浙江大学计算机学院
数字媒体与网络技术

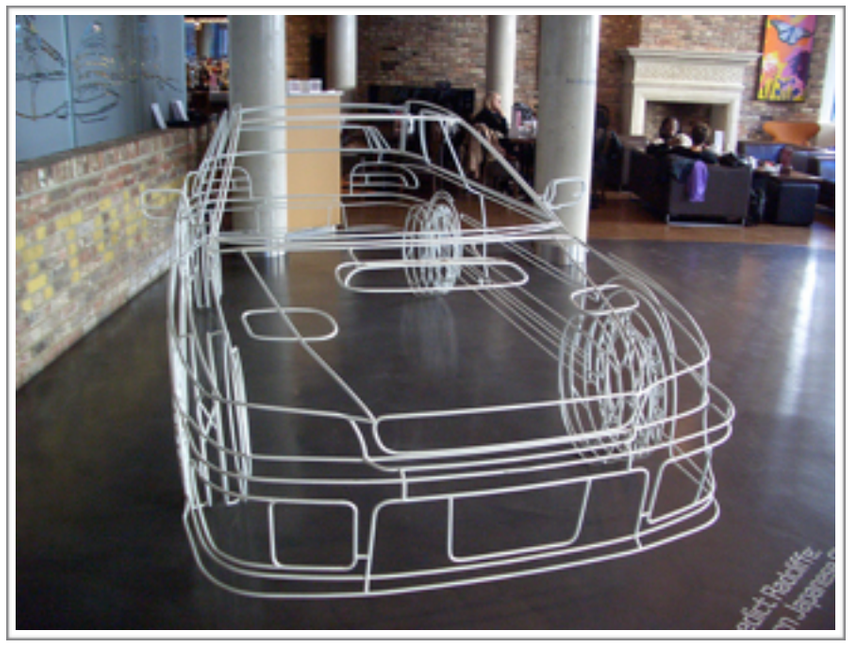
Digital Asset Management
数字媒体资源管理

8. Introduction to Digital Library

任课老师：张宏鑫
2014-11-11

Outline

- Basic Principles
- MetaData
- Intellectual Property
- Architecture of DL
- New Digital Library Architecture



Basic Principles

What is a Library?

- A place in which literary, musical, artistic, or reference materials are kept for use but not for sale

- A collection of such materials
  - books, manuscripts, recordings, or films

http://www.m-w.com/cgi-bin/dictionary?book=Dictionary&va=library&x=0&y=0

What is a Digital Library (DL)?

"...a managed collection of information, with associated services, where the information is stored in digital formats and accessible over a network" (Arms, p. 2)

What is a Digital Library (DL)?

"...a managed collection of information, with associated services, where the information is stored in digital formats and accessible over a network" (Arms, p. 2)

What is a Digital Library

- Library ++ (library+archive+museum+…)
- Distributed information system + organization + effective interface
- User community + collection + services
- Digital objects, repositories, IPR management, handles, indexes, federated search, hyper-base, annotation

http://www.cadal.zju.edu.cn/

中美百万册数字图书馆

- China-America Digital Academic Library
- 数字化信息传播快、访问便捷、易于检索、支持动态媒体数据

纸张 v.s. 字节

	纸张存储	电子存储
价格	10～15美分／页	<0.1美分／页
空间	1,000,000纸张	250G (400dpi)
复印	170小时	23小时
复印浙大图书馆 502万册图书	40年，4亿元	5年，4千万元

中美百万册数字图书馆

- 珍贵濒危文物数字化信息库
  - 信息检索查询与共享浏览
  - 协同讨论与合作交流
  - 访问控制与安全管理
- 敦煌石窟文物的数字化信息共享与交流

中美百万册数字图书馆

- 挑战：
  - 日点击超过 > 120,000次
  - 日数据下载量 > 40GB
  - 日新产生数据 > 10GB
  - 图书数据存储量 > 200TB

- 采用IBM System Storage DS8100

中美百万册数字图书馆

- 关键技术:
  - 数字化文物多媒体信息模型
  - 多媒体信息检索方法
  - 多媒体信息安全与版权控制

中美百万册数字图书馆

- 视频内容结构化与摘要生成
- 视频内容结构化
- 视频摘要生成
- 视频内容结构化和摘要的网络浏览

中美百万册
数字图书馆

- 多媒体信息检索
  - 图像分析系统
    - 视频分析系统
    - 音频分析系统
    - 高层语义的提取
- 功能
  - 基于文本的查询
  - 基于内容的查询
  - 智能化数据导航/浏览

Digital Libraries 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)

