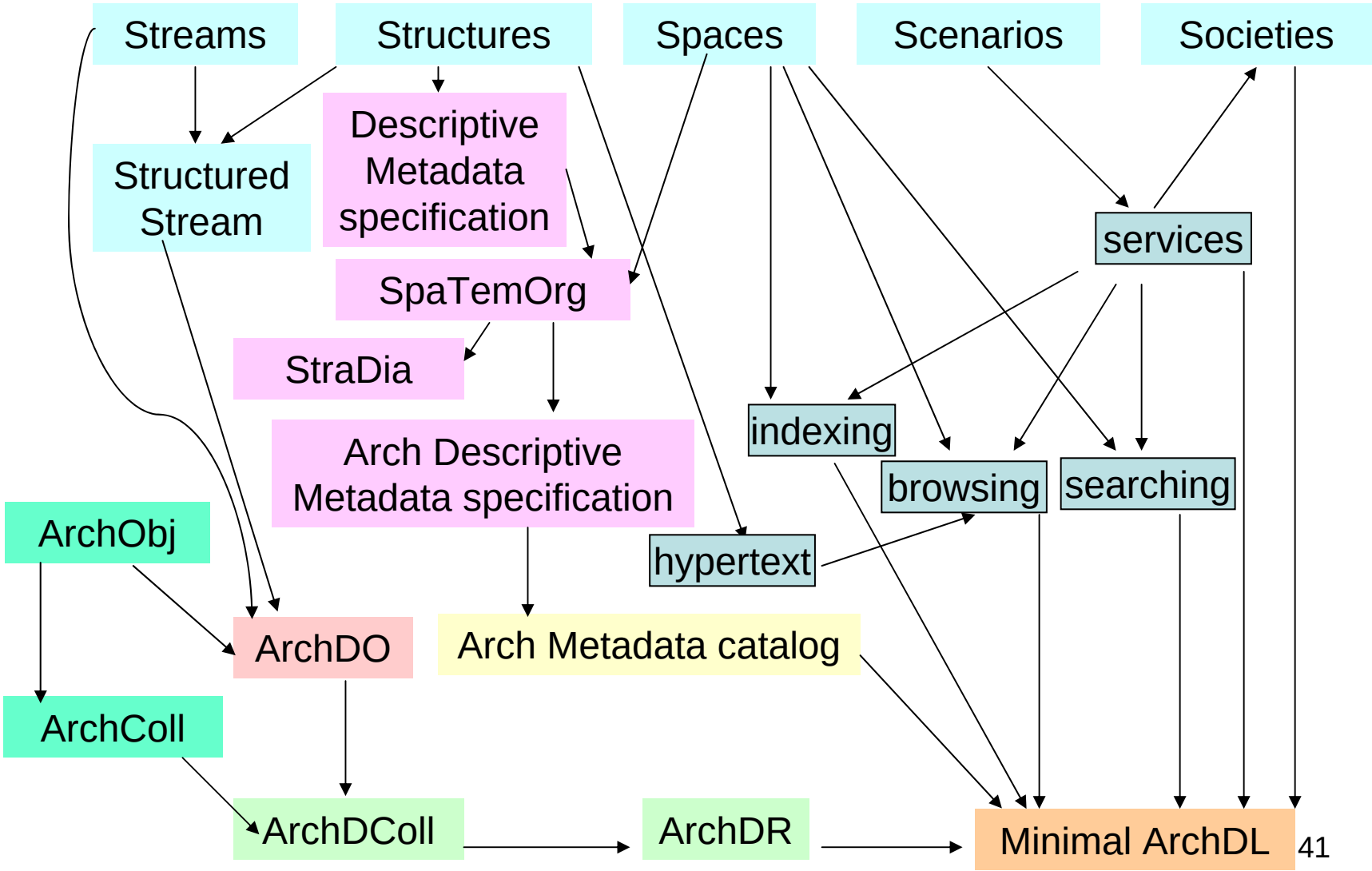
A Minimal DL in the 5S Framework

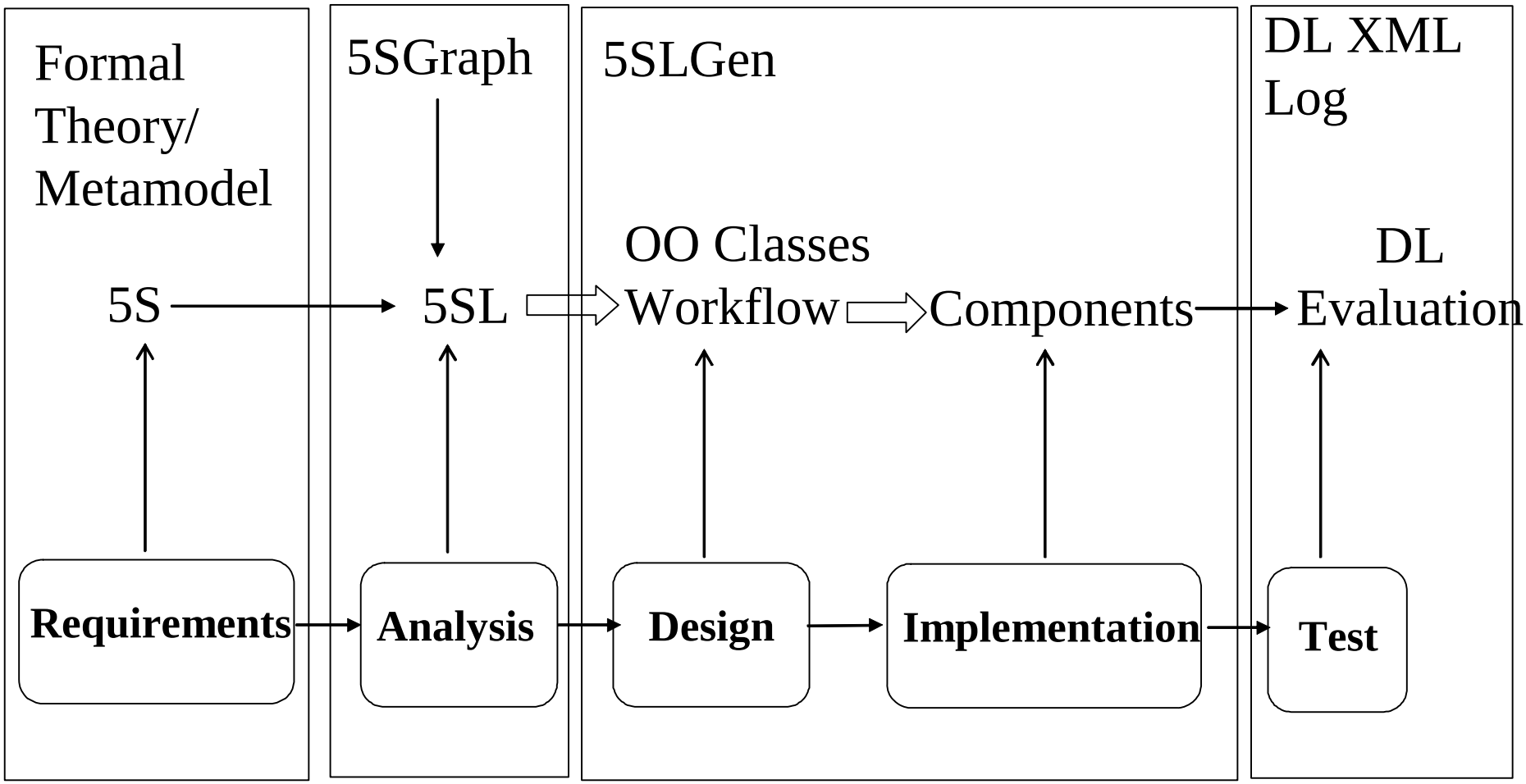


5SL – The Minimal DL Metamodel



A Minimal ArchDL in the 5S Framework




Formal Theory/ Metamodel
5SGraph
5SLGen
DL XML Log
5S
5SL
OO Classes
