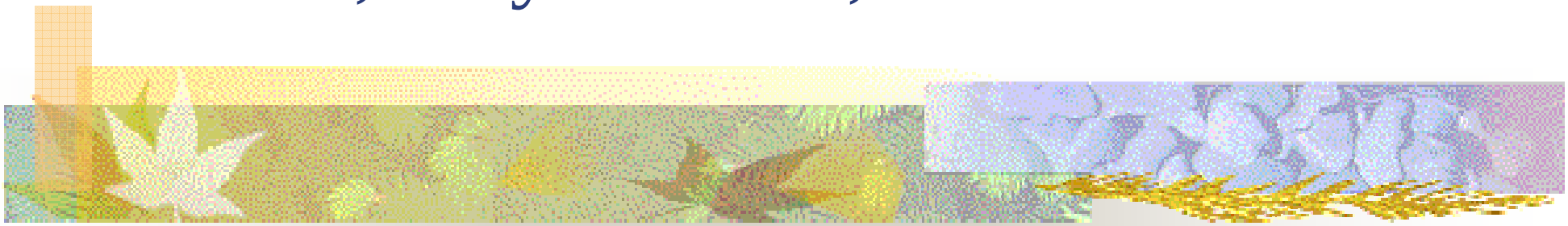
Towards a Quality Model for Digital Libraries

DELOS Evaluation Workshop
Padova, Italy Oct. 4, 2004



Edward A. Fox, Marcos André Gonçalves, Baoping Zhang, Layne T. Watson

Virginia Tech, Blacksburg, VA 24061 USA
fox@vt.edu http://fox.cs.vt.edu/talks

Acknowledgements (Selected)

- **Sponsors:** ACM, Adobe, AOL, CAPES, CONACyT, DFG, IBM, NLM, NSF (IIS-9986089, 0086227, 0080748, 0325579; DUE-0121679, 0136690, 0121741, 0333601), OCLC, VTLS
- **VT Faculty/Staff:** Debra Dudley, Weiguo Fan, Gail McMillan, Manuel Perez, Naren Ramakrishnan, …
- **VT Students:** Yuxin Chen, Shahrooz Feizabadi, Nithiwat Kampanya, S.H. Kim, Aaron Krowne, Bing Liu, Ming Luo, Paul Mather, Fernando Das Neves, Unni. Ravindranathan, Ryan Richardson, Rao Shen, Ohm Sornil, Hussein Suleman, Ricardo Torres, Wensi Xi, …

Special thanks

- Norbert Fuhr, Maristella Agosti, other organizers and support team, DELOS, EU
- Last time in Padova and Venice was 1972
- Have been frustrated by lack of interest in DL evaluation for several years
- How can we move DL into a science without agreement on evaluation?
- Would be happy to help connect this with IEEE TCDL, NSDL, and other efforts

Outline

- ***Major Points of the Presentation***
- Introduction to the 5S View of DLs
  - Informal Definition of DLs
  - Formal Definition of DLs
  - DL Ontology
- Defining a Quality Model for DLs
- Quality and the Information Life Cycle
- An XML Log Standard for DLs
- Conclusions and Future Work

Major Points of the Presentation

- New formalization for digital libraries: 5S
- Formalization of traditional (and new?) measures within our 5S framework
- Contextualization of these measures within the information life cycle, and some data
- Reminder of work on DL logging, in hopes this can be refined and adapted/adopted
- Encourage wider sharing of tools, collections, beyond original intent (e.g., INEX, ETDs)

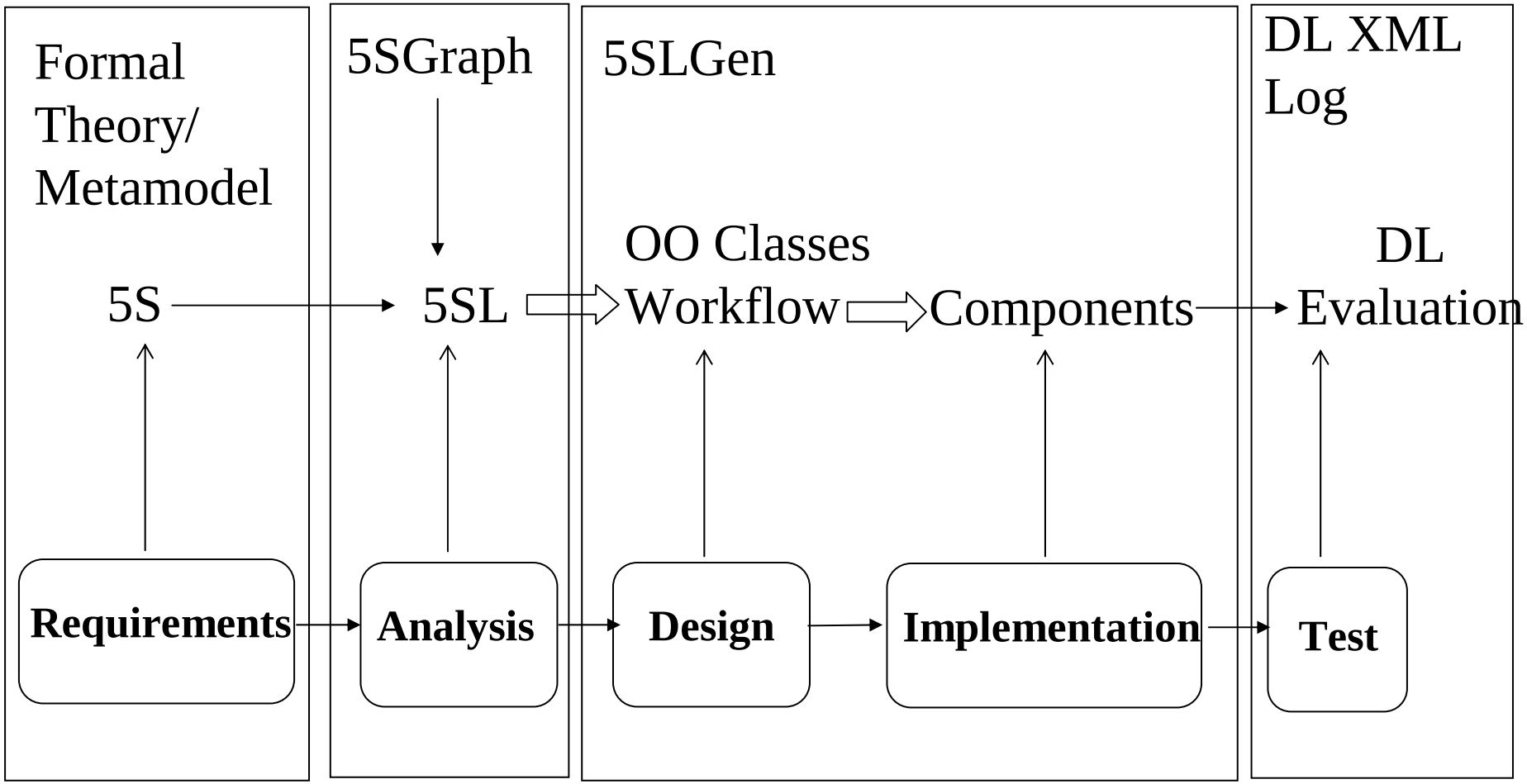
Outline

- Major Points of the Presentation
- ***Introduction to the 5S View of DLs***
  - Informal Definition of DLs
  - Formal Definition of DLs
  - DL Ontology
- Defining a Quality Model for DLs
- Quality and the Information Life Cycle
- An XML Log Standard for DLs
- Conclusions and Future Work