5S Layers

Societies

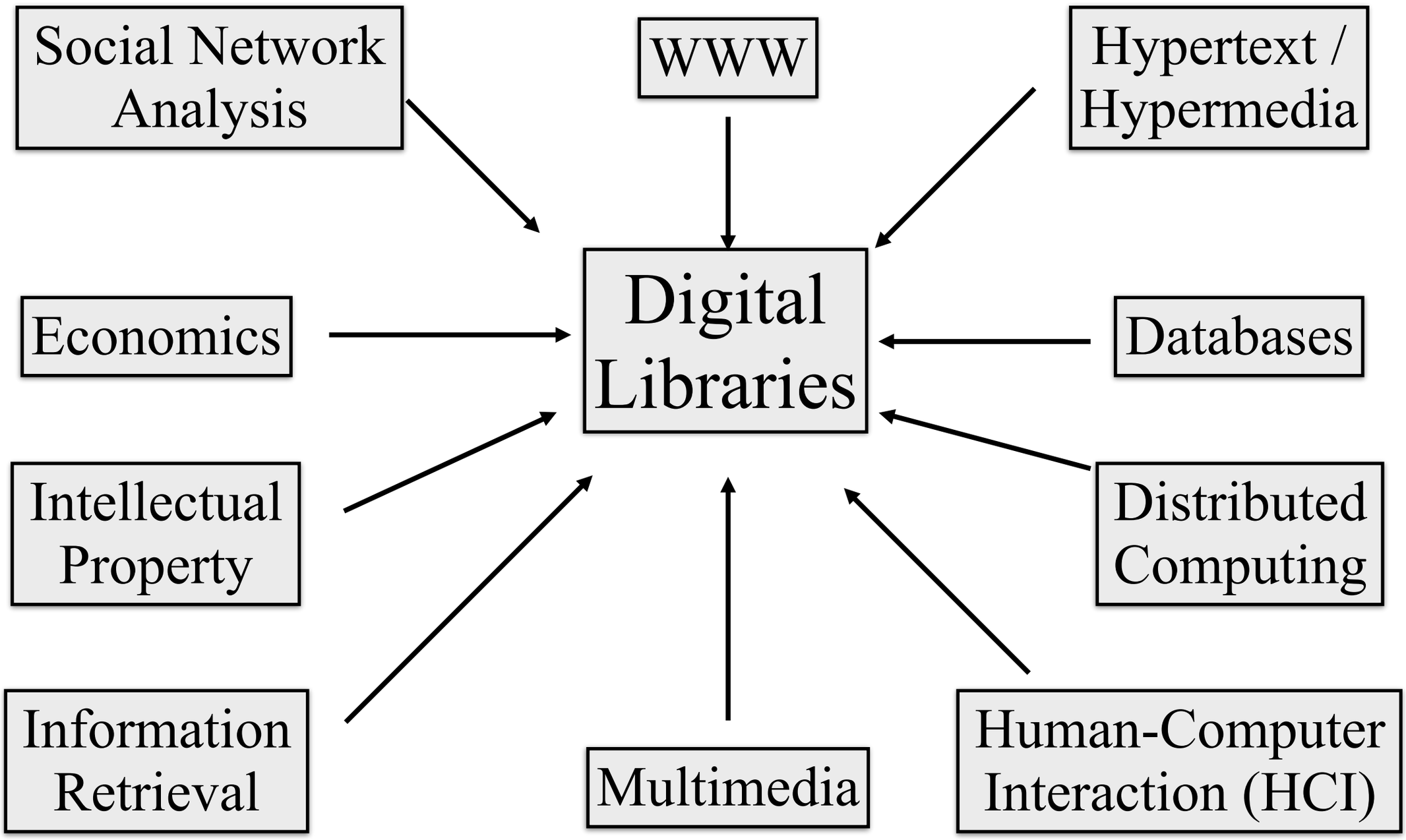
Scenarios

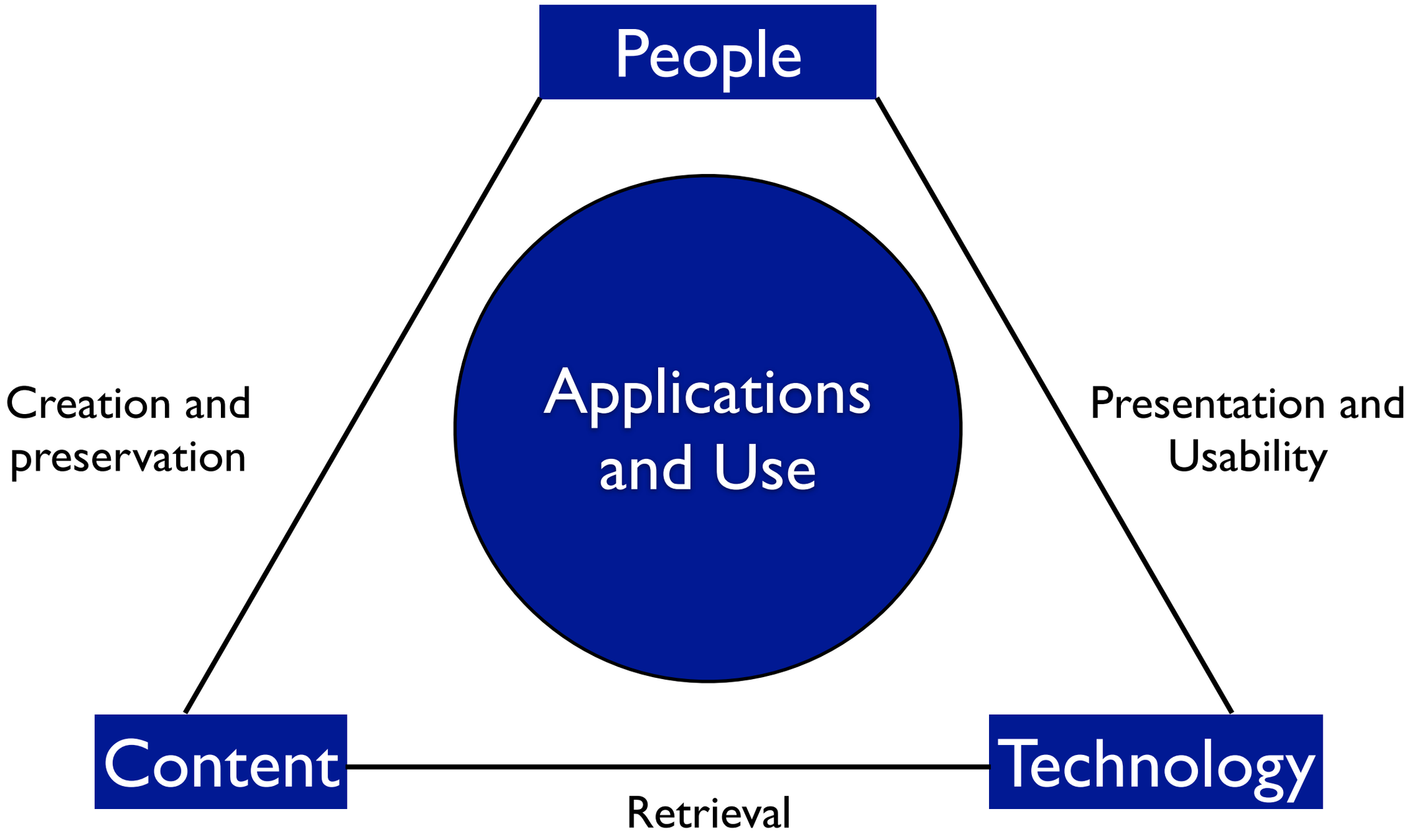
Spaces

Structures

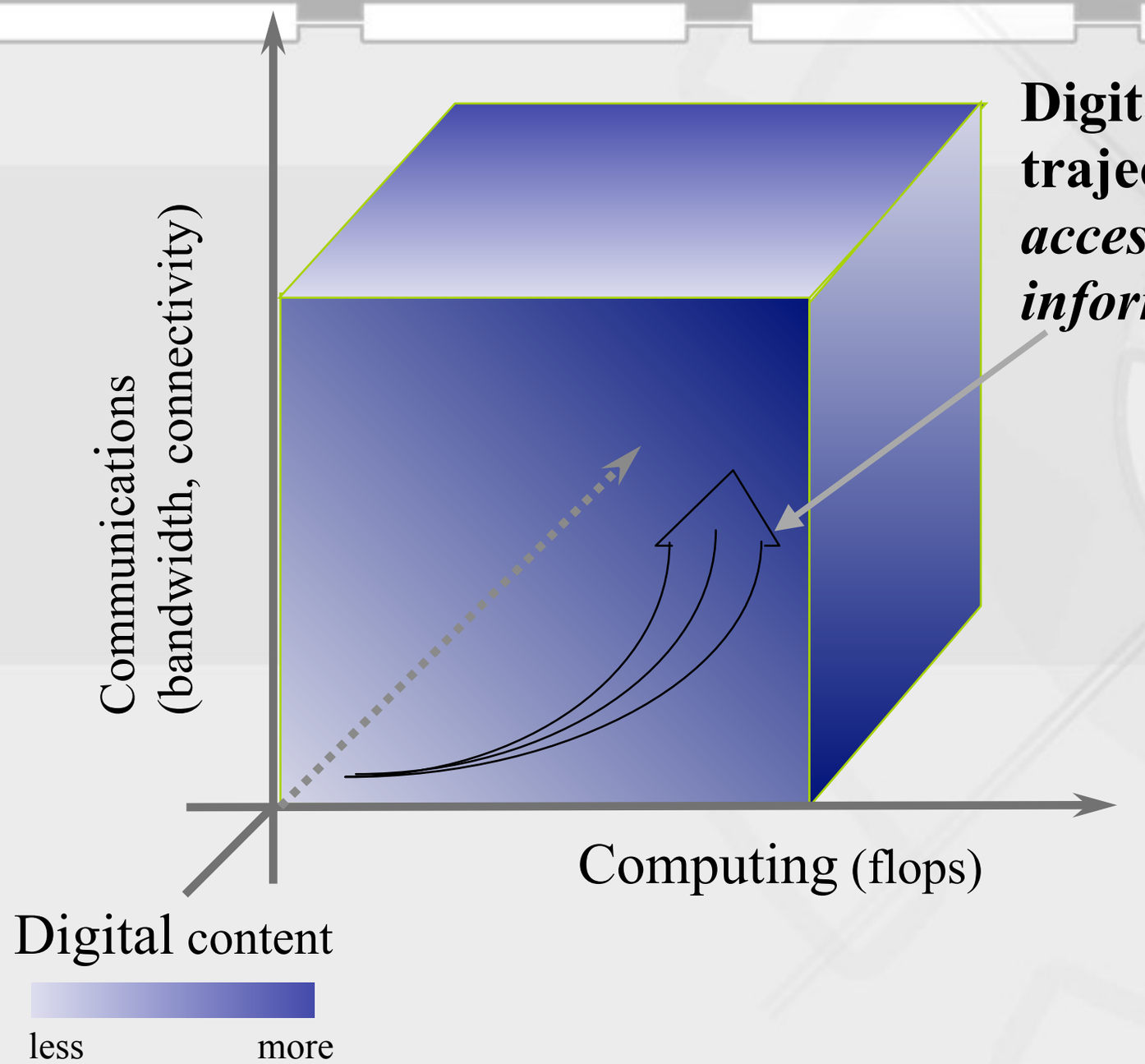
Streams

Research of Digital Library




People
Applications and Use
Creation and preservation
Presentation and Usability
Content
Technology
Retrieval

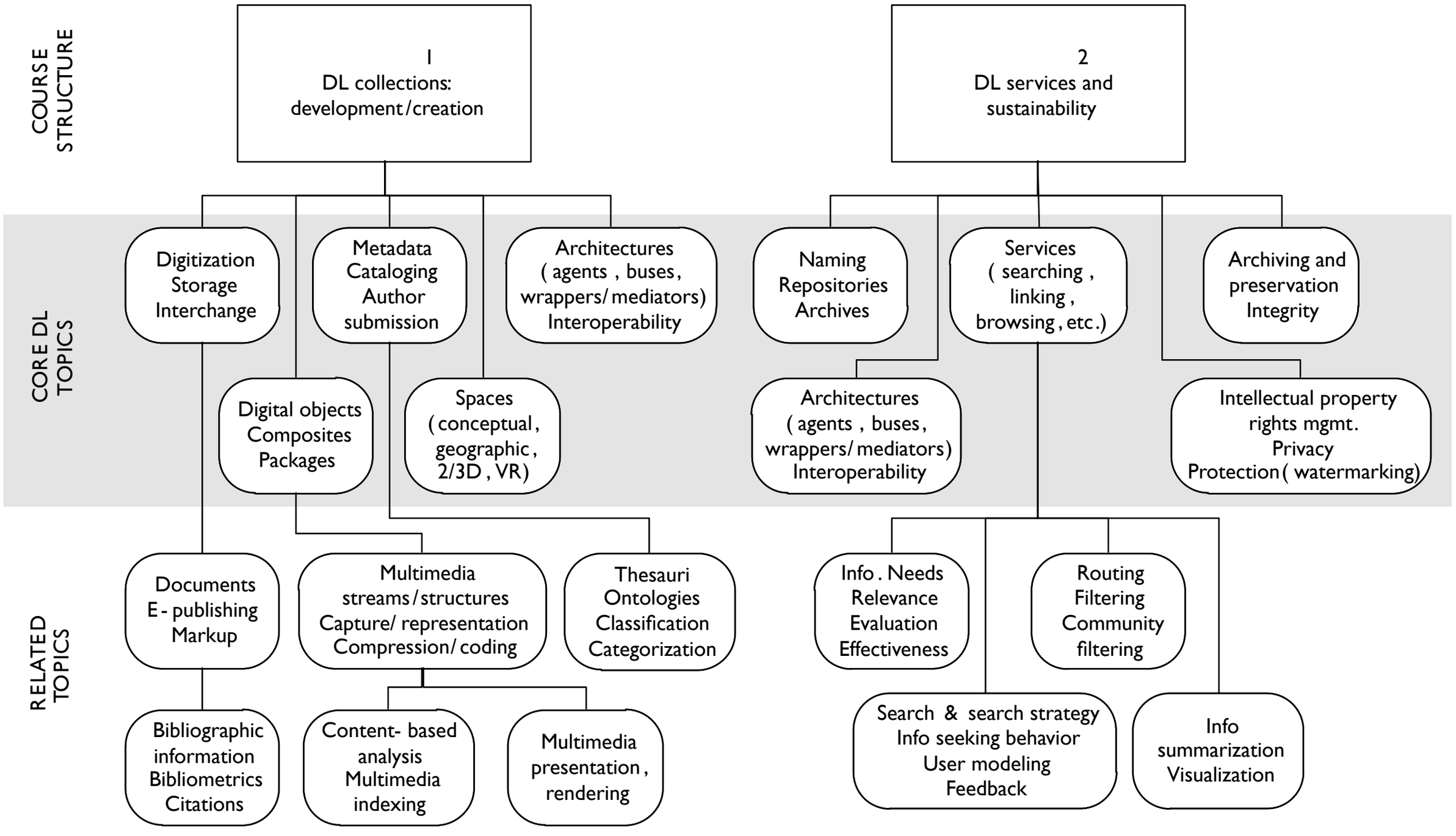
Locating Digital Libraries in Computing and Communications Technology Space