Workflow
Components
DL
Evaluation
Requirements
Analysis
Design
Implementation
Test

Outline

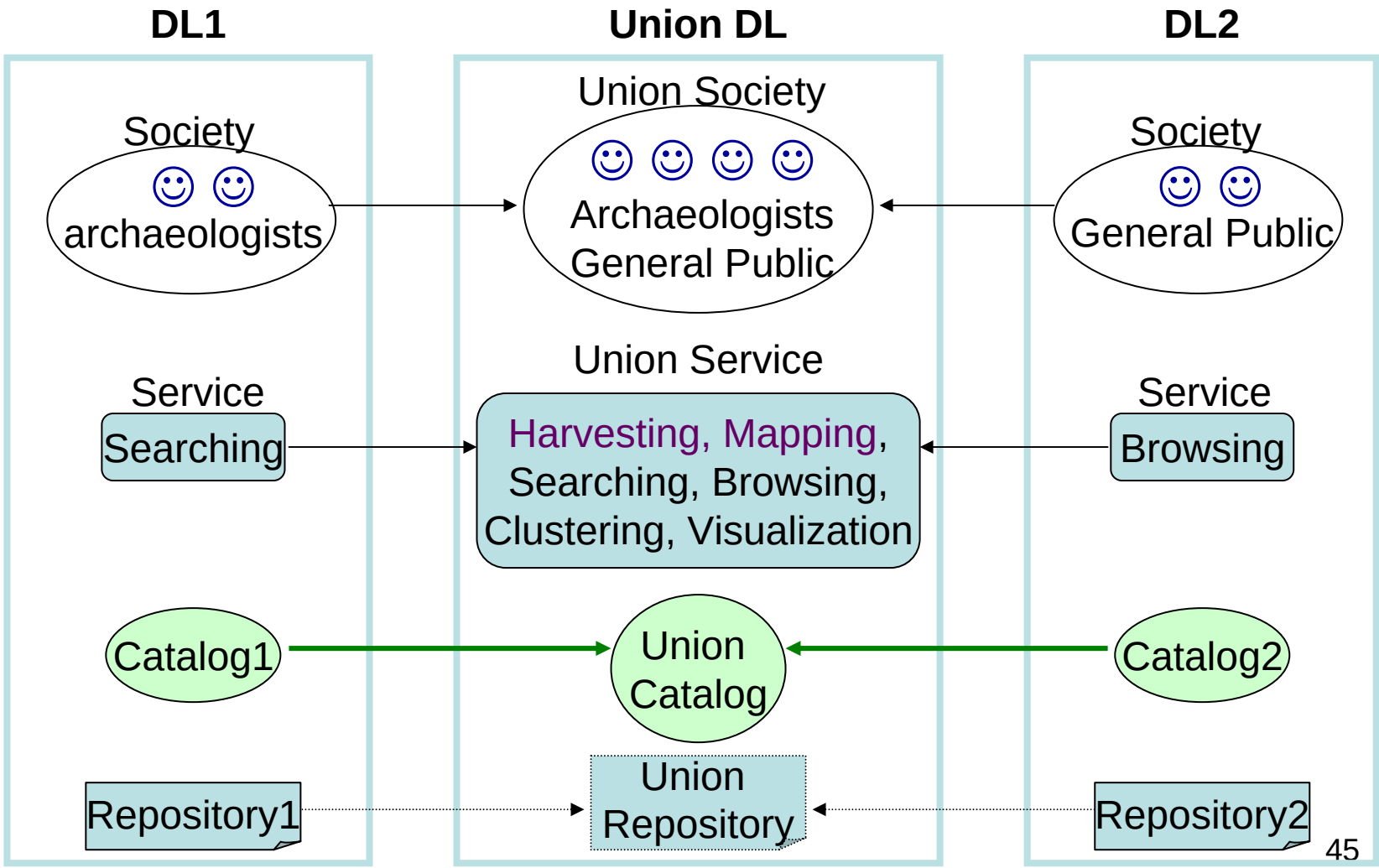
- Key Publications
- 5S Approach
- Modeling and Implementation
- **Integration**
- Quality
- Challenges for the Reference Model