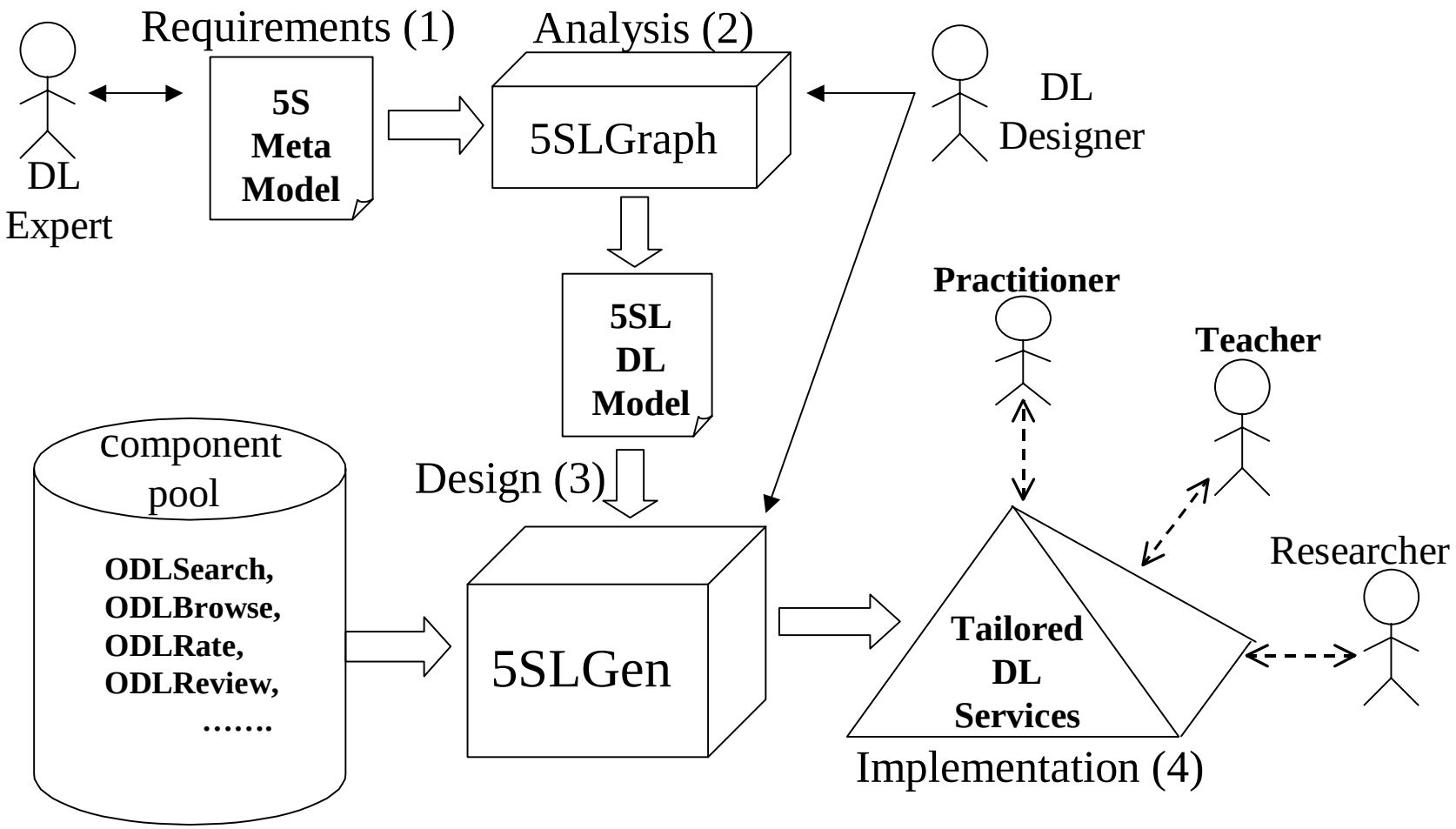
DL Services/Activities Taxonomy

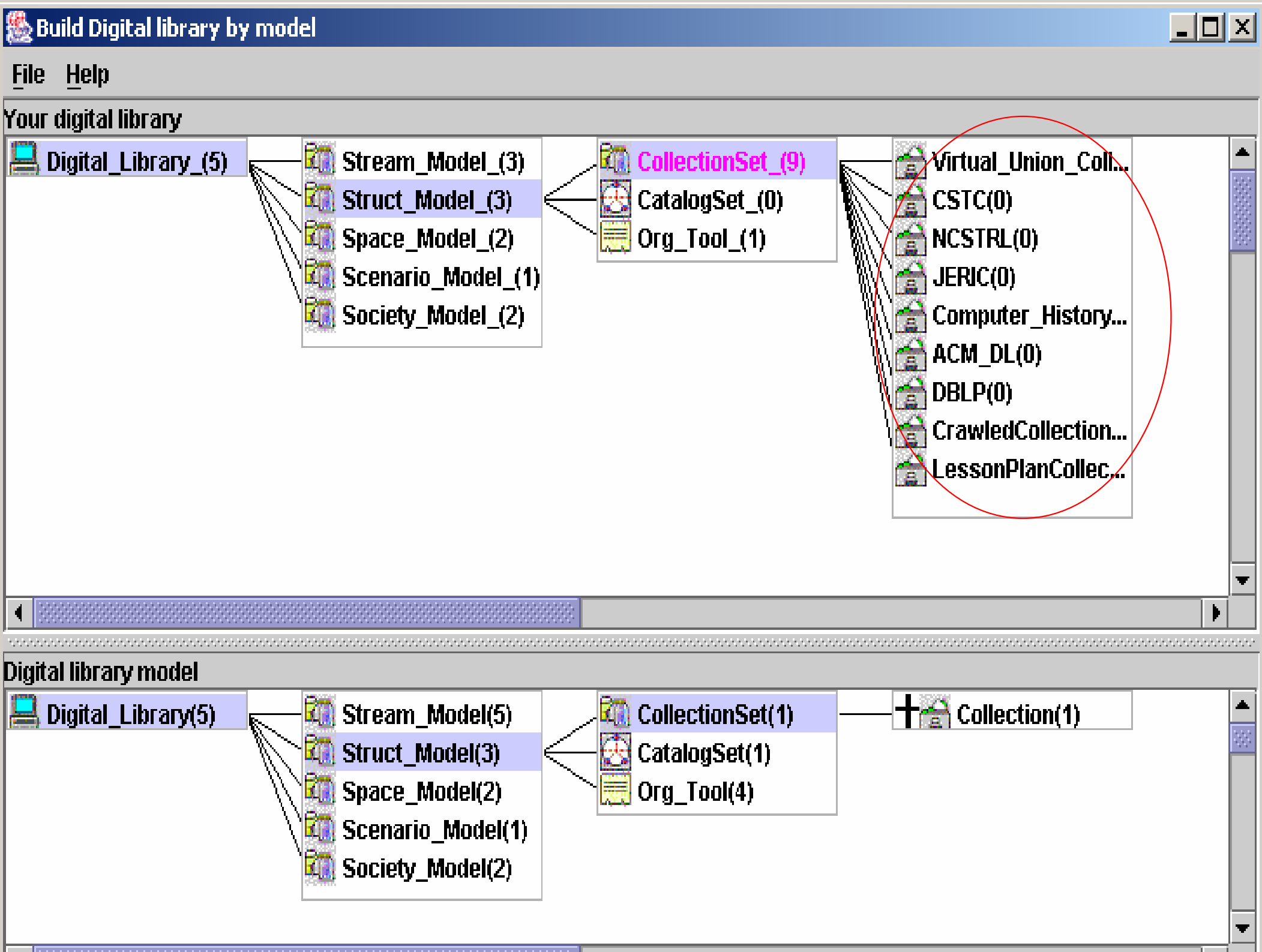
Infrastructure Services			**Information Satisfaction Services**
Repository-Building		***Add Value***	
Creational	Preservational		
Acquiring Cataloging Crawling (focused) Describing Digitizing Federating Harvesting Purchasing Submitting	Conserving Converting Copying/Replicating Emulating Renewing Translating (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

5S Framework and DL Development



5SLGen: Automatic DL Generation




Build Digital library by model
File
Help
Your digital library
Digital_Library_(5)
Stream_Model_(3)
Struct_Model_(3)
Space_Model_(2)
Scenario_Model_(1)
Society_Model_(2)
CollectionSet_(9)
CatalogSet_(0)
Org_Tool_(1)
Virtual_Union_Coll...
CSTC(0)
NCSTRL(0)
JERIC(0)
Computer_History...
ACM_DL(0)
DBLP(0)
CrawledCollection...
LessonPlanCollec...
Digital library model
Digital_Library(5)
Stream_Model(5)
Struct_Model(3)
Space_Model(2)
Scenario_Model(1)
Society_Model(2)
CollectionSet(1)
CatalogSet(1)
Org_Tool(4)
Collection(1)

Outline

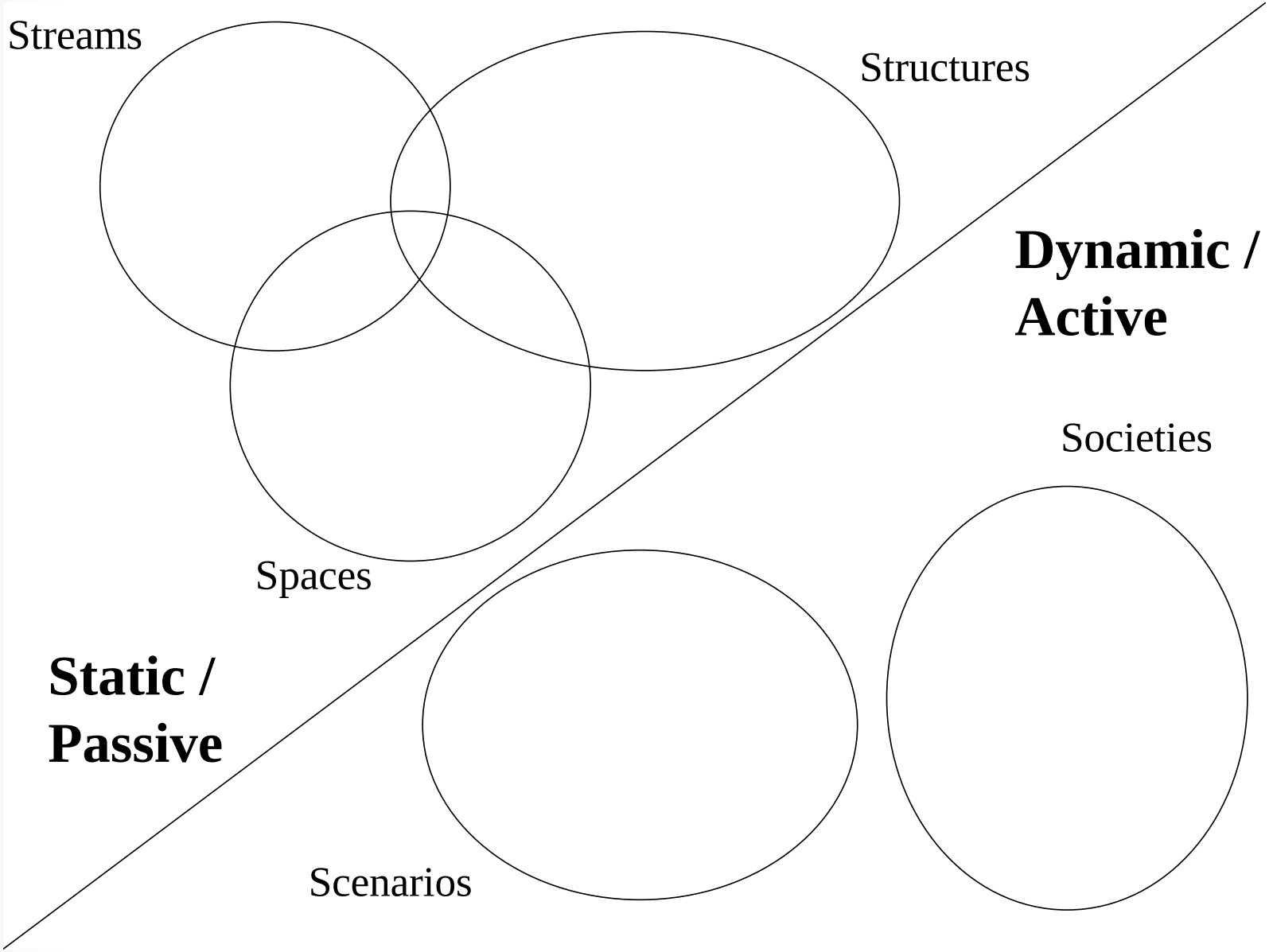
- Major Points of the Presentation
- ***Introduction to the 5S View of DLs***
  - ***Informal Definition of DLs***
  - Formal Definition of DLs
  - DL Ontology
- Defining a Quality Model for DLs
- Quality and the Information Life Cycle
- An XML Log Standard for DLs
- Conclusions and Future Work

Informal 5S 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**)


NATIONAL
GEOGRAPHIC
obiettivo
SPAZIO
un universo di immagini

Background: The 5S Model



5Ss

Models	Examples	Objectives
Stream	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; organization tools	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

Metamodels

- For “typical digital library”
  - Minimal DL
  - Starts with digital object (e.g., born digital)
- For scientific digital library, educational DL, cultural heritage DL, e-Gov DL, …
  - Ex.: archaeological DL - ETANA-DL
  - Starts with real object

Digital Objects (DOs)

- Born digital
- Digitized version of “real” object
  - Is the DO version the same, better, or worse?
  - Decision for ETDs: structured + rendered
- Surrogate for “real” object
  - Not covered explicitly in metamodel for a minimal DL
  - Crucial in metamodel for archaelogy DL

Metadata Objects (MDOs)

- MARC
- Dublin Core
- RDF
- IMS
- OAI (Open Archives Initiative)
- Crosswalks, mappings
- Ontologies
- Topics maps, concept maps

Repository

- Also called: digital rep., digital asset rep., digital object rep., institutional repository
- Stores and maintains digital objects (assets)
- Provides external interface for Digital Objects: Creation, Modification, Access
- Enforces access policies
- Provides for content type disseminations

Adapted from Slide by V. Chachra, VTLS

Other Key Definitions

- coll, catalog, service, archive, (minimal) DL
- See Gonçalves et al. in April 2004 *ACM Transactions on Information Systems (TOIS)*

Scope: see abstract

- Minimal DL: catalog, collection, digital object, metadata specification, repository, and services
- Quality dimensions: accessibility, accuracy, completeness, composability, conformance, consistency, effectiveness, efficiency, extensability, impact factor, pertinence, preservability, relevance, reliability, reusability, significance, similarity, and timeliness.
- Measurement characteristics: response time (with regard to efficiency), cost of migration (with respect to preservability), and number of service failures (to assess reliability)

Outline

- Major Points of the Presentation
- ***Introduction to the 5S View of DLs***
  - Informal Definition of DLs
  - ***Formal Definition of DLs***
  - DL Ontology
- Defining a Quality Model for DLs
- Quality and the Information Life Cycle
- An XML Log Standard for DLs
- Conclusions and Future Work

The 5S Formal Model

- A digital library is a 10-tuple (Streams, Structs, Sps, Scs, St2, Coll, Cat, Rep, Serv, Soc) in which:
  - Streams is a set of streams, which are sequences of arbitrary types (e.g., bits, characters, pixels, frames);
  - Structs is a set of structures, which are tuples, (G, ϕ), where G= (V, E) is a directed graph and ϕ: (V $\cup$ E) $\rightarrow$ *L* is a labeling function;
  - Sps is a set of spaces each of which can be a measurable, measure, probability, topological, metric, or vector space.