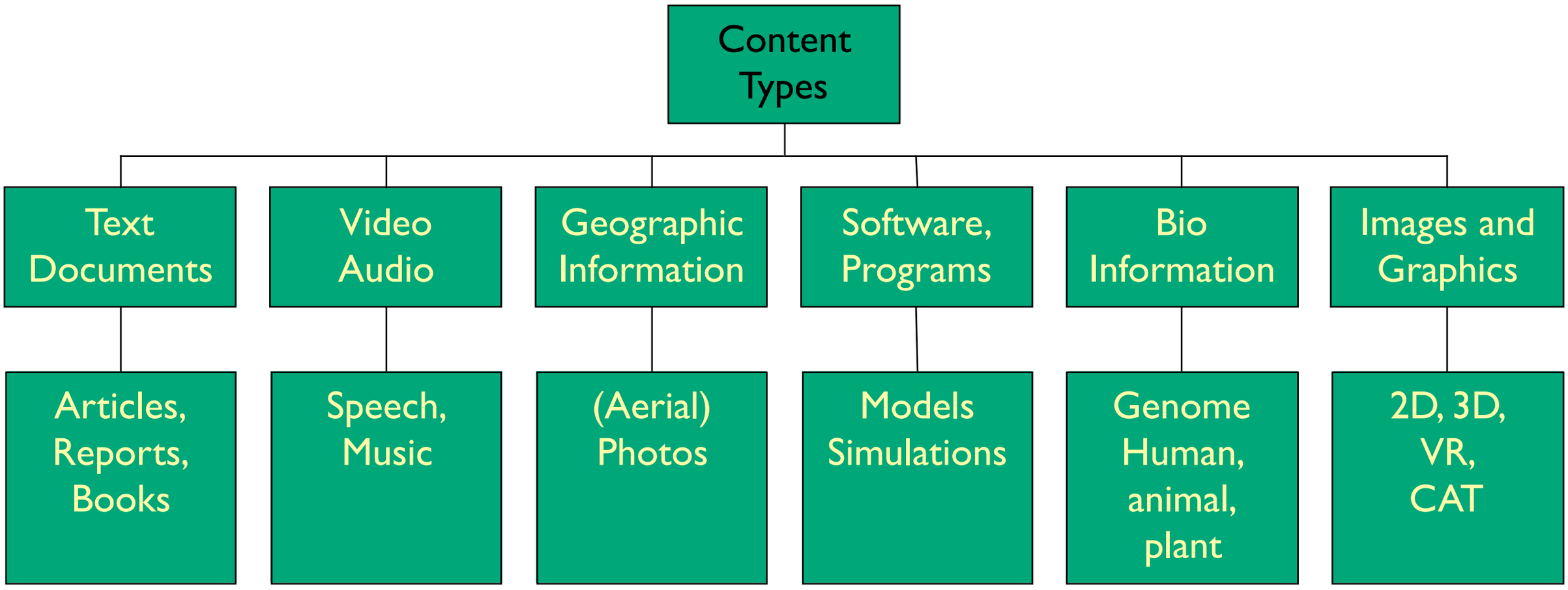
Digital Libraries technology trajectory: ***intellectual access to globally distributed information***

(from S. Griffin, NSF)

DL Framework



Digital Library Content



Core of Digital Library

- Collecting
  - Authoring, Repositories, Archives, Museums, …
- Organizing
  - Packaging of Data and Metadata, Storing
  - Naming/Identifying and Cataloging
  - Classification, Clustering, …
- Serving
  - Indexing, Linking, Summarizing, Visualizing
  - Browsing, Accessing, Searching, Filtering, Retrieving, Distributing, Using, …

