“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
<table>
<thead>
<tr>
<th>Digital</th>
<th>Information</th>
<th>Libraries</th>
<th>Retrieval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chair</td>
<td>Committee</td>
<td>Conference</td>
<td>Workshop</td>
</tr>
<tr>
<td>Education</td>
<td>Electronic</td>
<td>Dissertations</td>
<td>Theses</td>
</tr>
<tr>
<td>Research</td>
<td>Science</td>
<td>Systems</td>
<td>Technology</td>
</tr>
<tr>
<td>Open Initiative</td>
<td>Initiative</td>
<td>International</td>
<td>Multimedia</td>
</tr>
<tr>
<td>Access</td>
<td>Archive</td>
<td>Intelligent</td>
<td>Interactive</td>
</tr>
<tr>
<td>Knowledge</td>
<td>Learning</td>
<td>SIGIR</td>
<td>Tutorial</td>
</tr>
</tbody>
</table>
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
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
Quality and the Information Life Cycle

Creation

Utilization

Distribution

Seeking

Preservability

Accessibility

Significance

Accessing Filtering

Relevance

Searching

Browsing

Recommending

Active

Retrieval

Retaining

Archiving Networking

Semi-Active

Preservability

Semi-Inactive

Preservability

Inactive

Retrieval

Discard

Expiration
Digital Libraries
Shorten the Chain from

Author

Editor

Reviewer

Publisher

A&I

Consolidator

Library

Reader
DLs Shorten the Chain to

Roles

Author
Reader
Editor
Reviewer
Teacher
Learner
Librarian

User

Digital Library
Degree of Structure

Web  DLs  DBs

Chaotic  Organized  Structured
Locating Digital Libraries in Computing and Communications Technology Space

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

- digital object
  - atomic digital object
  - complex digital object
  - is of type
  - is composed of
A digital library = repository of collections and metadata + services
The World According to the Open Archives Initiative

Service Providers
- Discovery
- Current Awareness
- Preservation

Data Providers

Metadata harvesting
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

Core NSDL “Bus”

Collection Building

NSDL Collections

Portals & Clients

Usage Enhancement

Core Services:
- information retrieval
- browsing
- authentication
- personalization
- discussion
- annotation

Core Services:
- metadata gathering
- protocols
- harvesting

Other NSDL Services

Special Databases

NSDL Collections

Core Services:
The Ensemble Computing Portal

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

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

http://www.computingportal.org/
Ensemble: PDP-8 Overview

**Principles of Distributed Portals**

**Content**
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
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
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
<table>
<thead>
<tr>
<th>IM1. Information models and systems*</th>
<th>IM8. Distributed DBs</th>
</tr>
</thead>
<tbody>
<tr>
<td>IM2. Database systems*</td>
<td>IM9. Physical DB design</td>
</tr>
<tr>
<td>IM3. Data modeling*</td>
<td>IM10. Data mining</td>
</tr>
<tr>
<td>IM4. Relational DBs</td>
<td>IM11. Information storage and retrieval</td>
</tr>
<tr>
<td>IM5. Database query languages</td>
<td>IM12. Hypertext and hypermedia</td>
</tr>
<tr>
<td>IM6. Relational DB design</td>
<td>IM13. Multimedia information &amp; systems</td>
</tr>
</tbody>
</table>

* Core components
DL Curriculum Framework

Semester 1: DL collections: development/creation

Semester 2: DL services and sustainability

**COURSE STRUCTURE**

**Core DL Topics**

- Digitization
- Storage
- Interchange
- Metadata
- Cataloging
- Author submission
- Architectures (agents, buses, wrappers/mediators) Interoperability
- Naming
- Repositories
- Archives
- 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
- Intellectual property rights mgmt.
- Privacy Protection (watermarking)

- Bibliographic information
- Bibliometrics
- Citations
- Content-based analysis
- Multimedia indexing
- Multimedia presentation, rendering
- Search & search strategy
- Info seeking behavior
- User modeling
- Feedback
- Info summarization
- Visualization

**Core DL Topics**

- Digitization
- Storage
- Interchange
- Metadata
- Cataloging
- Author submission
- Architectures (agents, buses, wrappers/mediators) Interoperability
- Naming
- Repositories
- Archives
- 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
- Intellectual property rights mgmt.
- Privacy Protection (watermarking)

- 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:
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
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.
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

ETANA-DL UNION CATALOG

- Search
- Browse
- Recommend
- Note
- Personalize
- Review
- Visualizations
- Archaeology Specific