The 5S Formal Model (2)

- Scs = $\{sc_1, sc_2, \ldots, sc_d\}$ is a set of scenarios where each $sc_k = <e_{1k}(\{p_{1k}\}), e_{2k}(\{p_{2k}\}), \ldots, e_{d\_kk}(\{p_{d\_kk}\})>$ is a sequence of events that also can have a number of parameters $\{p_{ik}\}$. Events represent changes in computational states; parameters represent specific locations in a state and respective values.

- St2 is a set of functions Ψ: V× Streams→ (N × N) that associate nodes of a structure with a pair of natural numbers (a, b) corresponding to a portion of a stream.

The 5S Formal Model (3)

- Coll = $\{C_1, C_2, \ldots, C_f\}$ is a set of DL collections where each DL collection
  - Ck = $\{do_{1k}, do_{2k}, \ldots, do_{f\_kk}\}$ is a set of digital objects.
- Each digital object do is a tuple (h, SM, ST, StructuredStreams) where
  - h is a handle,
  - SM is a set of streams,
  - ST is a set of structural metadata specifications,
  - StructuredStreams is a set of StructuredStream functions defined from the streams in SM set and from the structures in the ST set.

The 5S Formal Model (4)

- Cat = $\{DM_{C\_1}, DM_{C\_2}, \ldots, DM_{C\_f}\}$ is a set of metadata catalogs for Coll where each metadata catalog $DM_{C\_k} = \{(h, mss_{hk})\}$, and $mss_{hk} = \{ms_{hk1}, ms_{hk2}, \ldots, ms_{hkn\_hk}\}$ is a set of descriptive metadata specifications. Each descriptive metadata specification ms_{hki} is a structure with atomic values (e.g., numbers, dates, strings) associated with nodes.

- A repository Rep = $\{(C_i, DM_{C\_i})\}$ (i=1 to f) is a set of pairs (collection, metadata catalog)
  - It is assumed there exists operations to manipulate them (e.g., get, store, delete).

The 5S Formal Model (5)

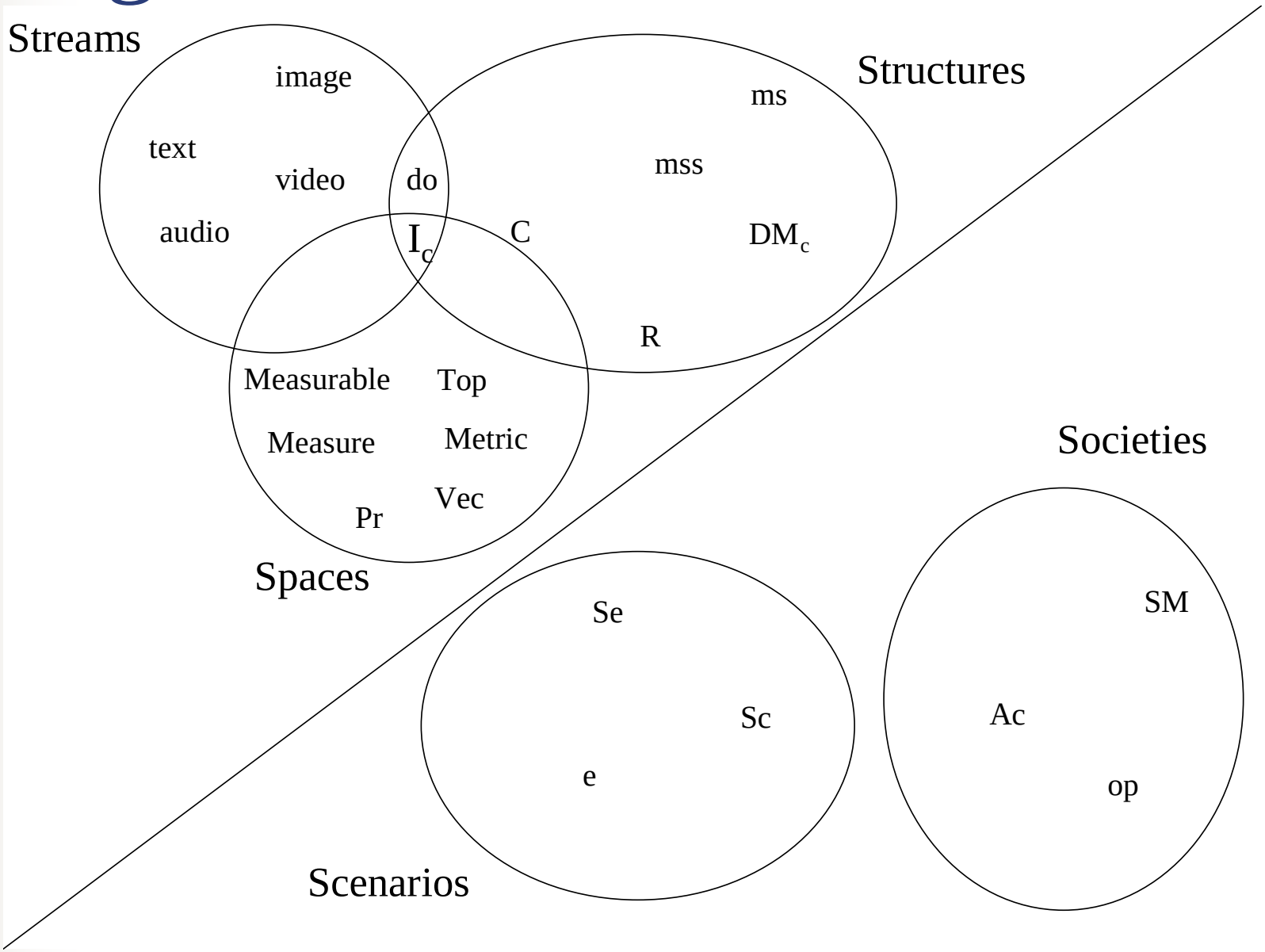
- Serv = {Se_1, Se_2, …, Se_s} is a set of services where each service Sek = {sc_{1k}, .., sc_{s_kk}} is described by a set of related scenarios.

- Soc = (C, R) where C is a set of communities and R is a set of relationships among communities. SM = {sm1, sm2, …, smj}, and Ac = {ac1, ac2, …, acr } are two such communities where the former is a set of service managers responsible for running DL services and the latter is a set of actors that use those services.
  - Being basically an electronic entity, a member smk of SM distinguishes itself from actors by defining or implementing a set of operations {op1k, op2k, …, opnk} $\subset$ smk. Each operation opik of smk is characterized by a triple (nik, sigik, impik), where nik is the operation's name, sigik is the operation's signature (which includes the operation's input parameters and output), and impik is the operation's implementation. These operations define the capabilities of a service manager smk.

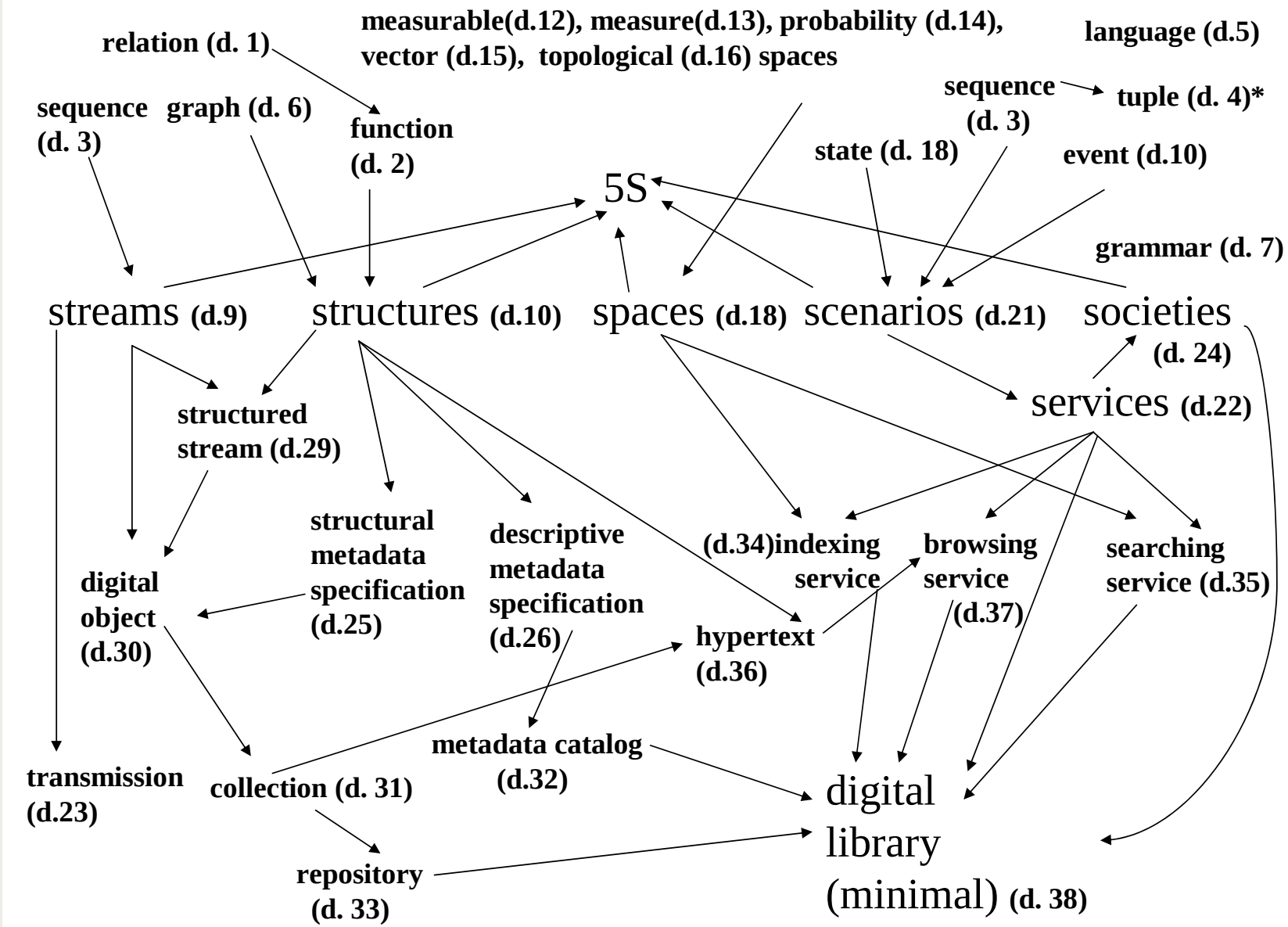
The 5S Formal Model (6)

- Soc = (C, R) where C is a set of communities and R is a set of relationships among communities. SM = {sm_1, sm_2, …, sm_j}, and Ac = {ac_1, ac_2, …, ac_r } are two such communities where the former is a set of service managers responsible for running DL services and the latter is a set of actors that use those services.
  - Being basically an electronic entity, a member sm_k of SM distinguishes itself from actors by defining or implementing a set of operations {op_{1k}, op_{2k}, …, op_{nk}} $\subset$ sm_k. Each operation op_{ik} of sm_k is characterized by a triple (n_{ik}, sig_{ik}, imp_{ik}), where n_{ik} is the operation's name, sig_{ik} is the operation's signature (which includes the operation's input parameters and output), and imp_{ik} is the operation's implementation. These operations define the capabilities of a service manager sm_k.