Digital Libraries Shorten the Chain from

Author

Editor

Reviewer

Publisher

A&I

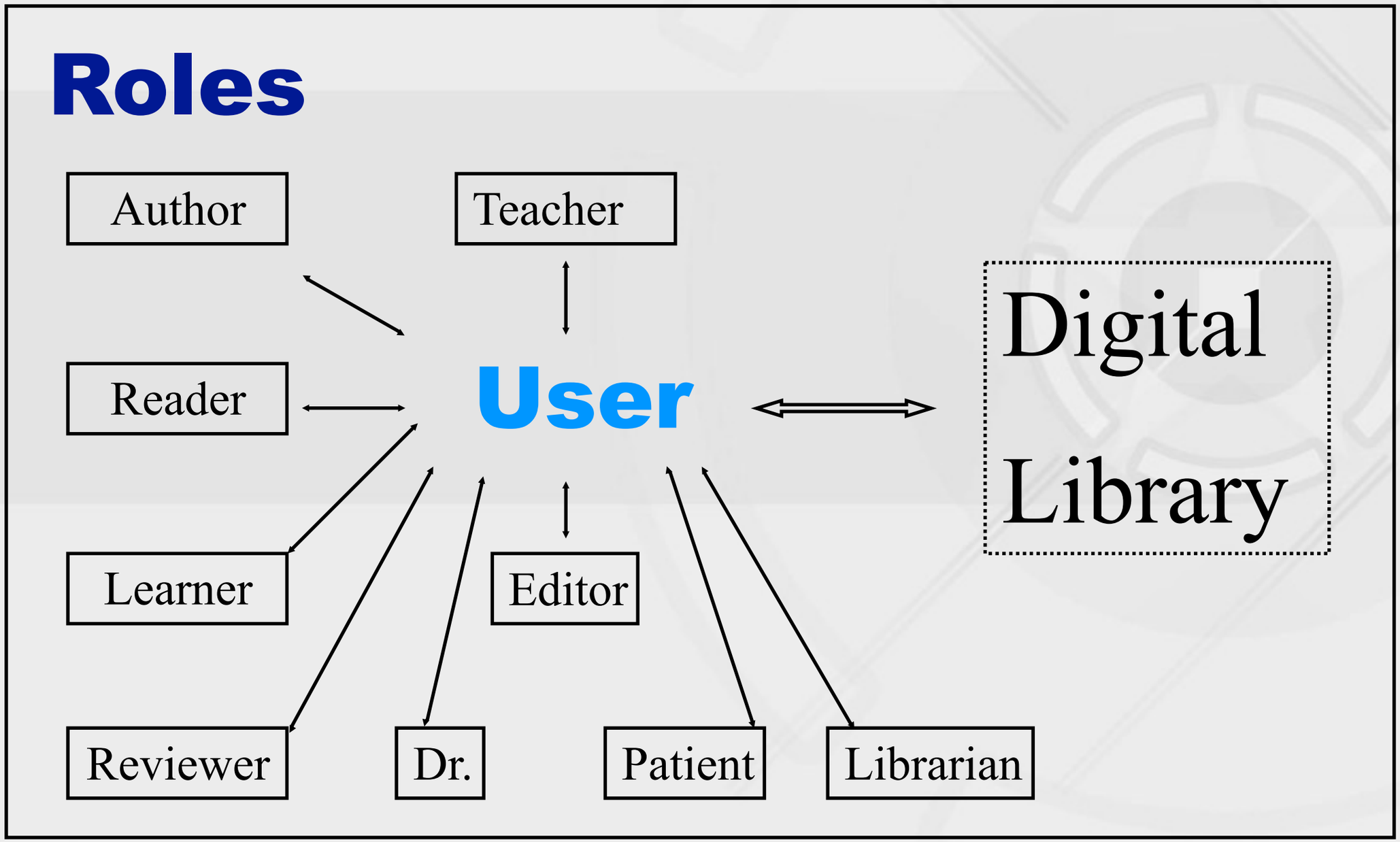
Consolidator

Library

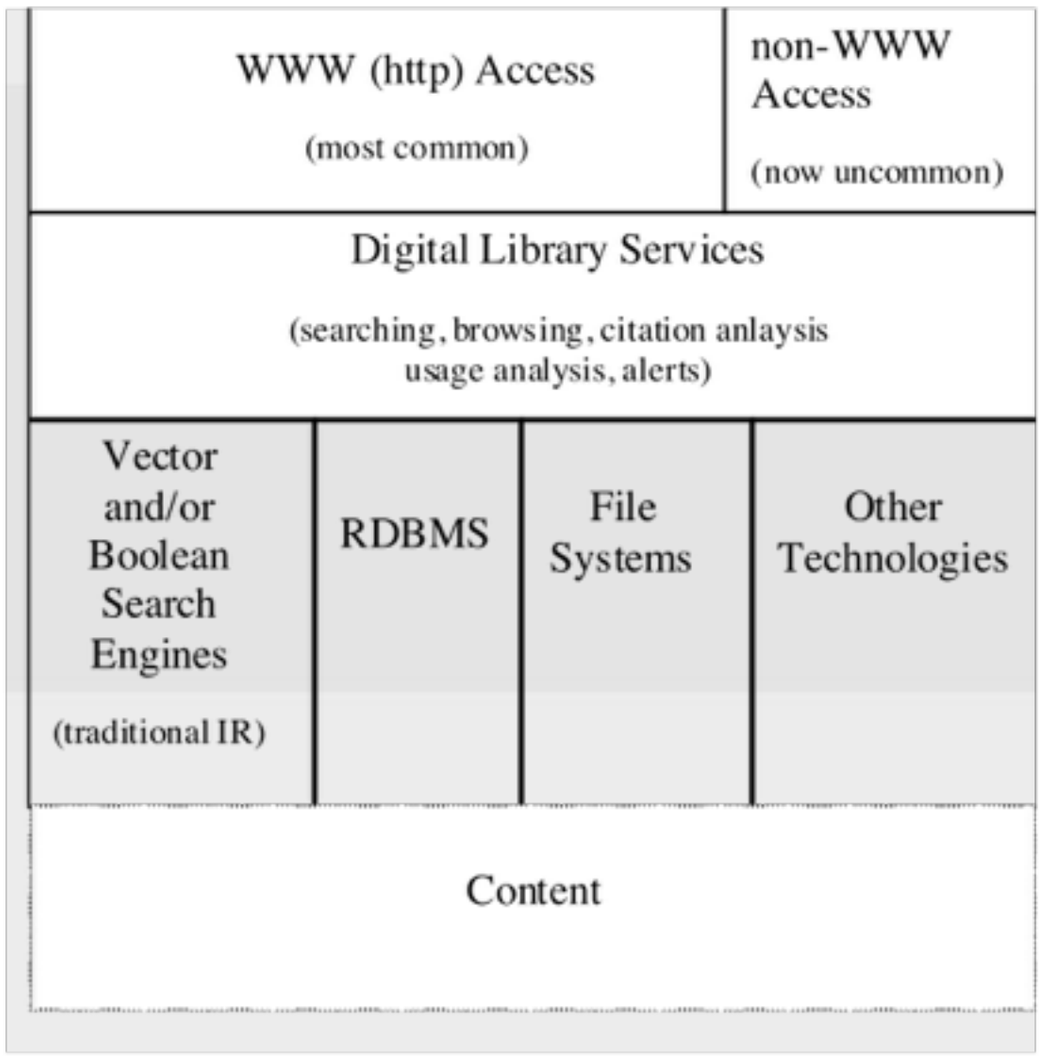
Reader

浙江大学计算机学院
数字媒体与网络技术

DL = Users Direct
(Organized Artifact Mediated Communication)



DL = Content + Services



- "Why not just use the WWW" ?
  - WWW by itself has low archival & management characteristics
- "Why not use a RDBMS?"
  - In the same way that a card catalog is not a TL, a RDBMS is candidate technology for use in DLs
- DL is the union of the content and services defined on the content

DL Components

Gateways

User Interfaces

MM/ HT Renderer

Workflow Mgr

Search Engines, Classifiers, …

DBMS

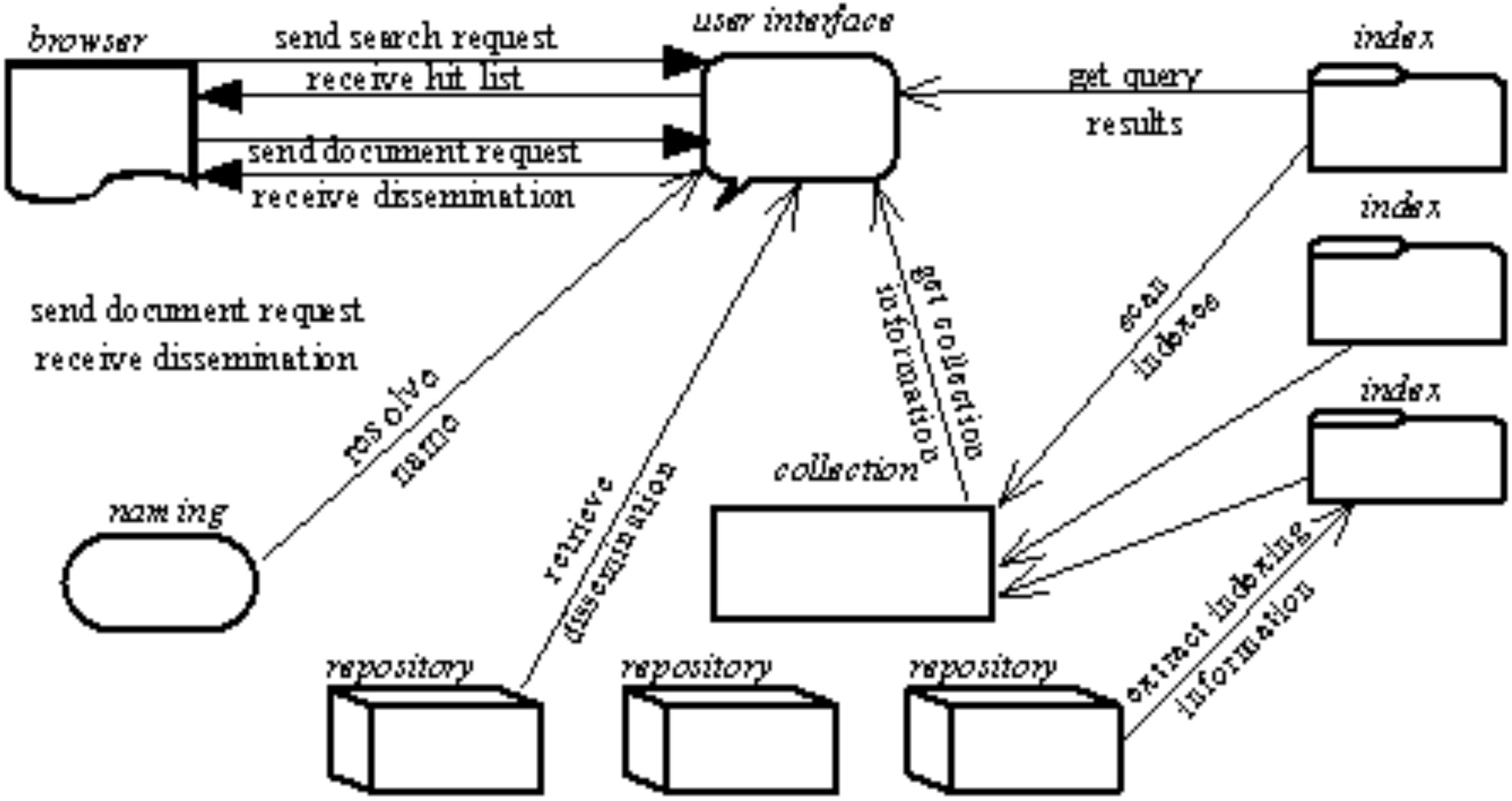
Rights Management

Data, MM Info

Repository

Core DL Components

Value-added service



Digital archives

- Archives differ from libraries
  - Containing primary sources of information
    - typically letters and papers directly produced by an individual or organization
    - rather than the secondary sources found in a library (books, etc);
  - Having their contents organized in groups rather than individual items.
  - Having unique contents

浙江大学计算机学院
数字媒体与网络技术

MetaData

Metadata vs. Data

- **Data** refers to digital objects or digital representations of objects
- **Metadata** is information about the objects (e.g. title, author, etc.): descriptive, interpretive, administrative, …
- Many digital library efforts, including the Open Archives Initiative (OAI), focus on metadata, with the implicit understanding that metadata usually contains useful links to the source digital objects
- Purists would argue metadata is just data

Metadata

- “data about data” is about as good as the definition gets...
- As a DL grows, metadata becomes more important
- Lack of metadata has different consequences
  - documentation: metadata can be regenerated automatically, or by hand
  - datasets, pictures: once lost, can be impossible to regenerate
    - LaRC windtunnel example

Types of Metadata

- Descriptive See http://www.loc.gov/standards/metadata.html
  - Discovery / description of objects
    - Title, author, abstract, etc.
- Structural
  - Storage & presentation of objects
    - 1 pdf file, 1 ppt file, 1 LaTeX file, etc.
- Administrative
  - Managing and preservation of objects
    - Access control lists, terms and conditions, format descriptions, “meta-metadata”

Digital Objects

- Digital Objects = digital information + metadata + handle
  - XML markup, digitized radio programs, computer programs…
- Rules and conventions for each category of digital objects
  - Grouping several digital objects for complex works
    - Reports = several chapters
  - Representing relation between works
    - Versions, translation, …
  - Naming
  - Need a user interface aware of rules and conventions…

Metadata Formats

- MARC is very rich
  - good candidate for an “archival” metadata format, from which simpler formats can be derived
- Dublin Core designed to be simple enough for the average author to generate by hand
  - only 15 core fields defined
- Other formats defined for specific purposes:
  - BibTeX: TeX/LaTeX publishing
  - refer: troff/nroff
  - RFC-1807: email exchange

Interesting Formats

- Library science
  - Machine Readable Catalogue (MARC): huge, extensive, all purpose, one size fits all format
    - pro: does everything
    - con: kids, don't try this at home!
- Computer science
  - application-specific formats: refer, BibTeX, RFC-1807, etc.
- DC - common ground?