DL Integration

- What is “DL Integration”
  - Hide distribution
  - Hide heterogeneity
  - Enable autonomy of individual component
- Why Integration
  - island-DLs
  - inability to seamlessly and transparently access knowledge across DLs
  - e.g., toward The European DL !

Utilize various autonomous DLs in concert

Global DL: Architecture of a Union DL



Formal Definition of DL Integration

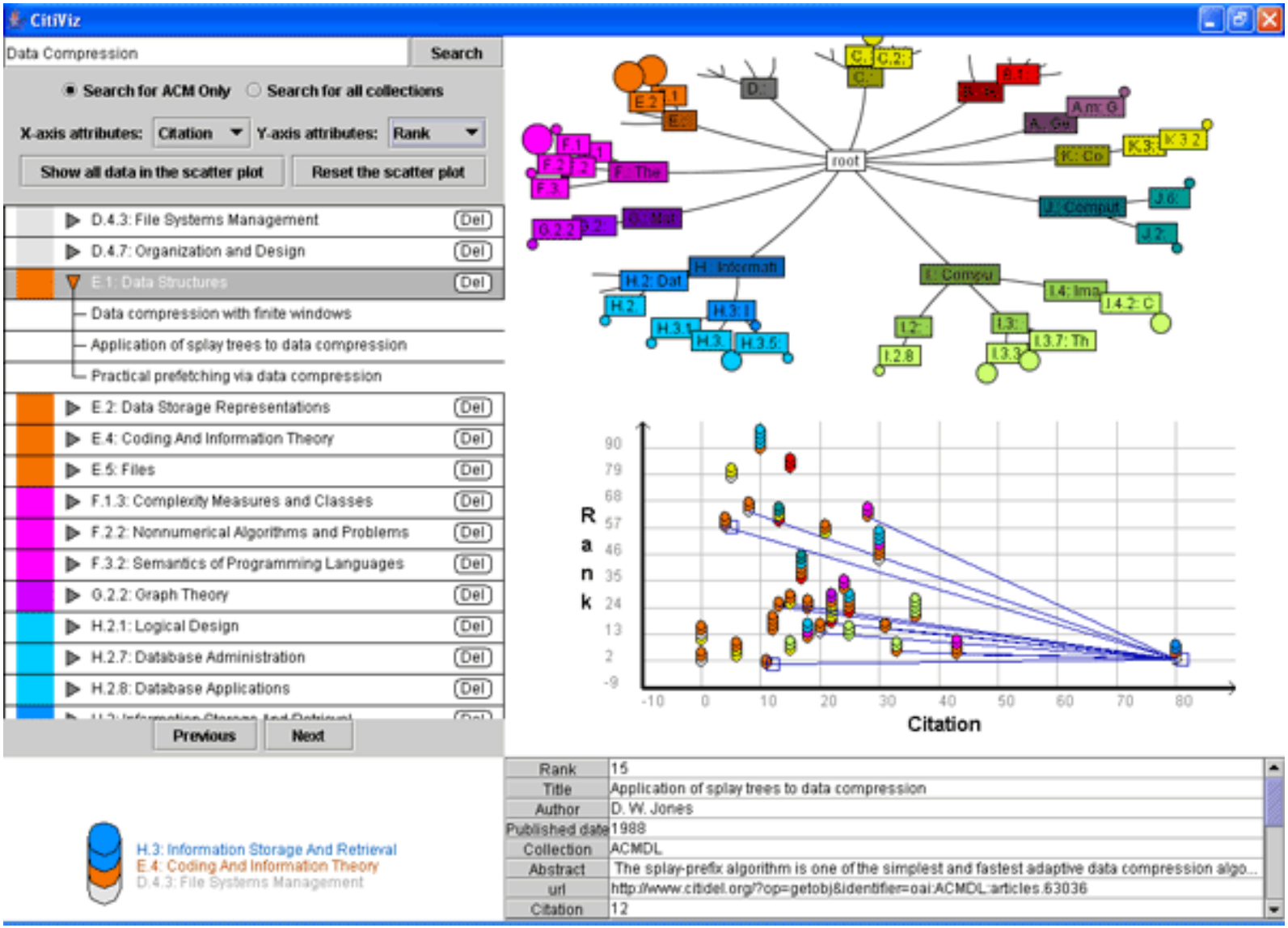
- $DL_i=(R_i, DM_i, Serv_i, Soc_i), 1 \leq i \leq n$
  - R_i *is a network accessible repository*
  - DM_i *is a set of metadata catalogs for all collections*
  - $Serv_i$ *is a set of services*
  - Soc_i *is a society*
- *UnionRep*
- *UnionCat*
- *UnionServices*
- *UnionSociety*

Formal Definition of DL Integration (Cont.)