Background


Streams
image
text
video
audio
do
I_c
C
Structures
ms
mss
DM_c
R
Measurable
Top
Measure
Metric
Vec
Pr
Spaces
Societies
SM
Ac
op
Se
Sc
e
Scenarios

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



Reducing confusion, misnaming

- A “document” is a stream, with a superimposed or externally understood structure, along with a use scenario.
  - Structures: grammatical, rhetorical, markup
- This could help us better address
  - “Semi-structured information”
  - “Unstructured information”

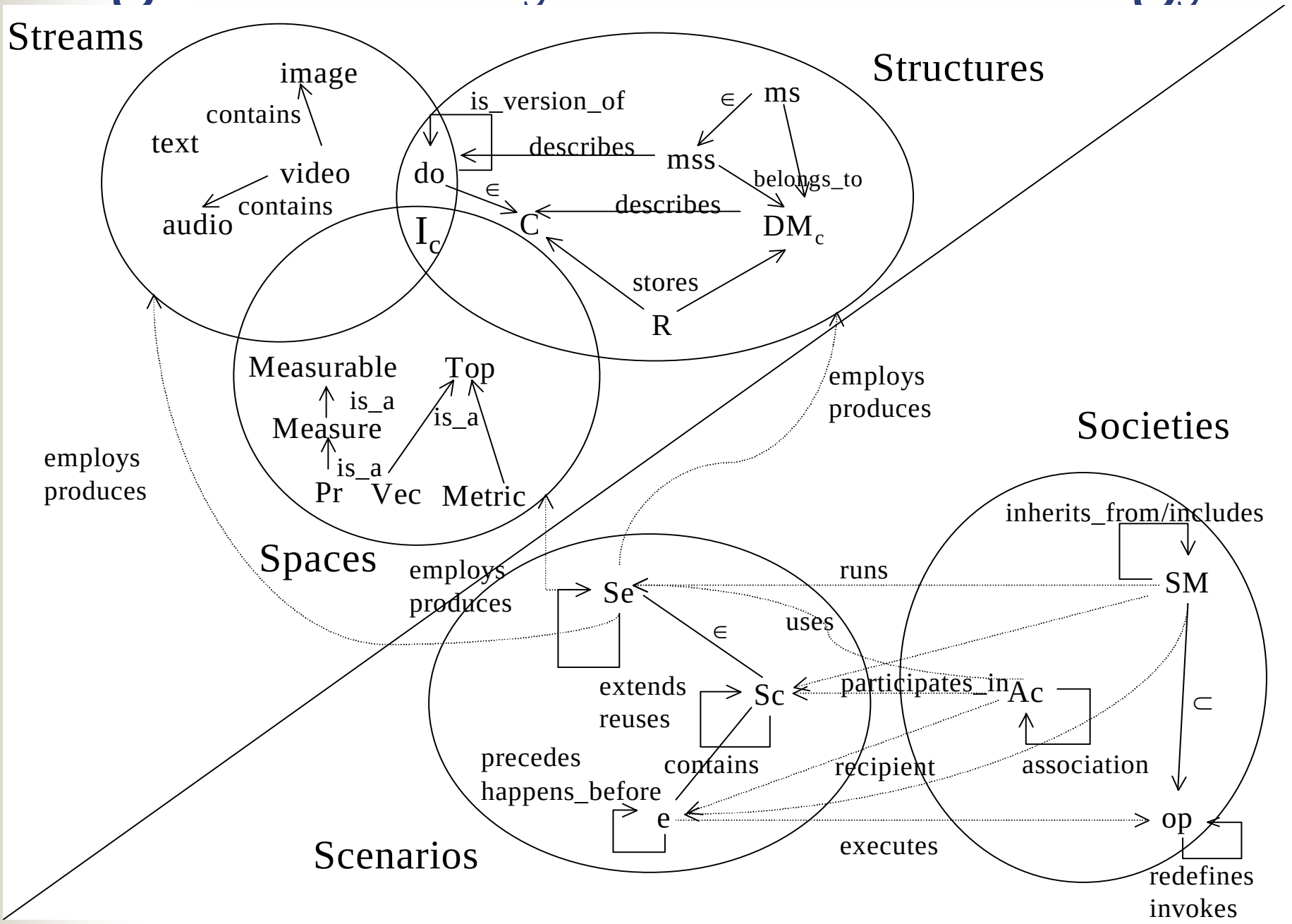
Glossary: Concepts in the Minimal DL and Representing Symbols

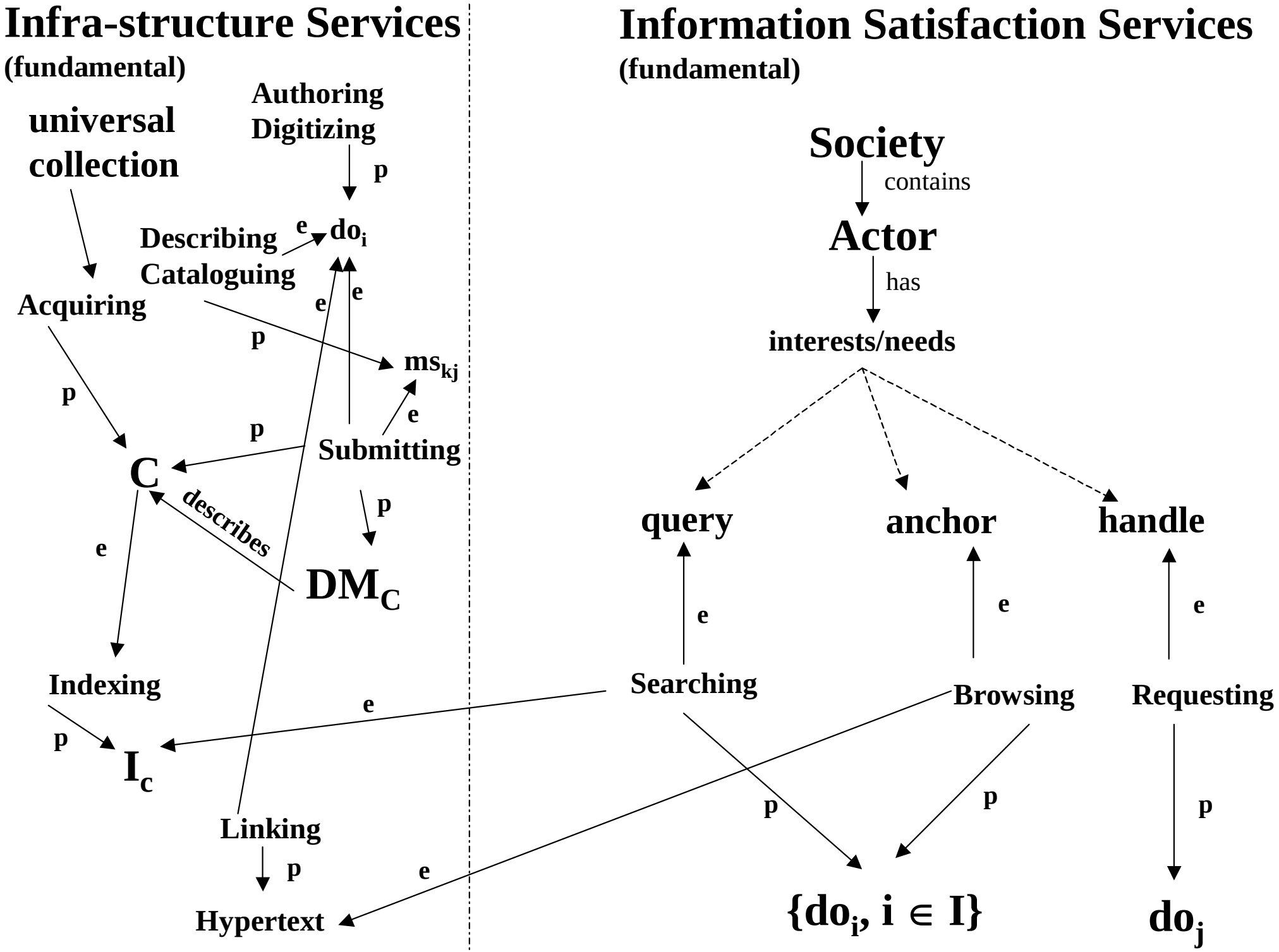
Concept	**Symbol**
Digital object	do
Metadata specification	ms
Set of metadata specifications	mss
Collection	C
Catalog	DM_C
Repository	S
Event	e
Scenario	Sc
Services	Se
Actor	Ac
Service Manager	SM
Operation	op
Society	Soc

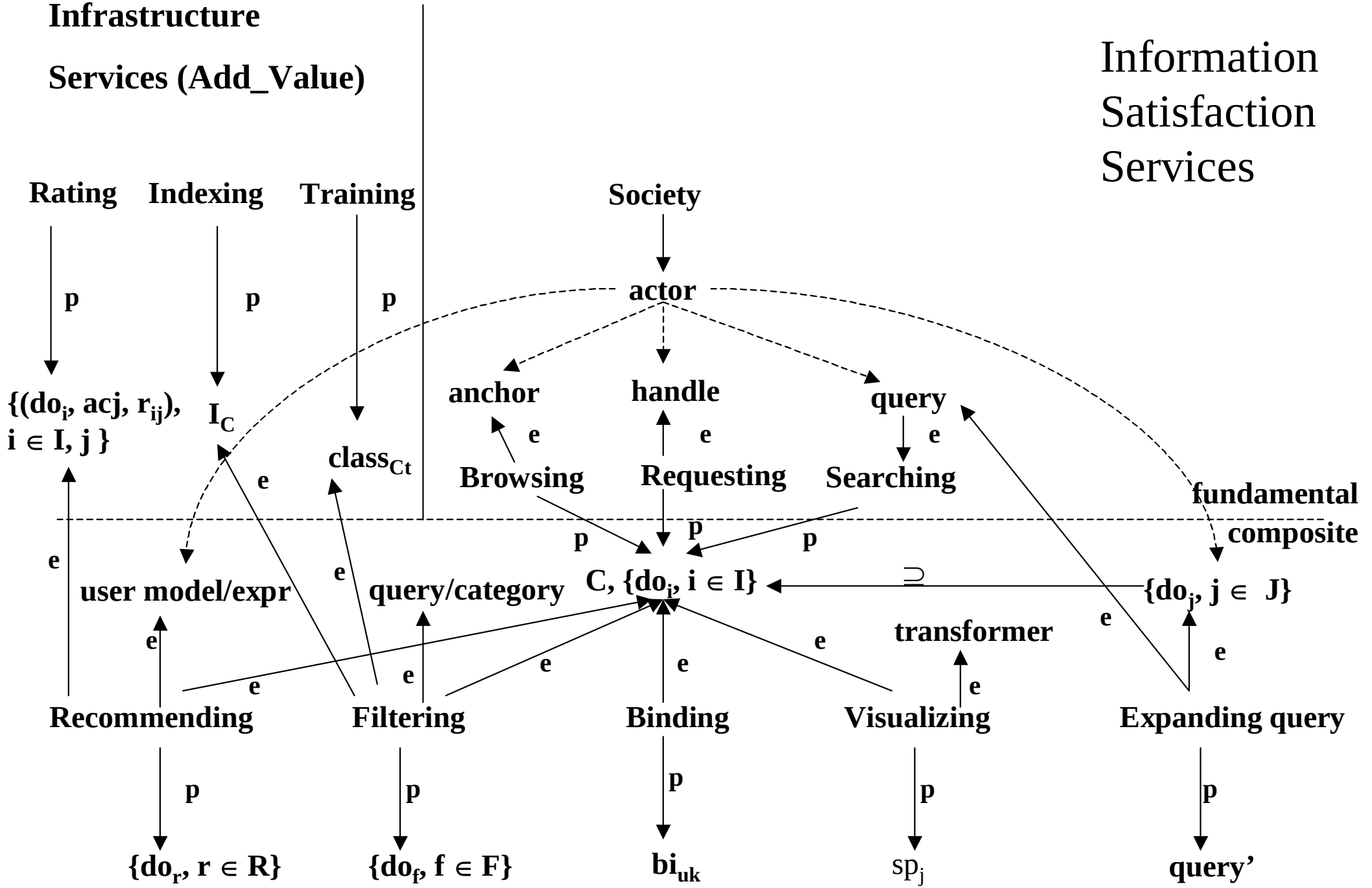
Outline

- Major Points of the Presentation
- ***Introduction to the 5S View of DLs***
  - Informal Definition of DLs
  - Formal Definition of DLs
  - ***DL Ontology***
- Defining a Quality Model for DLs
- Quality and the Information Life Cycle
- An XML Log Standard for DLs
- Conclusions and Future Work