Background and Primary Goal of DC

- Came out of a 1995 joint OCLC/NCSA workshop in Dublin, Ohio
- An attempt to improve resource discovery on the Web
  - resource discovery/description/evaluation
- Now adopted by many resource description communities
  - Museums, libraries, government agencies, and commercial organizations
- Building an interdisciplinary consensus about a core element set for resource discovery
  - Simplicity of creation and maintenance
  - Semantic Interoperability
  - International Consensus
  - Flexible extensibility
  - Metadata modularity on the Web

Dublin Core Element Sets

- Title
- Creator
- Subject
- Description
- Publisher
- Contributor
- Date
- Type

- Format
- Identifier
- Source
- Language
- Relation
- Coverage
- Rights

- 15 elements of descriptive metadata
- All elements are optional and repeatable
- Dublin Core is extensible
  - Offering a starting point for semantically richer descriptions

Dublin Core and RDF/XML

- Dublin Core is about semantics
  - What we are trying to say about resources
- RDF is about structure
  - Conventions for encoding the assertions about a resource that uses DC semantics
- XML
  - Syntax for encoding assertions in RDF
  - RDF-encoded DC metadata

Leader: : 01663ngm 22002771 4500:
005: : 19950927090218. 0 :
007: :vducgaiuu:
008: : 950927s1993 mau--- d vlfre d :

Ctrl Numb 001 200312310
Cntl Iden 003 OBgNWOET
ISBN Numb 020 -- a 0300056958
Catl Orig 040 __ a OBgNWOET
Tran c OBgNWOET
Lang Summ 041 -- b fre
Titl Main 245 00 a A la recontre de Philippe
GMD h [videorecording] /
Resp c Massachusetts Institute of Technology ; written by Gilberte Furstenburg ; directed by Janet H. Murray ; software programmed by Stuart
A. Malone.
Pubn City 260 __ a Cambridge, MA :
Publ b dist. by Annenberg/CPB.,
Date c 1993.
Desc Extn 300 __ a 1 laserdisc (CAV) :
Othr b sd., col. :
Dimn c 12 in. +
Accm e Teacher guide + 3 computer disks.
Note Genl 500 __ a Issued as videodisc.
Note Genl 500 __ a Title from cover.
Note Summ 520 __ a Provides an engaging way to sharpen comprehension skills. Students navigate through Paris neighborhoods and shops,dealing with friends, tradespeople, telephones and answering machines with the goal of finding an apartment for the hapless Philippe. Includes many helpful tools,such as self-testing exercises and an electronic glossary, visual and audio resources, including maps, telephones and newspapers which help students function within the story. Teachers can customize the program according to their students levels and abilities.
Note Targ 521 2_ a Senior high and college.
Note Targ 521 2_ a 09-adult.
Note Tech 538 -- a Macintosh computer ; system 6.0 or later ; 2 MB of RAM ; 3.5 MB of hard disk space ; videodisc player ; video monitor.
Subj Topc 650 _0 a Languages, Modern.
Subj Topc 650 _0 a Language and languages.
Subj Topc 658 _7 a Foreign languages, French.
Srce 2 nwoet
Locn Coll 852 1_ a OBgNWOET
SubA b Northwest Ohio Media Center
Clas h 200312310
BarC p 200312310

MARC

from:
http://m27-5.bgsu.edu/nwoetf/marc/phillippe.html

Dublin Core, pre-XML

<META NAME="DC.title" CONTENT="Metadata: Enabling the Internet">
<META NAME="DC.subject" CONTENT="(SCHEME=keyword) Metadata, Dublin Core, PICS, Resource Discovery">
<META NAME="DC.author" CONTENT="(TYPE=name) Renato Iannella">
<META NAME="DC.author" CONTENT="(TYPE=email) renato@dstc.edu.au">
<META NAME="DC.author" CONTENT="(TYPE=affiliation) DSTC Pty Ltd">
<META NAME="DC.author" CONTENT="(TYPE=name) Andrew Waugh">
<META NAME="DC.author" CONTENT="(TYPE=email) a.waugh@cmis.csiro.au">
<META NAME="DC.author" CONTENT="(TYPE=affiliation) CSIRO">
<META NAME="DC.publisher" CONTENT="(TYPE=name) DSTC Pty Ltd">
<META NAME="DC.date" CONTENT="(TYPE=creation) (SCHEME=ISO31) 1997-01-20">
<META NAME="DC.date" CONTENT="(TYPE=current) (SCHEME=ISO31) 1997-01-20">
<META NAME="DC.form" CONTENT="(SCHEME=imt) text/html">
<META NAME="DC.identifier" CONTENT="(TYPE=url) <http://www.dstc.edu.au/RDU/reports/CAUSE97/>">
<META NAME="DC.language" CONTENT="(SCHEME=iso639) en">

see Internet RFC-2731

from:
http://www.dstc.edu.au/RDU/reports/CAUSE97/

Dublin Core, XML-encoded

```
<?xml version="1.0"?>
      <!DOCTYPE rdf:RDF SYSTEM "http://purl.org/dc/schemas/dcmes-xml-20000714.dtd">
      <rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
            xmlns:dc ="http://purl.org/dc/elements/1.1/">
       <rdf:Description about="http://foo.edu/dl/report-1">
        <dc:title>Perpetual Motion Machine</dc:title>
        <dc:description>This report redefines physics.</dc:description>
        <dc:date>1998-10-10</dc:date>
        <dc:format>text/html</dc:format>
        <dc:language>en</dc:language>
        <dc:contributor>Kant, B. Reproduced</dc:contributor>
       </rdf:Description>
      </rdf:RDF>
```

example adapted from: http://www.purl.org/dc/documents/wd/dcmes-xml-20000714.htm

浙江大学计算机学院
数字媒体与网络技术

Intellectual Property

Intellectual Property (IP)

"Issues related to intellectual property law are the most serious problems facing digital libraries." - Lesk

Forms of IP Protection

- Applicable to DLs:
  - copyright
  - patent
- Less applicable to DLs:
  - trade secrets
    - you must sign a non-disclosure agreement before you can see the information
  - trademarks
    - the Internet has caused some issues to be revisited: *Sun Oil, Sun Microsystems, Sun Records: who gets www.sun.com?*