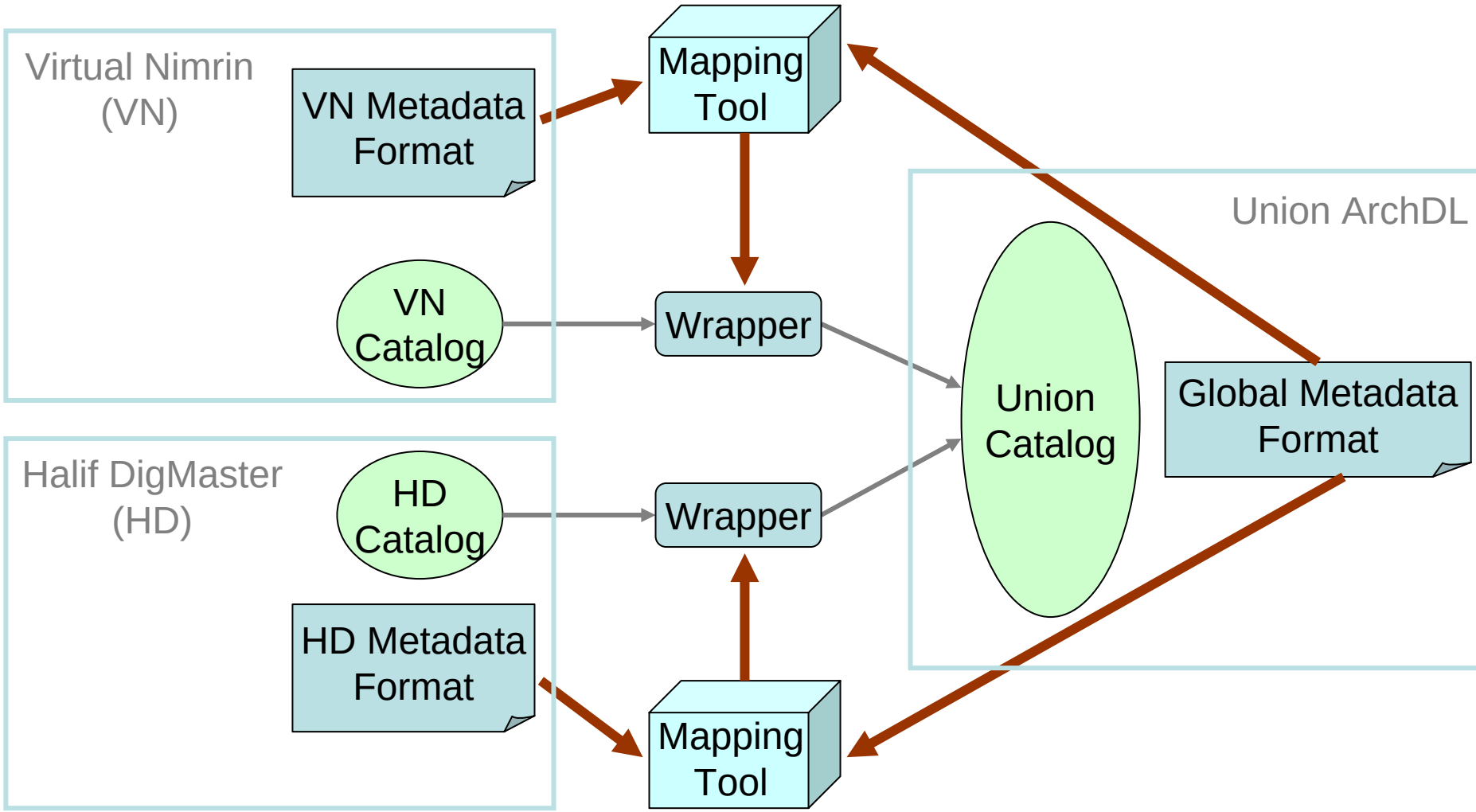
- DL integration problem definition:

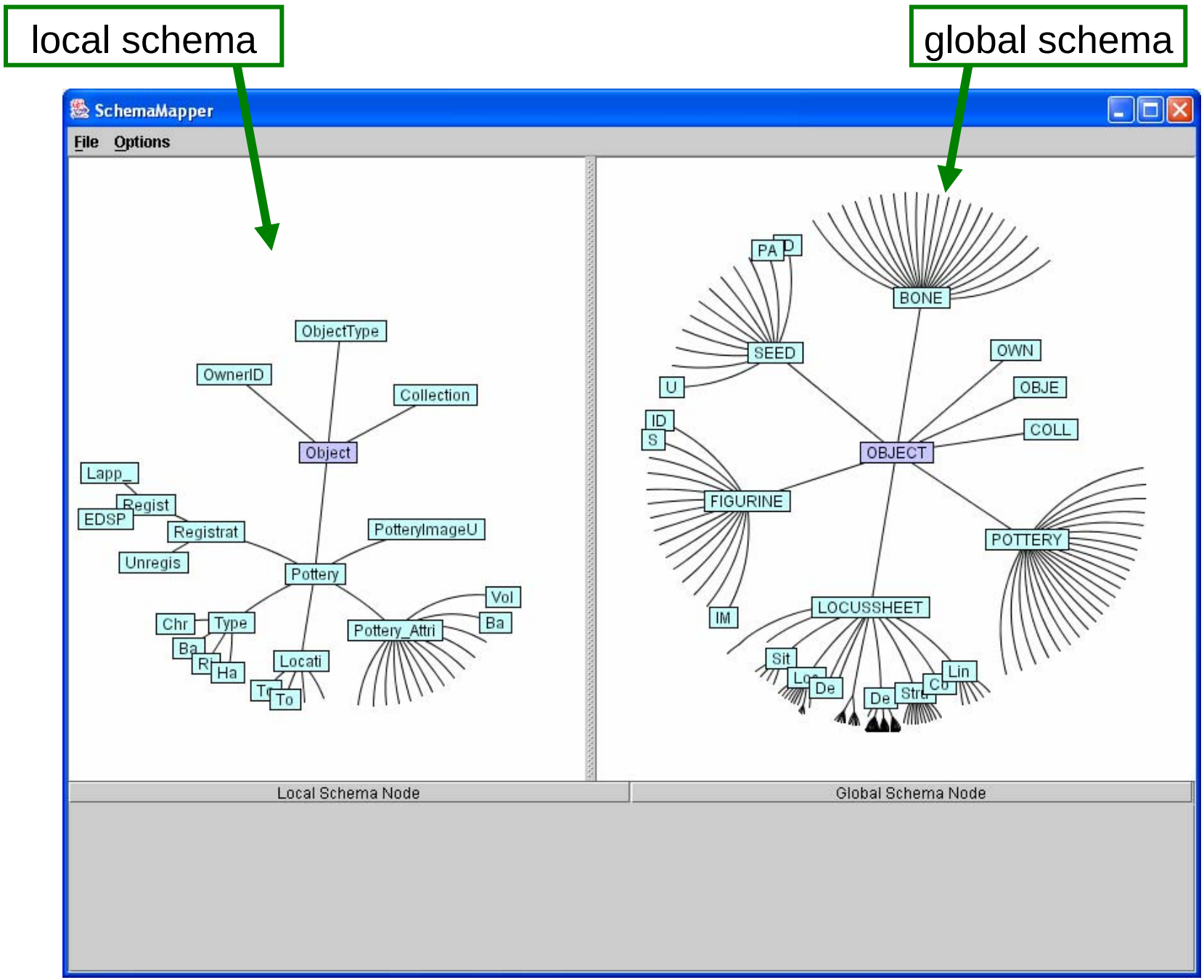
  *Given n individual libraries, integrate the n DLs to create a **UnionDL**.*

Example of Union Service: CitiViz



Union Catalog Integration




local schema
global schema
SchemaMapper
File Options
ObjectType
OwnerID
Collection
Object
Lapp_
Regist
EDSP
Registrat
Unregis
PotteryImageU
Pottery
Chr
Type
Ba
Ri
Ha
Locati
To
Vol
Ba
Pottery_Attri
BONE
SEED
U
ID
S
IM
OWN
OBJE
COLL
OBJECT
FIGURINE
POTTERY
LOCUSSHEET
Sit
De
Lin
Local Schema Node
Global Schema Node

Outline

- Key Publications
- 5S Approach
- Modeling and Implementation
- Integration
- **Quality**
- Challenges for the Reference Model