Digital Library Formal Ontology




Infra-structure Services
(fundamental)
universal collection
Authoring
Digitizing
p
Describing
Cataloguing
e
do_i
Acquiring
e
e
p
ms_kj
p
e
p
Submitting
C
describes
p
e
DM_C
Indexing
p
I_c
e
Linking
p
e
Hypertext
Information Satisfaction Services
(fundamental)
Society
contains
Actor
has
interests/needs
query
anchor
handle
e
e
e
Searching
Browsing
Requesting
p
p
p
{do_i, i ∈ I}
do_j


Infrastructure
Services (Add_Value)
Information
Satisfaction
Services
Rating
Indexing
Training
Society
actor
p
p
p
{(do_i, acj, r_ij), i ∈ I, j }
I_C
class_Ct
anchor
handle
query
e
e
e
e
Browsing
Requesting
Searching
fundamental
composite
p
p
p
C, {do_i, i ∈ I}
⊇
{do_j, j ∈ J}
user model/expr
query/category
transformer
Recommending
Filtering
Binding
Visualizing
Expanding query
p
p
p
p
p
{do_r, r ∈ R}
{do_f, f ∈ F}
bi_uk
sp_j
query'

Outline

- Major Points of the Presentation
- Introduction to the 5S View of DLs
  - Informal Definition of DLs
  - Formal Definition of DLs
  - DL Ontology
- ***Defining a Quality Model for DLs***
- Quality and the Information Life Cycle
- An XML Log Standard for DLs
- Conclusions and Future Work

Defining 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:
      - Proposing and formalizing new quality measures for DLs
      - Formalizing traditional measures within our 5S framework
      - Contextualizing these measures within the Information Life Cycle

Defining Quality in Digital libraries

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

Defining Quality in Digital Libraries

- Structure of this part of presentation
  - For each quality metric
    - Discussion about the metric
      - Meaning, use, etc.
    - Definition of numerical measure
    - Example of Use

Digital Objects: Accessibility

- A digital object is *accessible* by an DL actor or patron, if it exists in the collections of the DL, the repository is able to retrieve the object, and:
  1) an overly restrictive rights management property of a metadata specification does not exist for that object; or
  2) if it exists, the property does not restrict access to the particular society to which the actor belongs or to that actor in particular.

Digital Objects: Accessibility

- Accessibility $acc(do_x, ac_y)$ of digital object do_x to actor ac_y is:
  - 0, if there is no collection C in the DL so that $do_x \in C$;
  - otherwise $acc(do_x, ac_y) = \sum_{z \in struct\_streams(dox)} r_z(ac\_y))/|struct\_streams(do_x)|$, where:
    - $r_z(ac_y)$ is a rights management rule defined as an indicator function:
      - 1, if
        - z has no access constraints; or
        - z has access constraints and $ac_y \in cm_z$, where $cm_z \in Soc(1)$ is a community that has the right to access z; and
      - 0, otherwise

Digital Objects: Accessibility

- VT ETD Collection

First Letter of Author's Name	Unrestricted	Restricted	Mixed	Degree of accessibility for users not on the VT community
A	164	50	5	mix(0.5, 0.5, 0.167, 0.1875, 0.6)
B	286	102	3	mix(0.5,0.5, 0.13)
C	231	108	7	mix (0.11, 0.5, 0.5, 0.5, 0.33, 0.09, 0.33)
D	159	54	2	mix(0.875, 0.666)
E	67	26	1	mix(0.5)
F	88	39	2	mix(0.375, 0.09)
G	166	72	2	mix(0.666,0.5)
H	225	91	3	mix(0.66, 0.5, 0.235)
I	20	8	1	mix(0.5)
J	84	36	2	mix(0.5, 0.6)
K	166	69	2	mix(0.5, 0.5)
L	189	68	6	mix(0.153, 0.33, 0.5, 0.5, 0.94)
M	299	115	9	mix(0.5, 0.5, 0.5, 0.041, 0.5, 0.5, 0.5, 0.117, 0.5)

Digital Objects: Accessibility

N	74	16	1	mix(0.8)
O	45	19	2	mix(0.5, 0.125)
P	172	71	3	mix(0, 0, 0.33)
Q	13	6	0	mix = none
R	158	71	3	mix(0.66, 0.5, 0.5)
S	398	159	8	mix(0.66, 0.5, 0.5, 0.6, 0.33, 0.66, 0.33, 0.6)
T	111	49	1	mix(0.13)
U	9	7	0	mix = none
V	63	20	0	mix = none
W	191	81	5	mix (0.5, 0.22, 0.38, 0.875, 0.5)
X	11	5	0	mix = none
Y	38	9	3	mix(0.5, 0.5, 0.125)
Z	47	17	2	mix(0.5, 0.5)
All	3474	1368	73	

Digital Objects: Pertinence

- Let Inf(do_i) represent the ``information'' (not physical) carried by a digital object or any of its (metadata) descriptions, IN(ac_j) be the information need of an actor and $Context_{jk}$ be an amalgam of societal factors which can impact the judgment of pertinence by ac_j at time k.
  - These include among others, time, place, the actor's history of interaction, task in hand, and a range of other factors that are not given explicitly but are implicit in the interaction and ambient environment.

Digital Objects: Pertinence

- Let's define two sub-communities of actors, users and external-judges $\in$ Ac, as:
  - users: set of actors with an information need who use DL services to try to fulfill/satisfy that need
  - external-judges: set of actors responsible for determining the relevance of a document to a query.

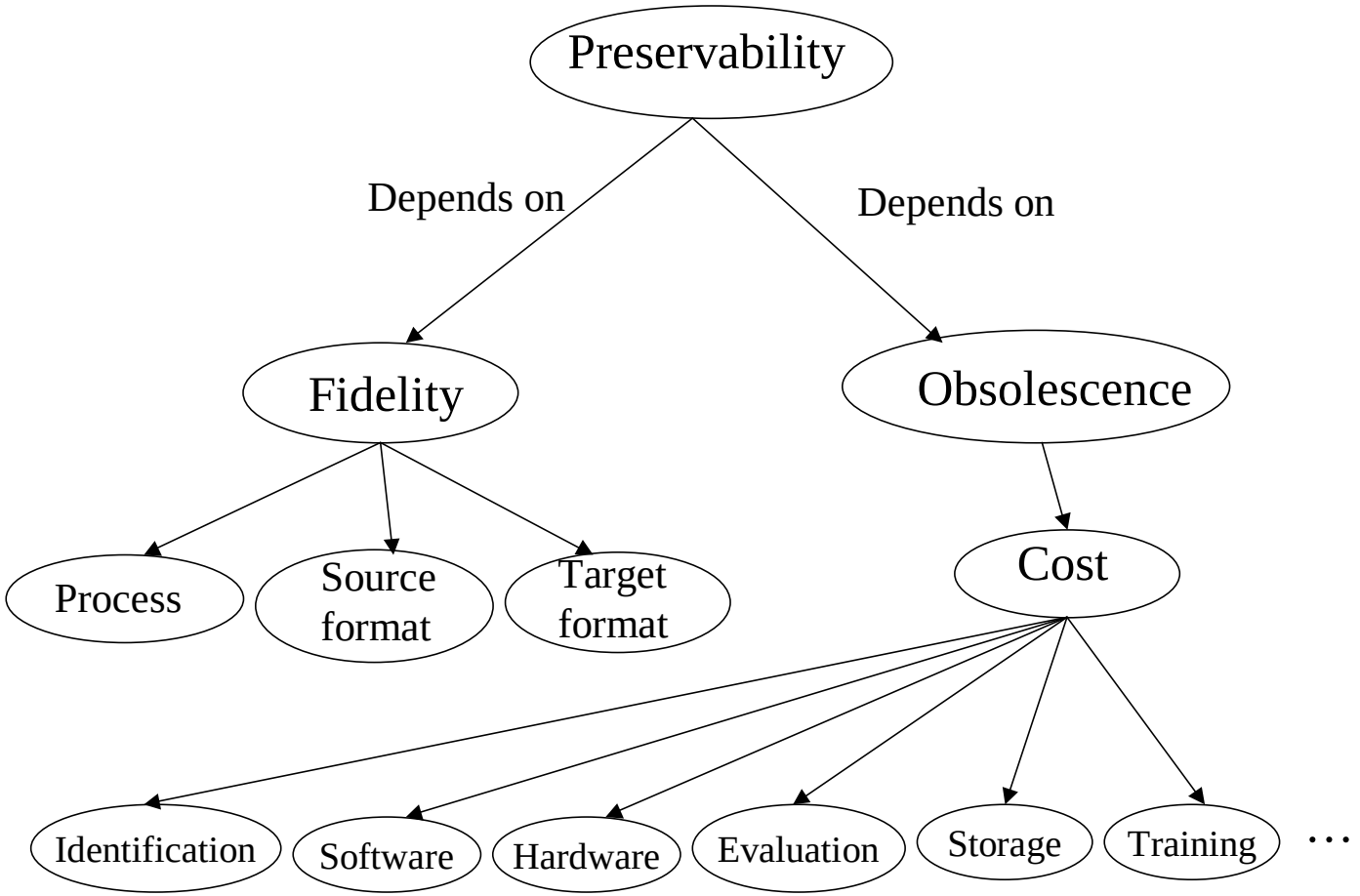
- Let's also constrain that a member of external-judges can not judge the relevance of a document to a query representing her own information need, i.e., at the same point in time users $\cap$ external-judges = $\varnothing$.

Digital Objects: Pertinence

- The pertinence of a digital object to a user ac_j is an indicator function Pertinence(do_i, ac_j): $Inf(do_i) \times IN(ac_j) \times Context_{jk}$ defined as:
  - 1, if $Inf(do_i)$ is judged by ac_j to be informative with regards to $IN(ac_i)$ in context $Context_{jk}$;
  - 0, otherwise

Digital Objects: Preservability

- Factors in Preservability



Digital Objects: Preservability

- Preservability(do_i, dl) =

  (fidelity of migrating(do_i,$format_x$, $format_y$), obsolescence(do_i, dl)).