DATABASE WRAPPERS

- Lahav
- Nimrin
- Umayri
- Hisban
- Megiddo
- Jalul
- New Sites

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

Streams
- Structured Stream
  - SpatTemOrg
  - StraDia

Structures
- Descriptive Metadata specification

Spaces

Scenarios
- indexing
- browsing
- searching

Societies
- services

ArchDL
- ArchDColl
- ArchDO
- ArchObj
- ArchColl
- ArchMetadata catalog
- Minimal ArchDL
From [Murthy et al. 2010], a Complex object is a tuple $cdo=(h, SCDO=DO \cup SM, S)$, where:

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

**Diagram:**
- Collection
  - Complex Object
    - Image Complex Object
      - composed by
        - Image Digital Object
          - Digital Object
            - Structured Stream
        - Feature Vector
        - Similarity Scores

5S Extensions to CBIR
Uma Murthy’s Dissertation, esp. Ch. 8

SI-DL metamodel - v1
• 5S analysis of SI
• Initial metamodel

SI-DL metamodel – v2
• Improved metamodel
• Case study

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

SuperIDR-v1
• Formative evaluation
• SuperIDR improvements
• Classroom study design

SuperIDR-v2
• Lucene indexing
• Usability improvements

SuperIDR-v3
• Image mgmt.
• Combined search
• Comparison
• Usability improvements

SI-DL design guidelines
Definitional dependencies among concepts in the SuperIDR SI-DL
Streams

- Text
- Audio
- Image
- Video
- Digital Object
- Collection
- Catalog
- Repository

Structures

- Metadata Specifications
- Index
- Topological Measure
- Metric
- Vector
- Probabilistic Measure
- Measurable

Spaces

- Service
- Operation
- Actor
- Association
- Event
- recipient

Content / People

- Societies
- Scenarios
- Contains
- Precedes
- Happens_before
- Reuses
- Extends
- Uses
- Participates_in
- Runs
-executes
-redefines
-Invokes
-inherits_from/includes

Scenario

- Extends
- Reuses
- Contains
- Happens before
- Event
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
<table>
<thead>
<tr>
<th>Infrastructure Services</th>
<th>Repository-Building</th>
<th>Add Value</th>
<th>Information Satisfaction Services</th>
</tr>
</thead>
<tbody>
<tr>
<td>Creational</td>
<td>Preservational</td>
<td>Annotating</td>
<td>Browsing</td>
</tr>
<tr>
<td>Acquiring</td>
<td>Conserving</td>
<td>Classifying</td>
<td>Collaborating</td>
</tr>
<tr>
<td>Cataloging</td>
<td>Converting</td>
<td>Clustering</td>
<td>Customizing</td>
</tr>
<tr>
<td>Crawling (focused)</td>
<td>Copying/Replicating</td>
<td>Evaluating</td>
<td>Renewing</td>
</tr>
<tr>
<td>Describing</td>
<td>Emulating</td>
<td>Extracting</td>
<td>Translating (format)</td>
</tr>
<tr>
<td>Digitizing</td>
<td>Renewing</td>
<td>Indexing</td>
<td>Providing access</td>
</tr>
<tr>
<td>Federating</td>
<td>Translating (format)</td>
<td>Measuring</td>
<td>Recommending</td>
</tr>
<tr>
<td>Harvesting</td>
<td></td>
<td>Publicizing</td>
<td>Requesting</td>
</tr>
<tr>
<td>Purchasing</td>
<td></td>
<td>Rating</td>
<td>Searching</td>
</tr>
<tr>
<td>Submitting</td>
<td></td>
<td>Reviewing (peer)</td>
<td>Visualizing</td>
</tr>
</tbody>
</table>

Add Value: Annotating, Classifying, Clustering, Evaluating, Extracting, Indexing, Measuring, Publicizing, Rating, Reviewing (peer), Surveying, Translating (language)
Tools/Applications

DL Expert

5S Meta Model

5S Graph

5SL DL Model

5SLGen

Tailored DL

Logging Module

XML Log

ODLSearch, ODLBrowse, ODLRate, ODLReview, ........

 DLCDesigner

Practitioner

Researcher

Teacher

Tailored DL
Architecture of a Union DL (ETANA.org)
Union Catalog Integration

Virtual Nimrin (VN)
- VN Metadata Format
- VN Catalog

Halif DigMaster (HD)
- HD Catalog
- HD Metadata Format

Mapping Tool
- Wrapper
- Union Catalog
- Global Metadata Format

Union ArchDL
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