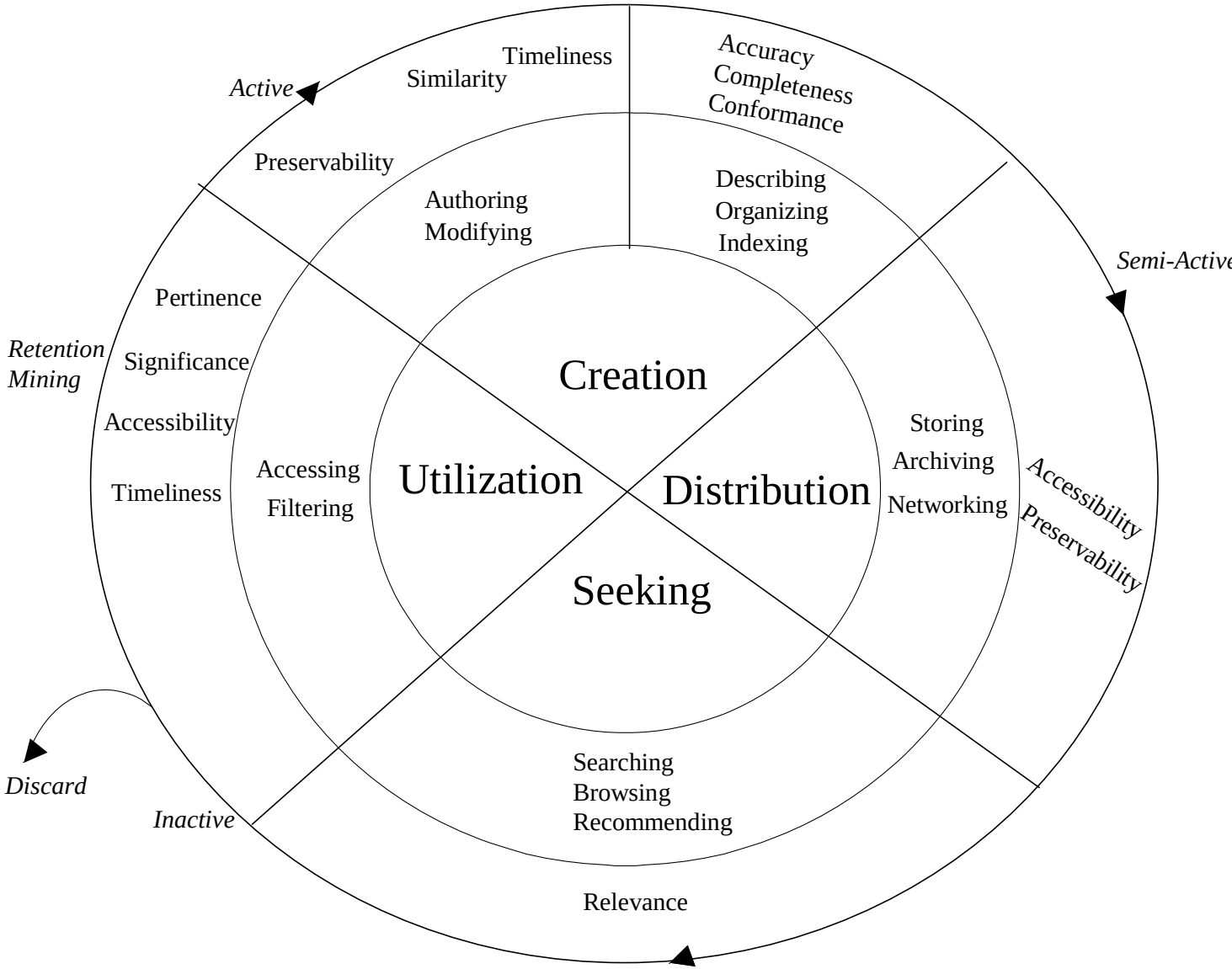
Describing Quality in Digital Libraries

- What's a "good" digital Library?
  - Central Concept: Quality!
  - Hypotheses of this work:
    - Formal theory can help to define "what's a good digital library" by:
    - New formalizations of quality indicators for DLs within our 5S framework
    - Contextualizing these measures within the Information Life Cycle

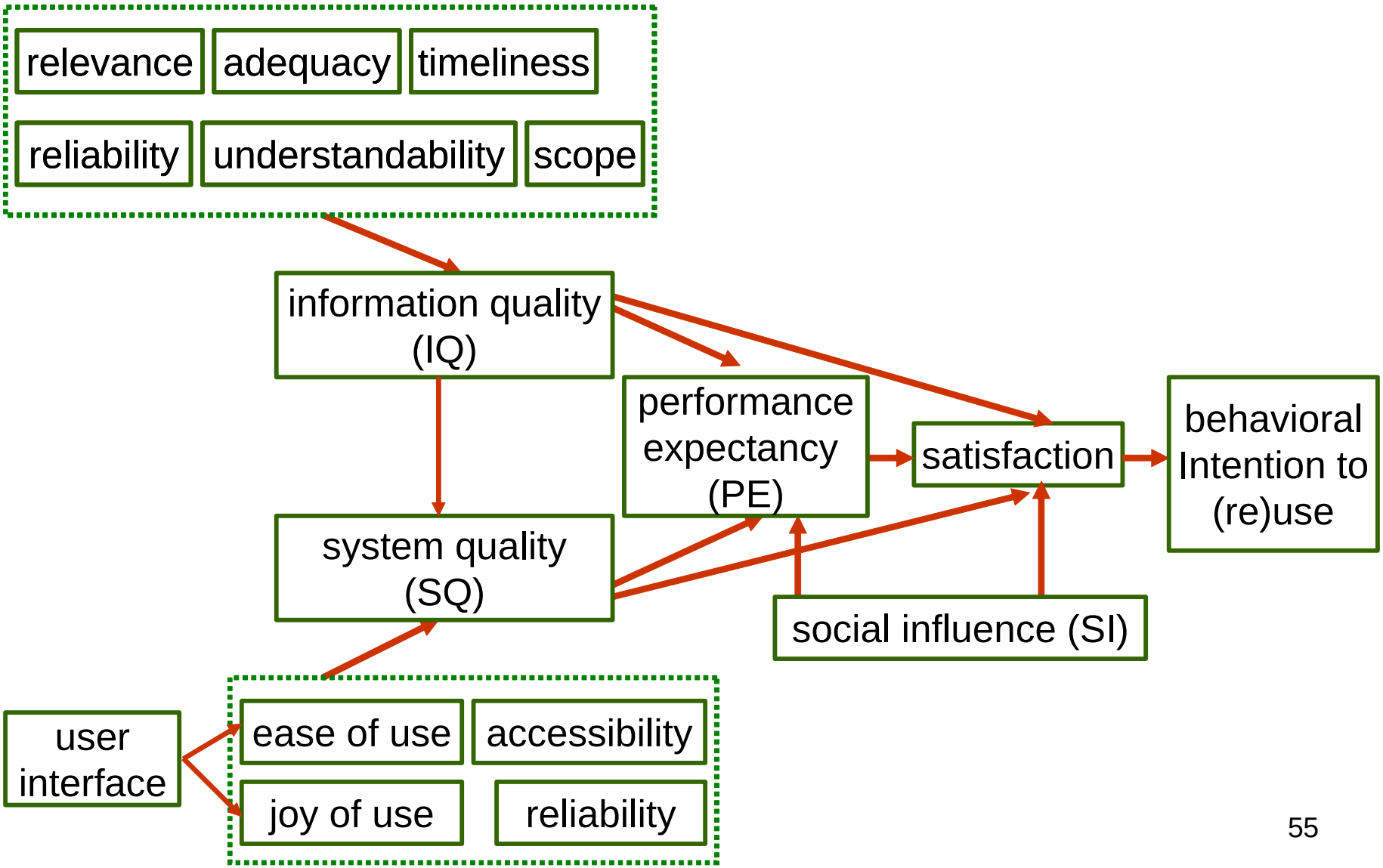
Quality Dimensions

DL Concept	Dimensions of Quality
Digital object	Accessibility Pertinence Preservability Relevance Similarity Significance Timeliness
Metadata specification	Accuracy Completeness Conformance
Collection	Completeness Impact Factor
Catalog	Completeness Consistency
Repository	Completeness Consistency
Services	Composability Efficiency Effectiveness Extensibility Reusability Reliability