  - fidelity(do_i, $format_x$, $format_y$) = 1/ distortion(p($format_x$, $format_y$))

  - obsolescence(do_i, dl) = cost of converting/migrating object within the context of the specific dl

Digital Objects: Relevance

- Relevance (do_i,q)

1, if doi is judge by *external-judge* to be relevant to q

0, otherwise

- Relevance Estimate
  - $Rel(do_i,q) = do_i^{\rightarrow} \bullet dj^{\rightarrow} / |do_i^{\rightarrow}| \times |q^{\rightarrow}|$
- Objective, public, social notion
  - Established by a general consensus in the field, not subjective, private judgment by an actor with an information need

Digital Objects: Similarity

- reflect the relatedness between two or more digital objects
- Used in many services (e.g., classification, find similar, etc)

Digital Objects: Similarity

- Metrics
  - Content-based
    - Cosine(di, dj)
      - $do_i^{\rightarrow} \bullet d_j^{\rightarrow} / |do_i^{\rightarrow}| \times |do_j^{\rightarrow}|$
    - Bag-of-words(d_i,d_j)
      - $|W(di) \cap W(dj)| / |W(di)|$
    - Okapi(di,dj) (see draft)

Digital Objects: Similarity

- Metrics
  - Citation-based
    - Co-citation
      - $cocit(d_i,d_j) = |Pd_i \cap Pd_j| / max\ P$
    - Bibliographic coupling
      - $bibcoup(d_i, d_j) = |Cd_i \cap Cd_j| / max\ Cd$
    - Amsler
      - $Amsler(di, dj) = |(Pdi \cup Cdi) \cap (Pdj \cap Cdj)| / max\ P \cup Cd$

Digital Objects: Similarity

Highest degree of cocitation	Publication	Year
A unified lattice model for static analysis of programs by construction or approximation of fixpoints	4th ACM SIGACT-SIGPLAN	1977
Active messages: a mechanism for integrated communication and computation	19th annual int. symposium on Computer architecture	1992
Improving direct-mapped cache performance by the addition of a small fully-associative cache and prefetch buffers	17th annual international symposium on Computer Architecture	1990
Computer programming as an art	CACM	1974
The SPLASH-2 programs: characterization and methodological considerations	22nd annual international symposium on Computer architecture	1995
ATOM: a system for building customized program analysis tools	ACM SIGPLAN '94	1994
Analysis of pointers and structures	Proceedings of the conference on Programming language design and implementation	1990
Revised report on the algorithmic language scheme	| ACM SIGPLAN Notices (Issue)	1986
The directory-based cache coherence protocol for the DASH multiprocessor	17th annual international symposium on Computer Architecture	1990

Digital Objects: Similarity

Highest degree of bibliographic coupling	**publication**	**date**
Query evaluation techniques for large databases	CSUR	1993
Compiler transformations for high-performance computing	CSUR	1994
On randomization in sequential and distributed algorithms	CSUR	1994
External memory algorithms and data structures: dealing with massive data	CSUR	2001
A schema for interprocedural modification side-effect analysis with pointer aliasing	TOPLAS	2001
Complexity and expressive power of logic programming	CSUR	2001
Computational geometry: a retrospective	ACM symposium on Theory of computing	1994
Research directions in object-oriented database systems	ACM SIGACT-SIGMOD-SIGART symposium	
Cache coherence in large-scale shared-memory multiprocessors: issues and comparisons	CSUR	1993

Digital Objects: Similarity

- Distributions

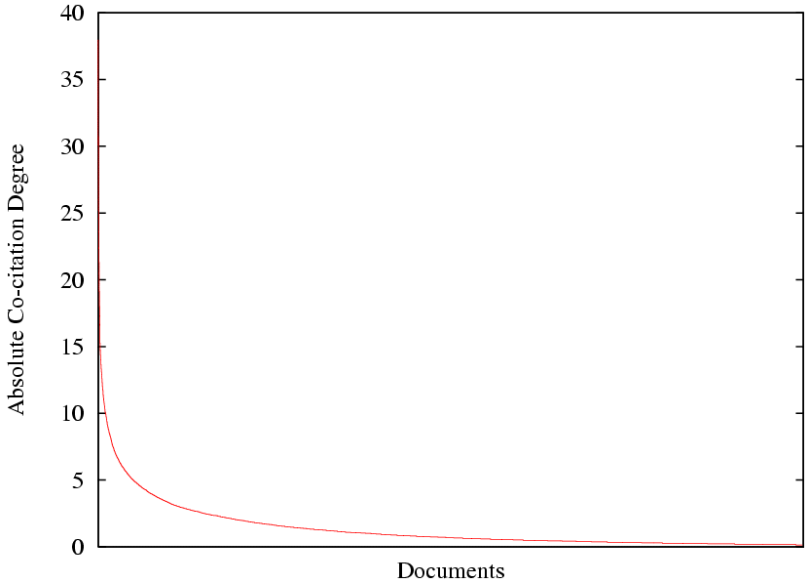


Figure 3(a)

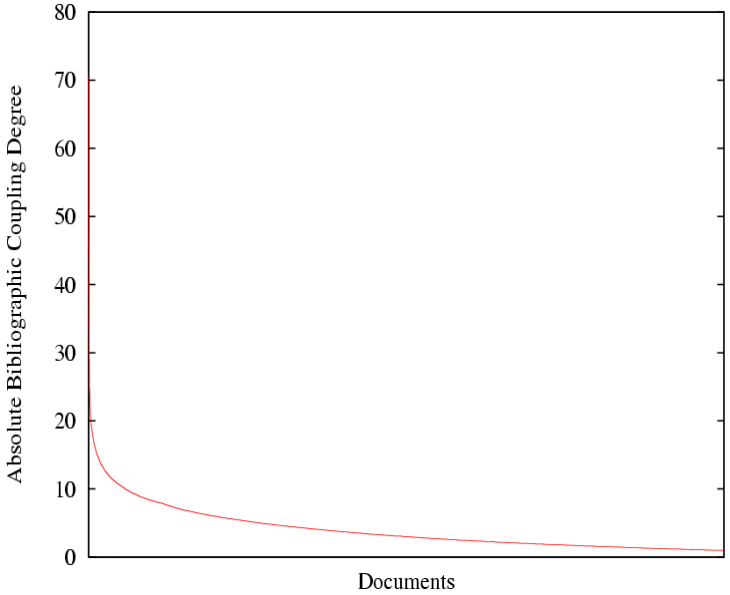
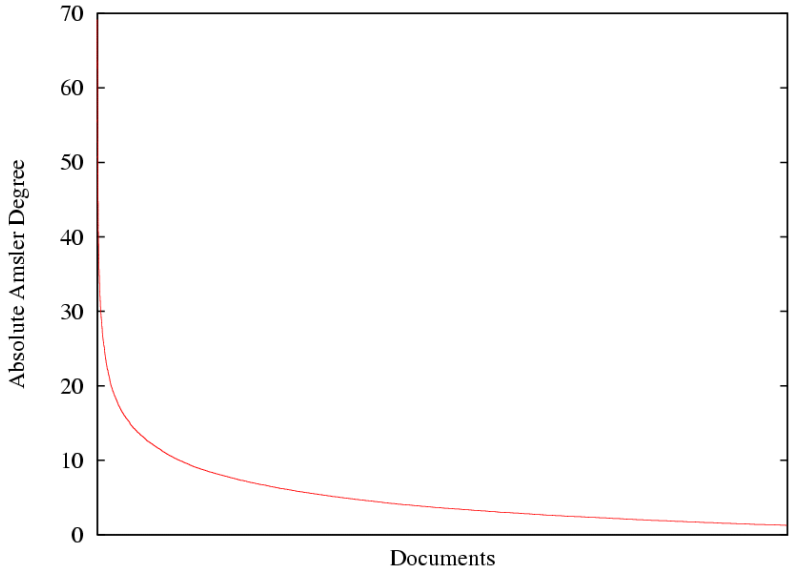


Figure 3(b)



Digital Objects: Similarity

- Application: Automatic classification with kNN

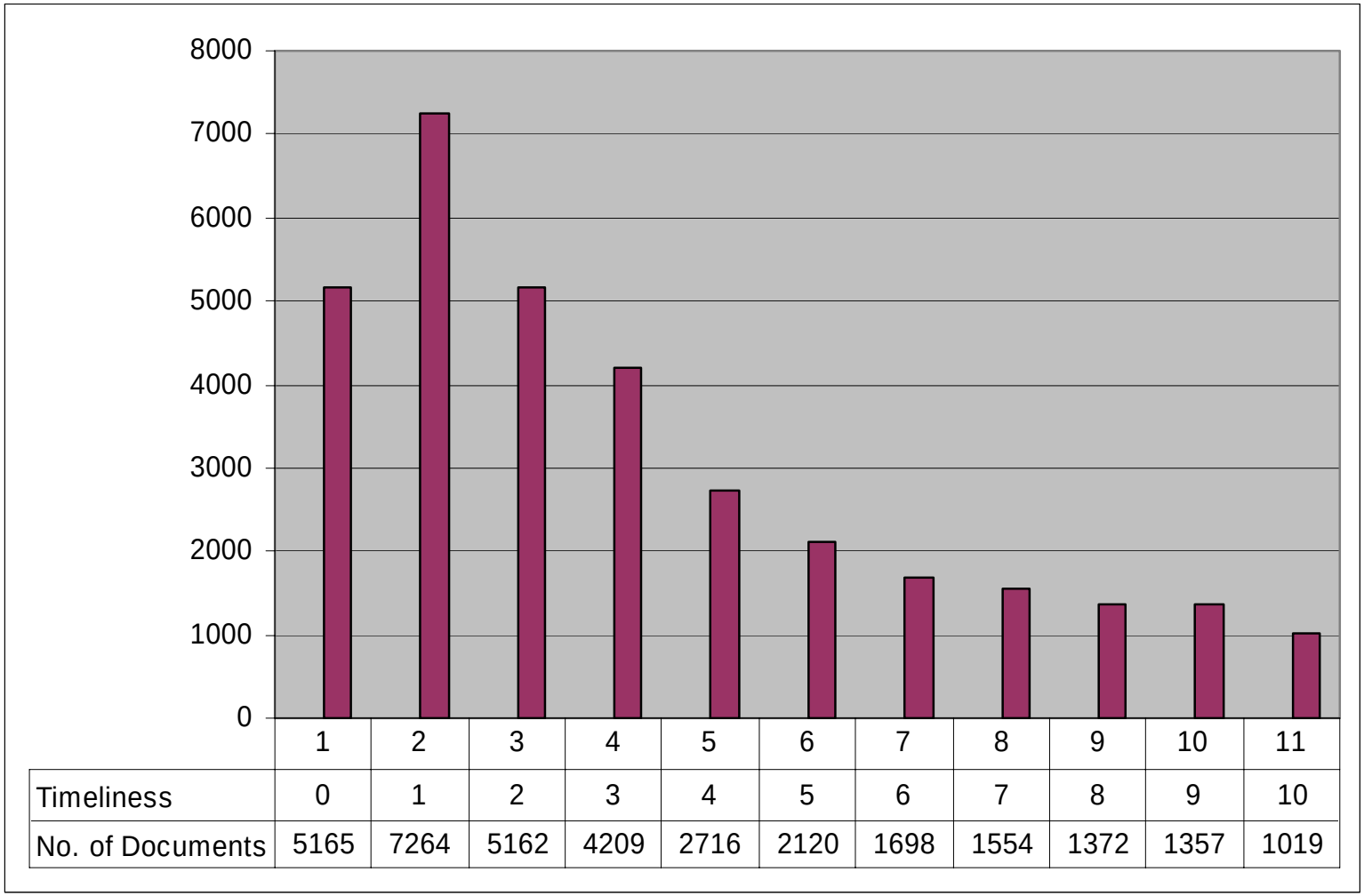
Evidence	**Macro F1 (30%)**
Abstract_BagOfWords	0.195
Abstract_Cosine	0.343
Abstract_Okapi	0.339
Bib_Coup	0.347
Amsler	0.412
Co-citation	0.273
Title_BagOfWords	0.492
Title_Cosine	0.525
Title_Okapi	0.525