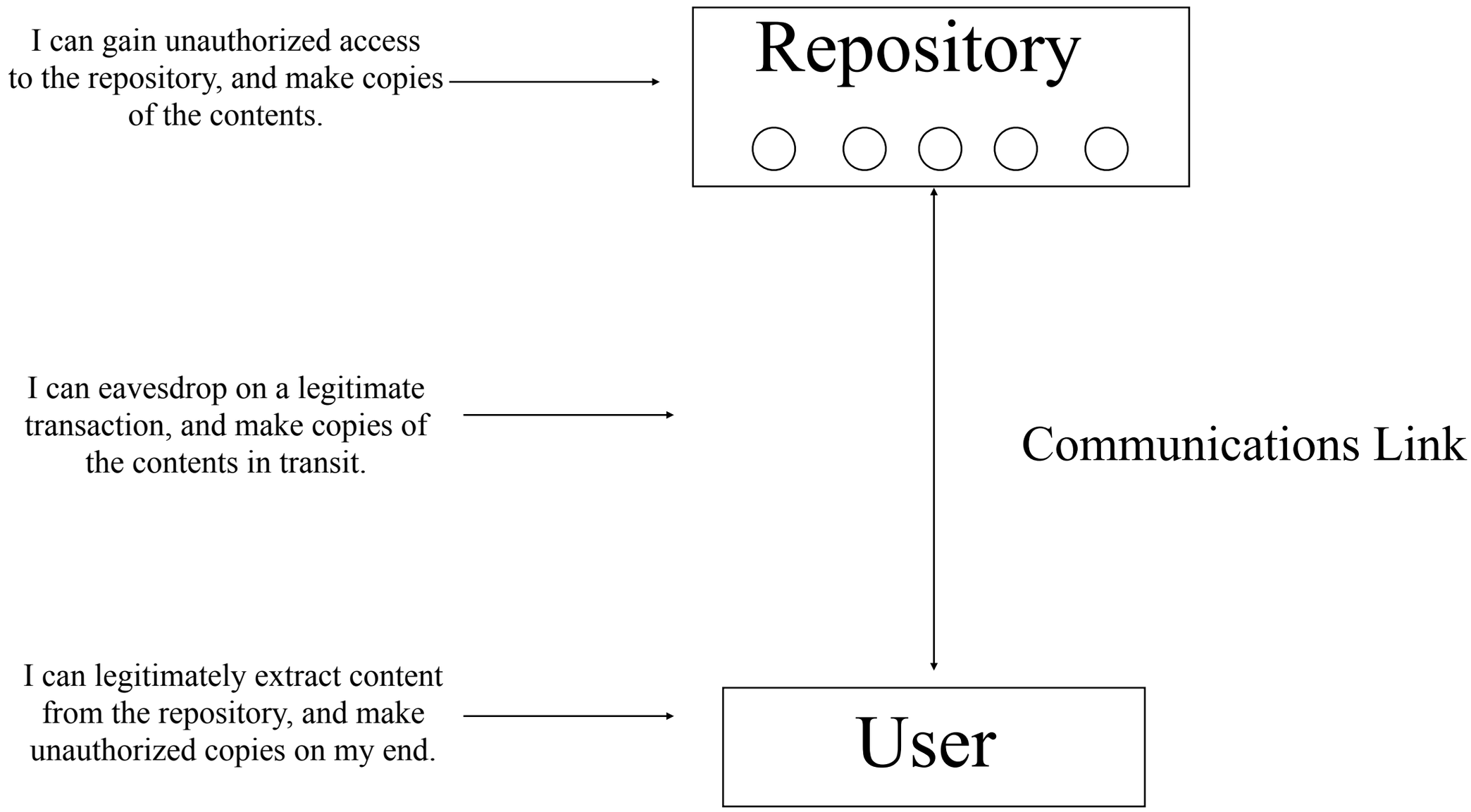
Software: Copyright and Patents

- Algorithms, methods, etc. can be patented
- Form of expression, representation, source code, etc. can be copyrighted
- Submarine / "Time bomb" patent
  - a popular program / method could be independently created, can be well established, then you discover you have violated a patent!
    - under old law, patents were 17 years after *issue* date, not *filing* date...
  - Pardo (spreadsheet) patent, LZW (GIF) patent
  - League of Programming Freedom (lpf.ai.mit.edu) - fights s/w patents

IP and Security in a DL Setting

- Typical things we would want to do in a DL:
  - authenticate the user
  - authenticate the DL
  - authenticate the contents
  - secure the contents
    - storage and transmission
  - negotiate terms and conditions of use

Places to Steal Information



Methods of Protection

Adapted from Lesk text book and Kohl, Lostpeich, Kaplan, 1997

- Fractional access
  - information is doled out in small quantities (presumably a small percentage of the total available)
  - its “hard” to capture enough to information to be useful
- Control of interface
  - access is only through proprietary interface
  - WWW is phasing out this approach

Methods of Protection

- Hardware locks / special hardware
  - the information (easy to copy) depends on a separate h/w device that is not easily copyable;
  - or, the output device has a unique key for decryption (audio board, video board, etc.)
- Repositories
  - keep a list of copyrighted stuff
  - generate checksums
  - to see if object X has been copied, compare checksums

Methods of Protection

- Steganography ("watermarks")
  - most applicable in images, it is the digital equivalent to what we currently do with currency, drivers licenses, etc.
  - premise: add enough extra information to uniquely identify the object
    - do not damage consumption of object
    - must not be removed by compression algorithms
  - for text, small adjustments in character spacing is made (fig. 10.4 in Lesk)
    - does not appear to be ASCII or SGML friendly... (?)

Methods of Protection

- Economic incentives
  - –make it not worth the time or the effort to steal
    - provide extra copies very cheap
    - site licenses
  - –recover costs in other ways
    - advertising
    - request donations

Securing the Content and Session

- Alternatives more commonly seen in a WWW DL environment are:
  - securing the session
    - to protect against eavesdroppers
    - can also be used to authenticate the repository and the user
  - securing the content
    - to protect against copying the object
    - can also be used to authenticate the object
  - Can be used together
    - but first, lets briefly review cryptography...

Securing the Session

- Two common WWW methods of secure communication:
  - Secure Socket Layer (SSL)
    - Freier, Karlton, Kocher, 1996
      - http://wp.netscape.com/eng/ssl3/ssl-toc.html
    - encrypts all network traffic, regardless of protocol (http, smtp, ftp, etc.)
  - Secure HTTP (SHTTP)
    - Rescorla & Schiffman, 1999
      - http://www.ietf.org/rfc/rfc2660.txt
    - a superset of http, protects only http traffic

Session Security Drawbacks

- From a providers point of view, you have protected only against eavesdroppers
  - –if your repository is compromised, your information can be copied
  - –nothing prevents legitimate clients from making local copies too...
- SSL & SHTTP are not rich enough to express terms and conditions

Securing the Content

- If the content were secured, then it would be safe from repository attacks, eavesdroppers, and local copying...
  - securing the session is not precluded; SSL/SHTTP would still be available for the very paranoid
- Moving security higher up the OSI model allows for richer expressions, such as terms and conditions

Superdistribution

- If the content is secure, then let copies be made!
  - the content can only be accessed by obtaining the right key, presumably through some authentication/payment mechanism
  - demo software with significant functionality "turned off" that can be turned on by purchasing a key is an example of *superdistribution*

Content Security Advantages

- Client verifies document authenticity
  - –not “librarian”
- Customer authenticates only to purchasing entity, not to publisher or librarian
- Customer drives the messaging / signing process
- Checksums and signatures verify integrity of the contents

From: Kohl, Lotspiech & Kaplan, 1997

Content Security Advantages

- The clearing house only "sells" keys; it does not decrypt content
  - only the purchaser sees the content
- Dedicated opener/viewer software for viewing cryptolopes discourages client side copying
  - the phrase "dedicated software" is almost always bad, and could be considered a showstopper...

Architecture of Digital Library

1. Kahn/Wilensky Framework(KWF)

Digital library objects are more than collections of bits

objects = metadata + data

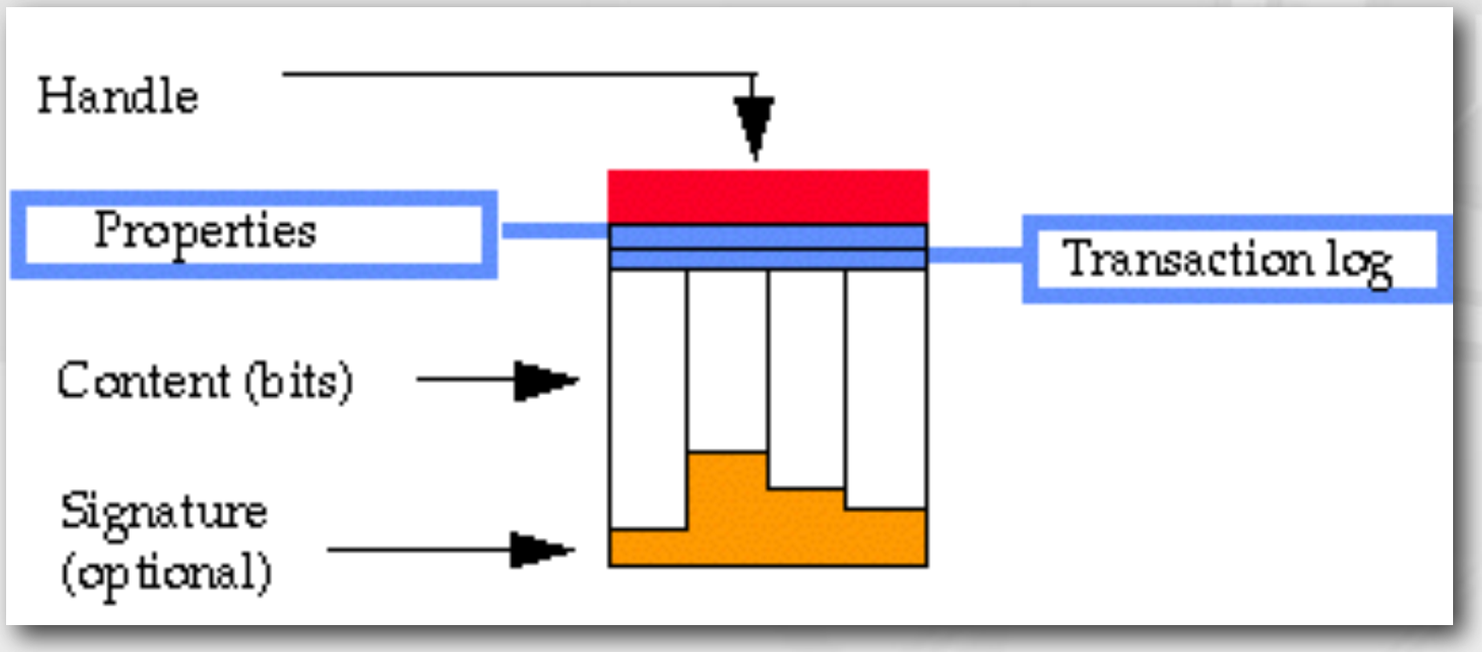


figure 2 in http://www.dlib.org/dlib/July95/07arms.html

Users want intellectual works, but not only digital objects

- The **D**igital **L**ibrary architect's needs should not inconvenience the users' needs
- recombination of objects
  - what is an object in your world view?