Quality and the Information Life Cycle



DL Success Model



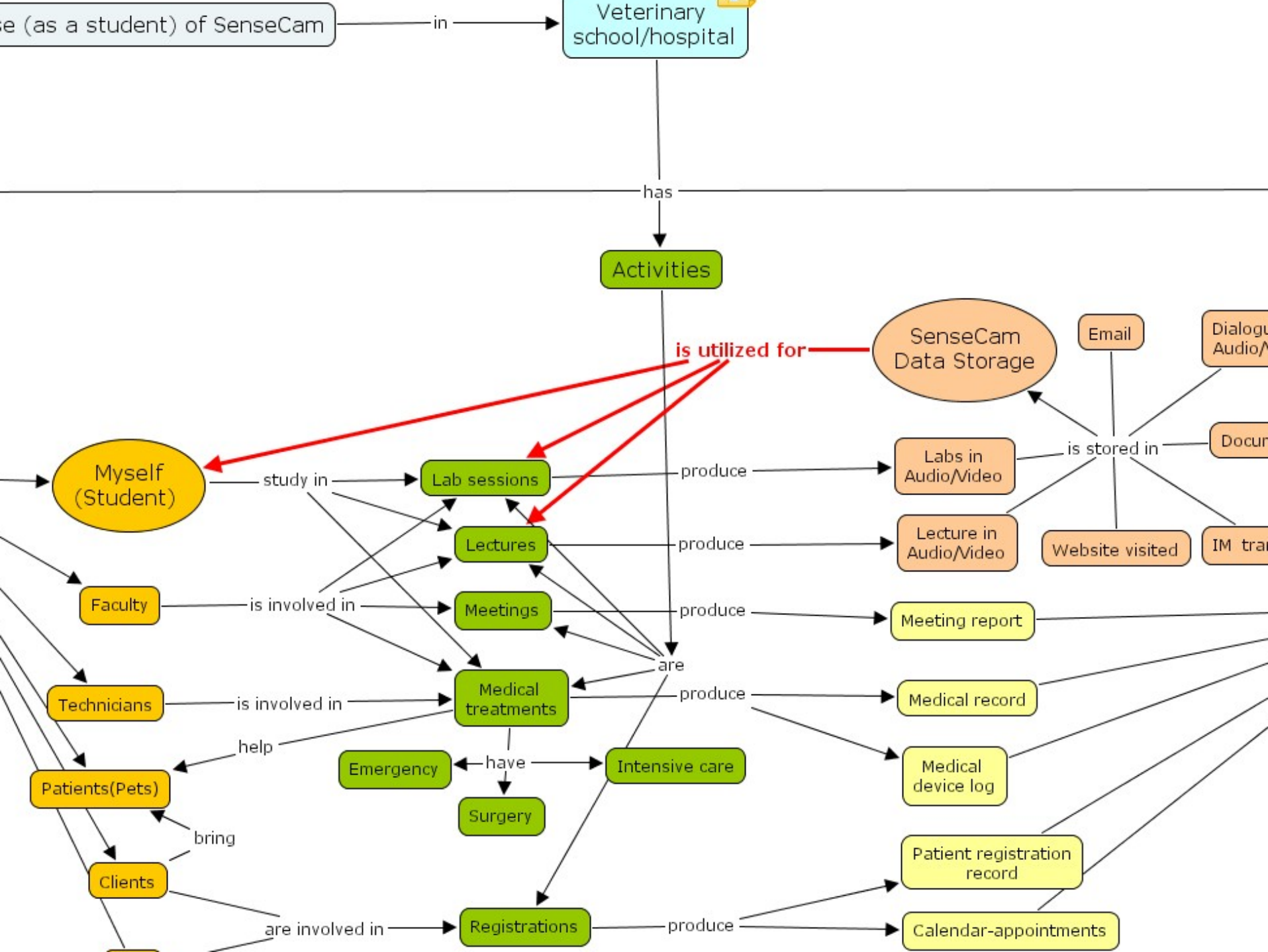
DL quality dimension	DL success manifest variable	5S and DL concept	DL success construct
accessibility *accuracy* *completeness* *consistence* *conformance* *pertinence* *preservability* *relevance* *significance* *similarity* *timeliness*	*adequacy* *relevance* *reliability* *scope* *timeliness* *understandability*	*stream, structure* *digital object* *metadata collection* *catalog* *repository*	*information quality (IQ)*
composability *efficiency* *effectiveness* *extensibility* *reusability* *reliability*	*accessibility* *reliability* *ease of use* *joy of use*	*society, scenario, space* *service*	*system quality (SQ)* *performance expectancy (PE)*
	DL visibility	*society*	*social influence (SI)*