Digital Object: Timeliness

- (current time or time of last freshening) – time of the latest citation, if object is ever cited
- age = (current time or time of last freshening) – (creation time or publication time) , if object is never cited
- Time of last freshening = time of the creation/publication of most recent object in the collection to which do_i belongs

Digital Objects: Timeliness

- ACM Digital Library



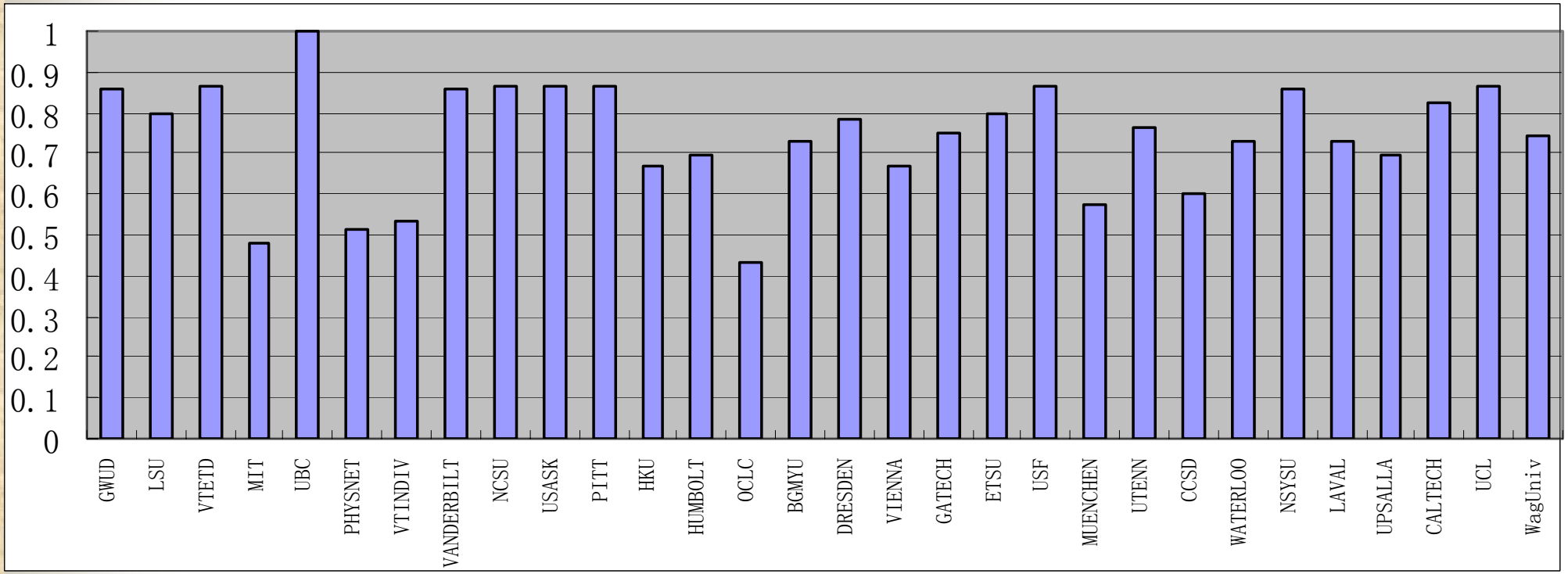
	1	2	3	4	5	6	7	8	9	10	11
Timeliness	0	1	2	3	4	5	6	7	8	9	10
No. of Documents	5165	7264	5162	4209	2716	2120	1698	1554	1372	1357	1019

Metadata Specifications and Metadata Format: Completeness

- Refers to the degree to which values are present in the description, according to a metadata standard. As far as an individual property is concerned, only two situations are possible: either a value is assigned to the property in question, or not.
- Metric
  - Completeness(ms_x) = 1 - (no. of missing attributes in ms_x/ total attributes of the schema to which ms_x conforms)

Metadata Specifications and Metadata Format: Completeness

- OCLC NDLTD Union catalog

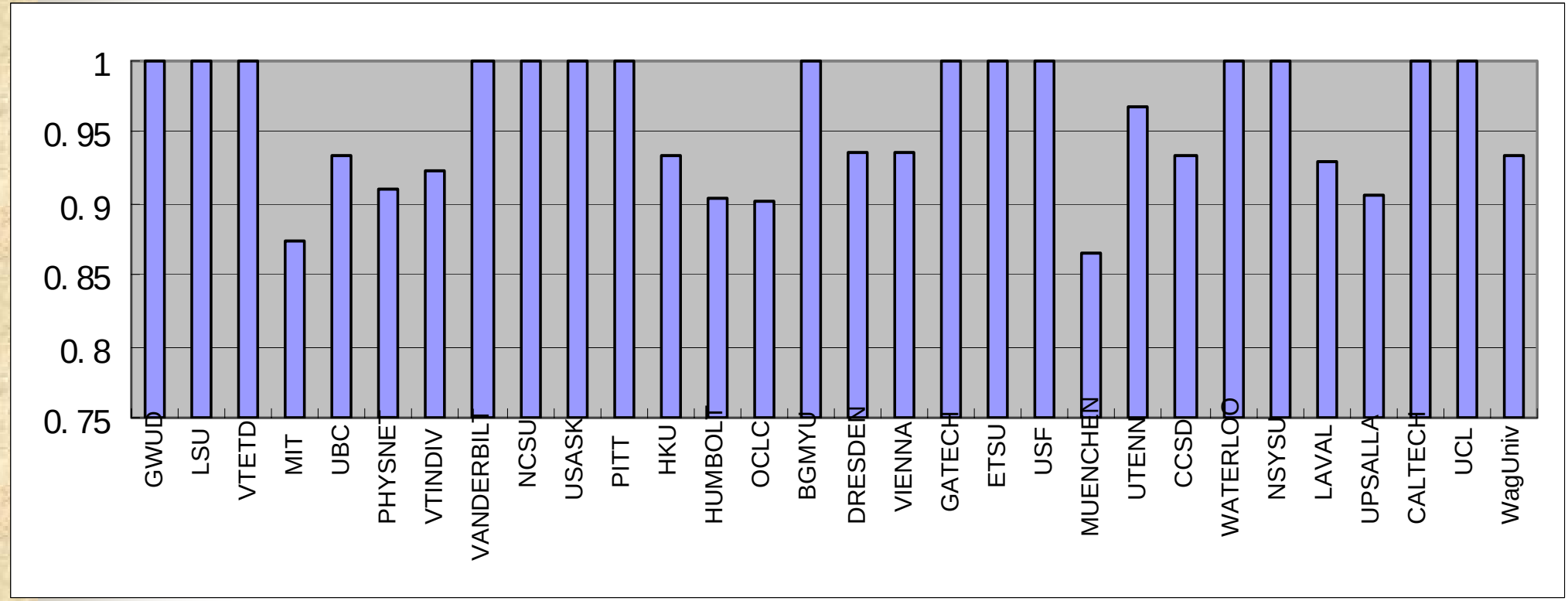


Metadata Specifications and Metadata Format: Conformance

- An attribute att_{xy} of a metadata specification ms_x is conformant to a metadata format/standard if:
  - it appears at least once, if att_{xy} is marked as *mandatory*, and;
  - its value is from the domain defined for att_{xy};
  - it does not appear more than once, if it is not marked as *repeatable*.
- Metric
  - Conformance(ms_x) = (Σ($\forall$attribute att_{xy} of ms_x) degree of conformance of att_{xy})/ total attributes).

Metadata Specifications and Metadata Format: Conformance

- Based on ETD-MS



Collection, Metadata Catalog, and Repository: Collection Completeness

- A *complete* DL collection is one which contains all the pertinent existing digital objects.

- Metric
  - completeness(C_x) = |C_x| /| “ideal collection”|

Collection, Metadata Catalog, and Repository: Collection Completeness

ACM Guide	
Journal (articles)	256527
Proceeding (papers)	299850
Book (chapters)	107870
Theses	46098
Tech. Reports	25081
Bibliographies	2
Plays	1
	735429

Collection	Degree of Completeness
ACM Guide	1
DBLP	0.652
CITIDEL(DBLP + ACM + NCSTRL + NDLTD-CS)	0.467
IEEE-DL	0.168
ACM-DL	0.146

Catalog Completeness/Consistency

- Completeness(DM_C)=

1 – (no. of do's without a metadata specification/size of the described collection)

- Consistency(DM_C)=

0, if there is at least one set of metadata specifications assigned to more than one digital object

1, otherwise

Repository Completeness and Consistency

- Completeness (Rep) =

Number of collections in the repository/ideal number of collections

- Consistency(Rep) =

1, if the consistency of all the repositories' catalogs with respect to their described collection is 1

0, otherwise

Services: Efficiency/ Effectiveness

- Effectiveness
  - Very common measures: Precision, Recall, F1, 10-precision, R-Precision
  - Other services may have different measures: e.g., Recommending, etc.
- Efficiency (duration of a service event):
  - Let t(e) be the time of an event e, e_{ix} and e_{fx} be the first and the last event of service se_x . The efficiency of service se_x is defined as:
    - $Efficiency(se_x) = t(e_{fx}) - t(e_{ix})$

Services: Extensibility and Reusability

- A service *Y reuses* a service *X* if the behavior of *Y* incorporates the behavior of X.
- A service *Y* extends a service *X* if it subsumes the behavior of *X* and potentially includes additional subflows of events.

Services: Extensibility and Reusability (2)

- Metrics
  - Macro-Reusability(Serv) = ($\sum$ reused(se_i), $se_i \in$ Serv)/ |Serv|, where reused is a indicator function defined as : 1, if $\exists$ sm_j, se_j reuses s_i; 0, otherwise.
  - Micro-Reusability(Serv) = ($\sum$ LOC(sm_x) * reused(se_i), $sm_x \in$ SM, $se_i \in$ Serv, se_x runs se_i)/ |$\sum$LOC(sm), $\forall$sm $\in$ SM|, where LOC corresponds to the number of lines of code of a service manager

Services: Extensibility and Reusability

Service	Component Based	LOC for implementing service	LOC reused from component	Total LOC
Searching – Back-end	Yes	-	1650	1650
Search Wrapping	No	100	-	100
Recommending	Yes	-	700	700
Recommend Wrapping	No	200	-	200
Annotating – Back-end	Yes	50	600	600
Annotate Wrapping	No	50	-	50
Union Catalog	Yes	-	680	680
User Interface Service	No	1800	-	1600
Browsing	No	1390	-	1390
Comparing (objects)	No	650	-	650
Marking Items	No	550	-	550
Items of Interest	No	480	-	480
Recent Searches/Discussions	No	230	-	230
Collections Description	No	250	-	250
User Management	No	600	-	600
Framework Code	No	2000	-	2000
Total		**8280**	**3630**	**11910**

Macro-Reusability = 3/16 = 0.187

Micro-Reusability = 3630 / 11910 = 0.304

Services: Reliability

- Def: 1 – no. of failures/no. of accesses
- Failure is an event that
  - was supposed to happen in a scenario but did not;
  - did happen, but did not execute some of its operations
  - did happen, where the operations were executed, but the results were not the expected ones.