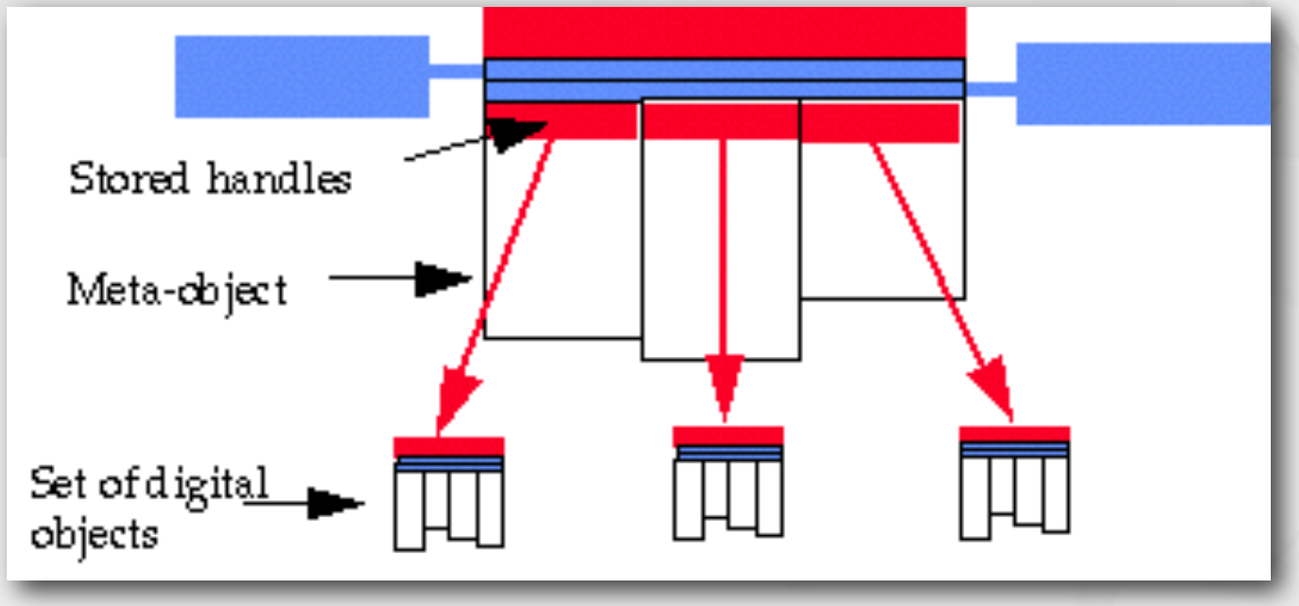
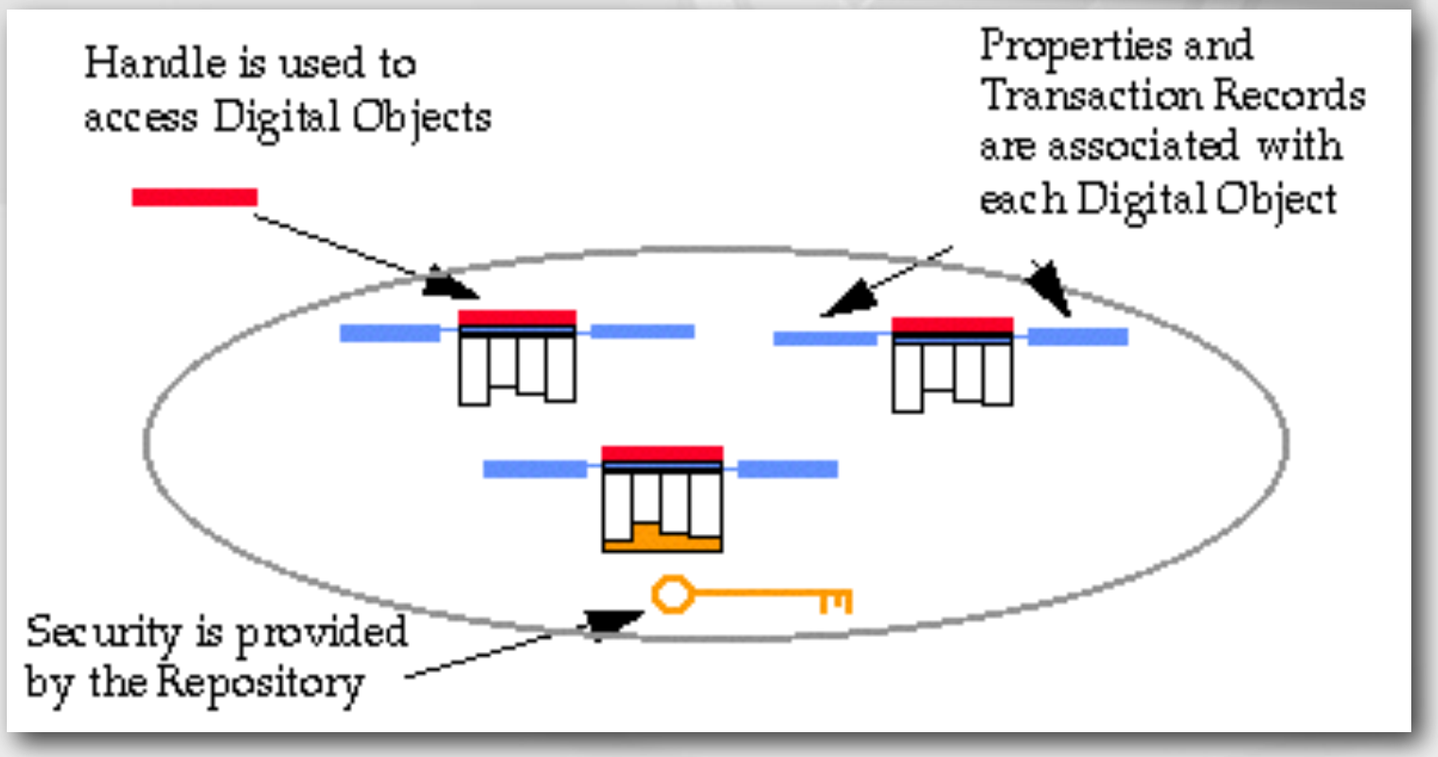


figure 4 in http://www.dlib.org/dlib/July95/07arms.html

Repositories must look after the information they hold

- "Repository Access Protocol"
  - –Kahn Wilensky Framework
    - http://www.cnri.reston.va.us/home/cstr/arch/k-w.html

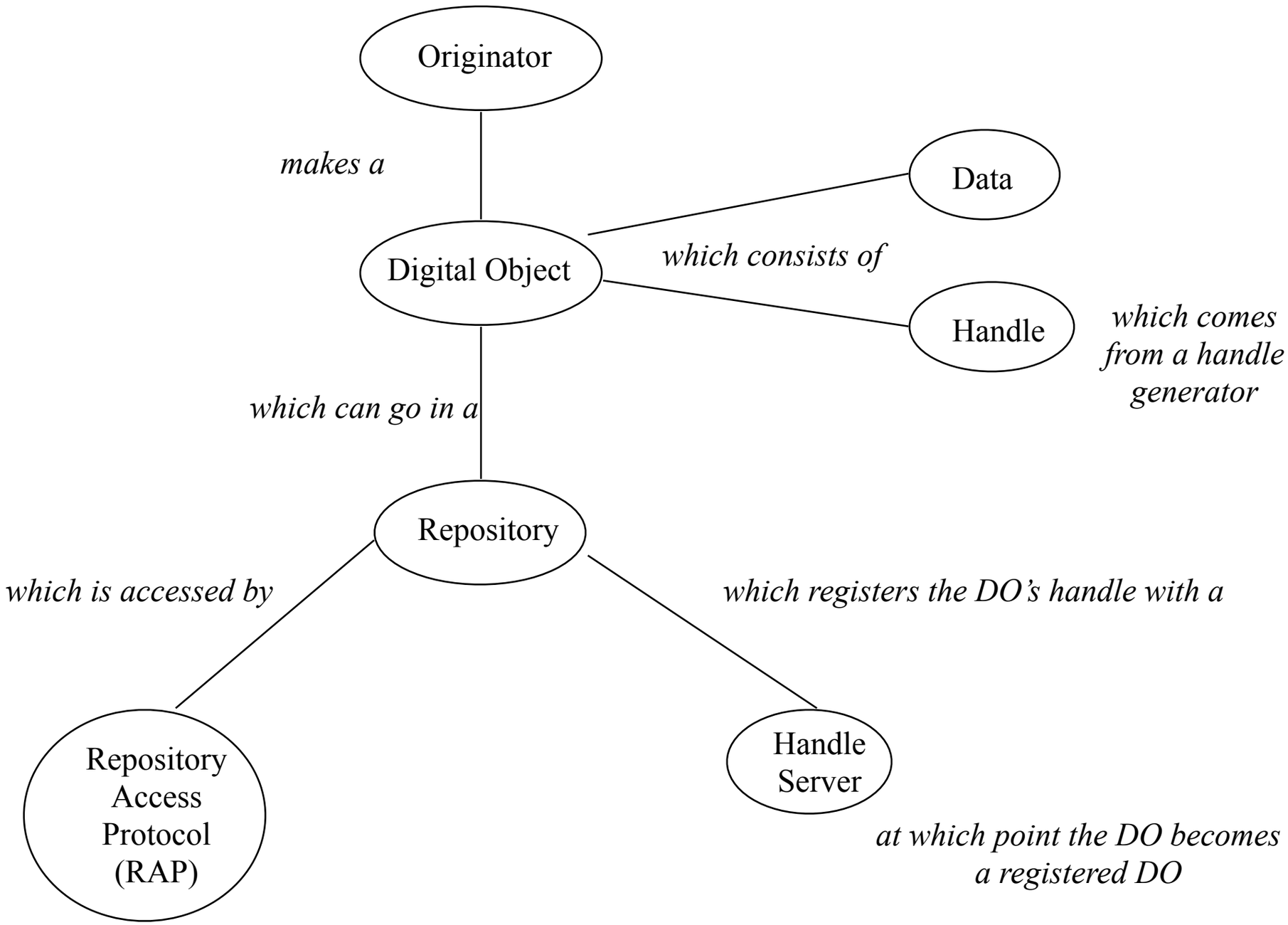
http://www.dlib.org/dlib/July95/07arms.html