Outline

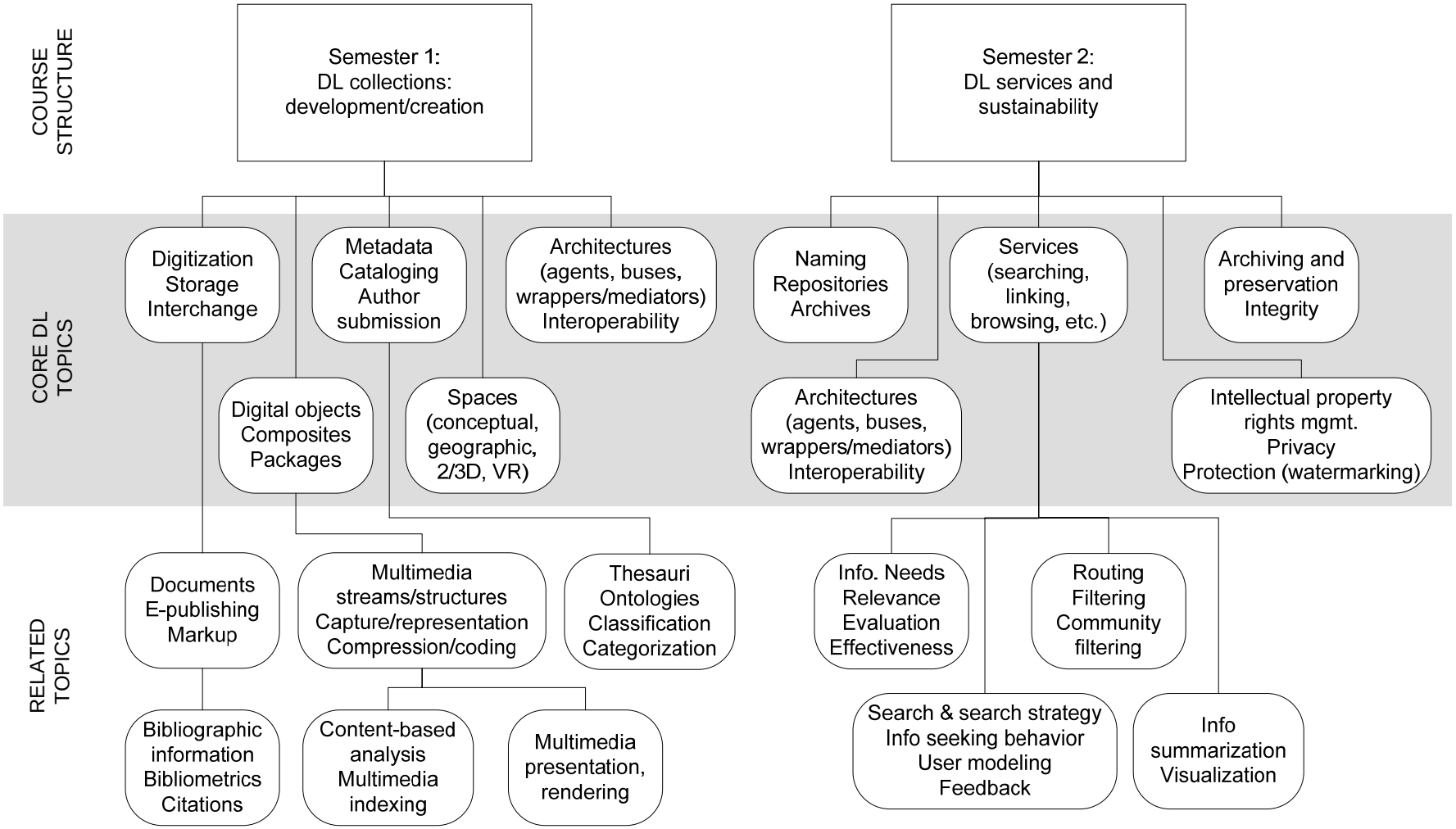
- Key Publications
- 5S Approach
- Modeling and Implementation
- Integration
- Quality
- **Challenges for the Reference Model?**

Practical Systems

- Commercial: IBM, VTLS, …
- Open Source
  - Greenstone
  - CWIS (for NSDL)
  - Institutional repositories
    - DSpace
    - Fedora


se (as a student) of SenseCam
in
Veterinary school/hospital
has
Activities
is utilized for
SenseCam Data Storage
Email
Dialogu Audio/
Docur
Myself (Student)
study in
Lab sessions
produce
Labs in Audio/Video
is stored in
Lectures
produce
Lecture in Audio/Video
Website visited
IM tra
Faculty
is involved in
Meetings
produce
Meeting report
are
Technicians
is involved in
Medical treatments
produce
Medical record
help
Emergency
have
Intensive care
Medical device log
Patients(Pets)
Surgery
bring
Clients
Patient registration record
are involved in
Registrations
produce
Calendar-appointments

DL Curriculum Framework



Selected Links - http://fox.cs.vt.edu

- CITIDEL (computing education resources)
  - www.citidel.org
- NDLTD (electronic theses and dissertations worldwide)
  - www.ndltd.org and etdguide.org
- NSDL (National Science Digital Library)
  - www.nsdl.org
- Virginia Tech Digital Library Research Laboratory (DLRL, www.dlib.vt.edu)
  - 5S, AmericanSouth.Org, CSTC, DL-in-a-box, ENVISION, ETANA, MARIAN, NDLTD, NSDL, OAD, ODL, …)

Mappings of Manifesto – 5S

- Warehouse
- DLS generator
- DL system admin
- Surrogate
- Functionality
- Policy
- Component pool
- 5Sgen
- Uses pool + 5Sgen
- Identifier
- Services
- Constraint

Extend Manifesto Toward 5S

- User -> society (and include agents)
- DL designer -> DL expert specifying metamodel + digital librarian specifying model
- Annotation -> superimposed information mark + new digital object
- Architecture, component, access/ presentation -> scenario, service (with constraints), event, interaction

Add to Manifesto?

- Precise (or even formal) definitions
- DL metamodel plus model
- Hypertext
- Catalog
- Repository (as in OAI, Fedora, DSpace – not just the current where seems only in OAIS)
- Workflow (declarative + management), life cycle
- Ontology -> structure (graph), beyond content
- Streams, Spaces, Context(s)
- How relates to: content management system, institutional repositories, info-viz systems, …: sub/supersets of Manifesto?

Note: IBM shrink-wrap DL had to be renamed CMS.

Offers

- Can share large bibliography re DL
- NSF-funded DL curriculum effort can help disseminate this widely
- IEEE-TCDL could help broaden the discussion
- Doug Gorton thesis might cover another pilot implementation
- OAI might be involved (beyond PMH)

Questions?
Discussion?

Thank You!