Services: Reliability

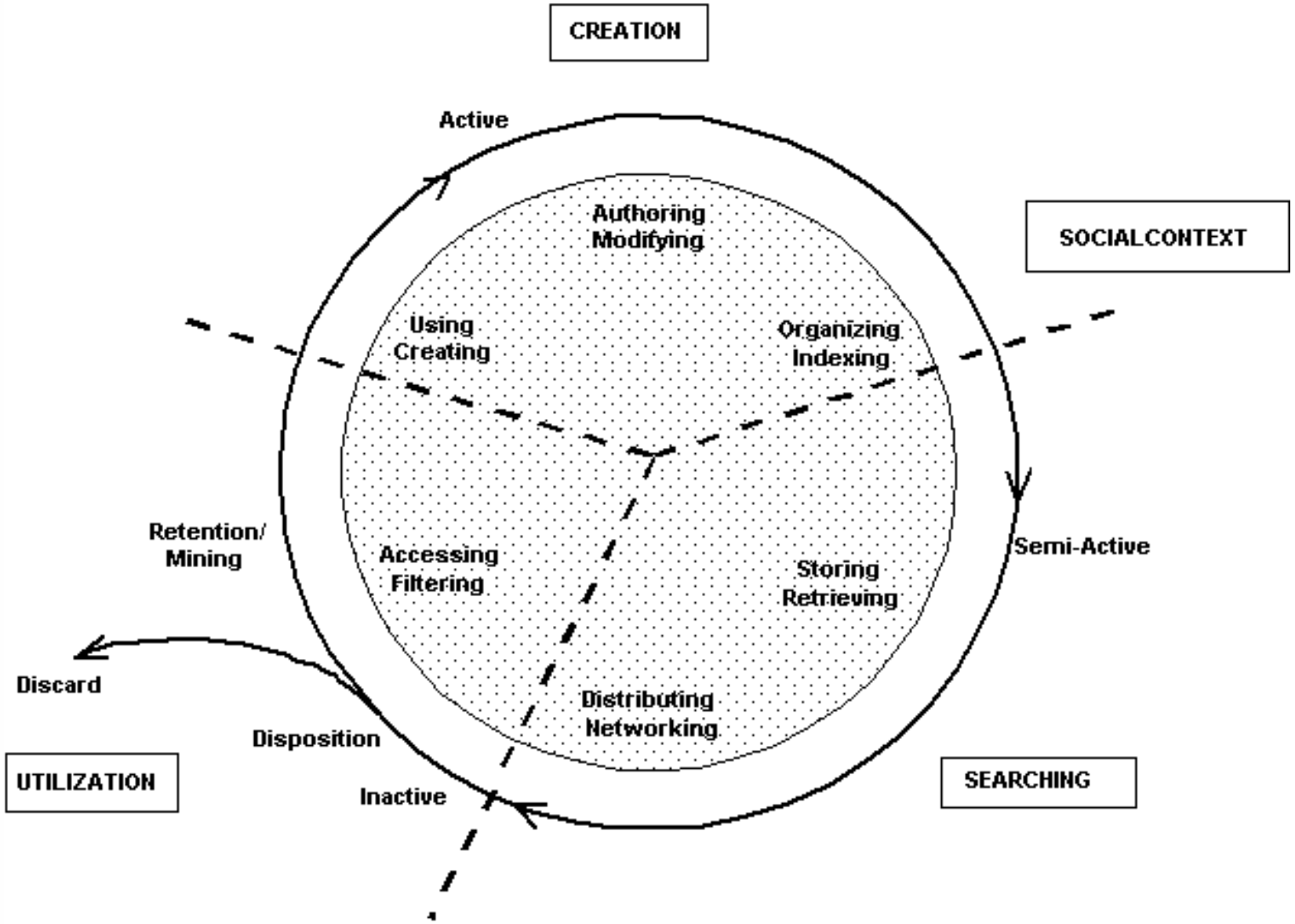
- CITIDEL (NSDL collection – computing/IT)

CITIDEL service	No. of failures/no. of accesses	Reliability
searching	73/14370	0.994
browsing	4130/153369	0.973
requesting (getobject)	1569/318036	0.995
structured search	214/752	0.66
contributing	0/980	1

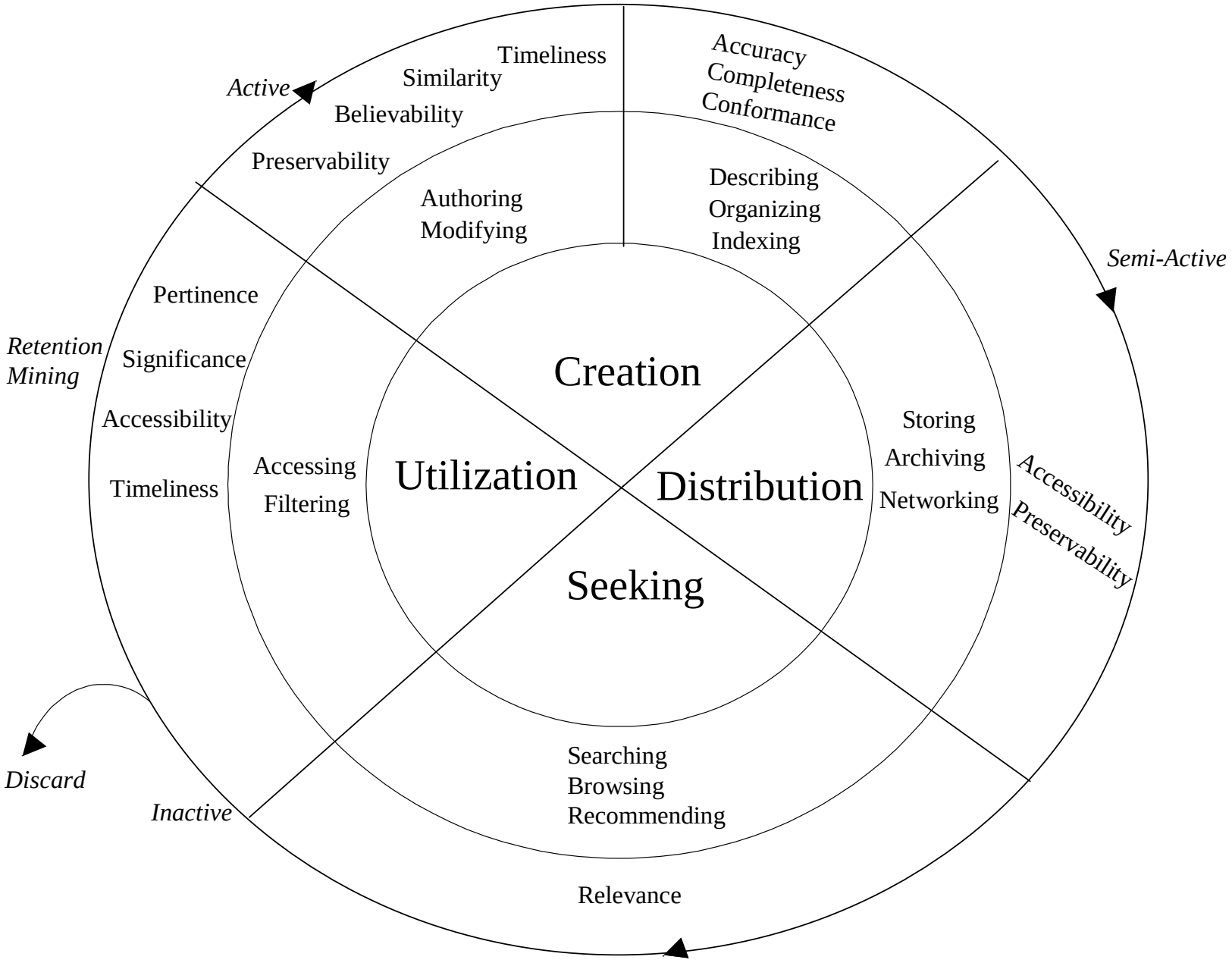
Outline

- Major Points of the Presentation
- Introduction to the 5S View of DLs
  - Informal Definition of DLs
  - Formal Definition of DLs
  - DL Ontology
- Defining a Quality Model for DLs
- ***Quality and the Information Life Cycle***
- An XML Log Standard for DLs
- Conclusions and Future Work

Information Life Cycle



Quality and the Information Life Cycle



Outline

- Major Points of the Presentation
- Introduction to the 5S View of DLs
  - Informal Definition of DLs
  - Formal Definition of DLs
  - DL Ontology
- Defining a Quality Model for DLs
- Quality and the Information Life Cycle
- ***An XML Log Standard for DLs***
- Conclusions and Future Work

XML Log Standard for DLs: Pubs

1. Marcos André Gonçalves, Ganesh Panchanathan, Unnikrishnan Ravindranathan, Aaron Krowne, Edward A. Fox, Filip Jagodzinski, and Lillian Cassel. The XML Log Standard for Digital Libraries: Analysis, Evolution, and Deployment. Proc. JCDL'2003, Third Joint ACM / IEEE-CS Joint Conf. on Digital Libraries, May 27-31, 2003, Houston, 312 - 314
2. Marcos André Gonçalves, Ming Luo, Rao Shen, Mir Farooq Ali, and Edward A. Fox. An XML Log Standard and Tool for Digital Library Logging Analysis. In Proc. Research and Advanced Tech. for Digital Libraries, 6th European Conf., ECDL 2002, Rome, Sep. 16-18, 2002, eds. Maristella Agosti and Constantino Thanos, LNCS 2458, Springer, pp. 129-143.

XML Log Standard for DLs, Quality

DL Concept	Dimensions of Quality	Log can be used to measure?
Digital object	Accessibility	No
	Pertinence	Yes
	Preservability	No
	Relevance	Yes
	Similarity	No
	Significance	No
	Timeliness	No
Metadata specification	Accuracy	No
	Completeness	No
	Conformance	No
Collection	Completeness	No
	Impact Factor	No
Catalog	Completeness	No
	Consistency	No
Repository	Completeness	No
	Consistency	No
Services	Composability	No
	Efficiency	Yes
	Effectiveness	Yes
	Extensibility	No
	Reusability	No
	Reliability	Yes

Outline

- Major Points of the Presentation
- Introduction to the 5S View of DLs
  - Informal Definition of DLs
  - Formal Definition of DLs
  - DL Ontology
- Defining a Quality Model for DLs
- Quality and the Information Life Cycle
- An XML Log Standard for DLs
- ***Conclusions and Future Work***

Conclusions and Future Work

- "Study of User Quality Metrics for Metasearch Retrieval Ranking": new grant in 2004 IMLS NLG led by Martin Halbert of Emory University
- Development of more usage-oriented measures
  - Current measures are mostly system-oriented
- Development of Quality ToolKit (5SQual) for DL managers with following features:
  - Mapping tool to map local log format to standard XML Log format
  - Components to implement all measures
  - Visualization of data and measures
  - Broken into several logical pieces to be used in the different phases of the information life cycle

Questions/Discussion?