Key KWF Terms

- digital objects (DOs)
  - a unit of exchange for the DL with a particular data structure and characteristics
- repository
  - the place where DOs live
- handles
  - a unique, persistent name for a DO

Kahn/Wilensky Framework(KWF)



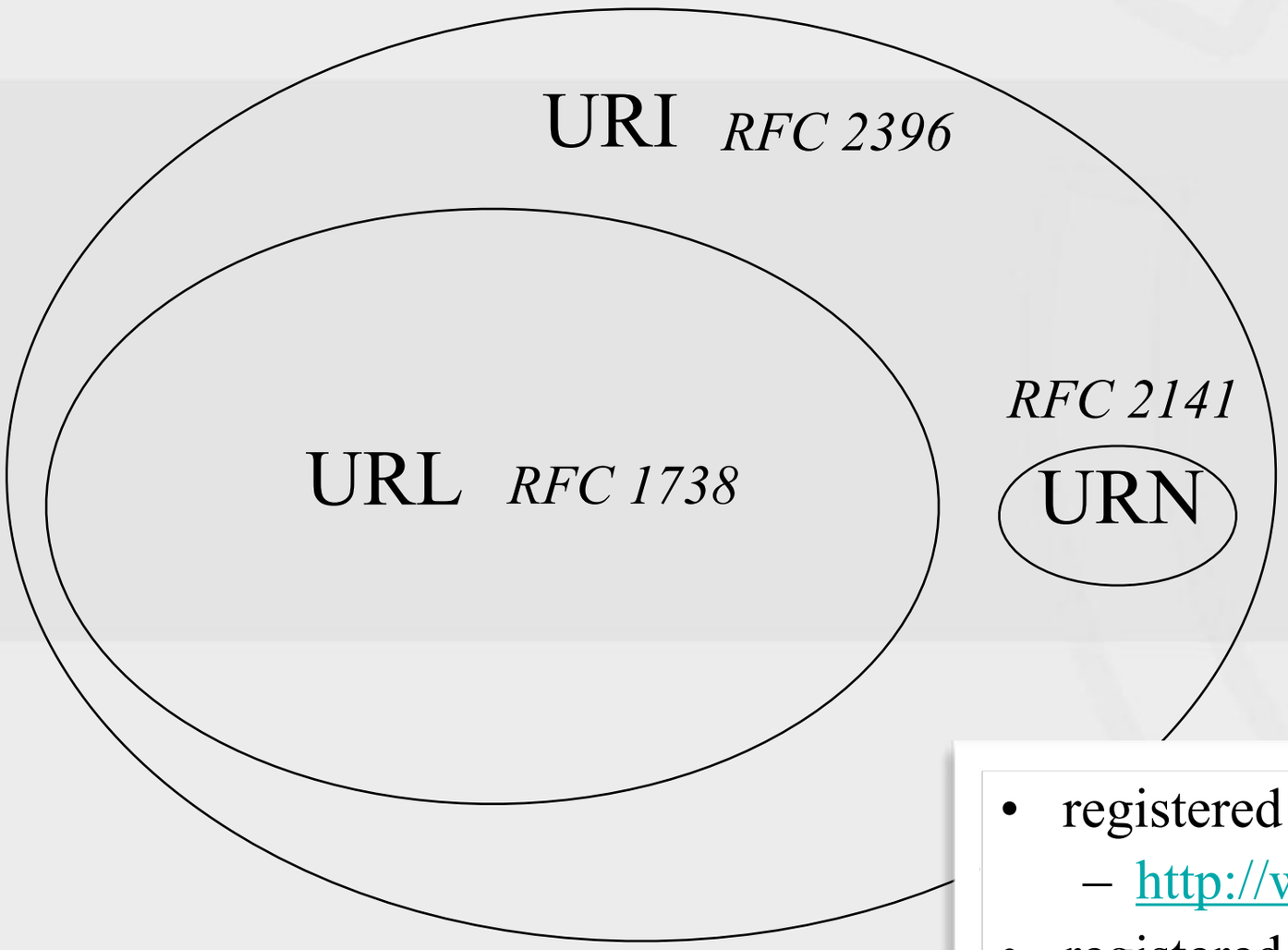
Digital Objects

- Digital object = data + key-metadata
  - data is classified; core classes include:
    - bit-sequence / set-of-bit-sequences
    - digital-object / set-of-digital-objects
    - handle / set-of-handles
  - other types can be defined, and registered with a global type registry
    - definition and registration left undefined
    - similar to MIME?
  - key-metadata includes handle, possibly other metadata (left undefined in KWF)

Digital Objects

- Typed data; example from KWF:
  - a DO subtype: computer-science-tech-report
  - with metadata: author, institution, series, etc.
- Composite DOs:
  - a DO with data of type digital-object
  - non-composite DOs are *elemental* DOs
  - composite DOs can be used to collect similar works together
    - composite DO that contains a DO for each work of Shakespeare...

Uniform Resource Identifiers



- registered URI schemes
  - http://www.iana.org/assignments/uri-schemes
- registered URN namespaces
  - http://www.iana.org/assignments/urn-namespaces

Handles

- Handles can be thought of as a Uniform Resource Name (URN) implementation
  - historical comparison of efforts
    - http://www.dlib.org/dlib/february96/02arms.html

- the handle system (http://www.handle.net/)
  - persistence
  - location independence
  - multiple instances

Repositories

- "A network accessible storage system in which digital objects may be stored for possible subsequent access or retrieval" (KWF)
- A *stored* DO is a DO that resides in a repository
- A *registered* DO is a DO that the repository has registered with a handle server
  - storing and registering can be the same or different processes

Repositories

- A repository keeps a *properties record* for each DO
  - contains key-metadata and any other metadata the repository chooses to keep
- A *repository of record* (ROR) is the first repository that a DO is placed in
  - ROR authorizes additional instances of the DO
- A *dissemination* is the result of an access service request

Repository Access Protocol (RAP)

- "Protocol" may be misleading, its really just the skeleton for a protocol
- RAP is designed to be simple
  - repositories themselves should be simple
- KWF defines 3 basic operation classes:
  - ACCESS_DO
  - DEPOSIT_DO
  - ACCESS_REF
    - this is the catch-all operation for all meta-services...

RAP

- RAP is fleshed out more in Cornell CS 95-TR1540
- Where KWF suggested that the operations would take “metadata”, “key-metadata”, and “digital object” as arguments, TR1540 splits some of those into separate operations
- RAP could be implemented as a subset of a more sophisticated protocol (Dienst, Z39.50, etc.)
  - prelude to the Open Archives Initiative (OAI) metadata harvesting protocol

RAP

Operation	Origin	Description
ACCESS_DO	KWF	requests a dissemination of a DO
VERIFY_DO	TR1540	verify that a DO is in the repository
ACCESS_META	TR1540	access a metadata element of a DO
MUTATE_DO	TR1540	modify data for a DO
MUTATE_META	TR1540	modify metadata for a DO (not key-metadata!)
DEPOSIT_DO	KWF	put a DO in a repository
REPLICATE_DO	TR1540	copy a DO to another repository (new handle)
REINSTANTIATE_DO	TR1540	copy a DO in the same repository (still gets a new handle)
DELETE_DO	TR1540	delete DO from repository and handle from handle server
ACCESS_REF	KWF	return references to servers that perform operations on this repository

Architecture of Digital Library
2. Open Archives Initiative (OAI)

Tiered Model of Interoperability

Mediator services

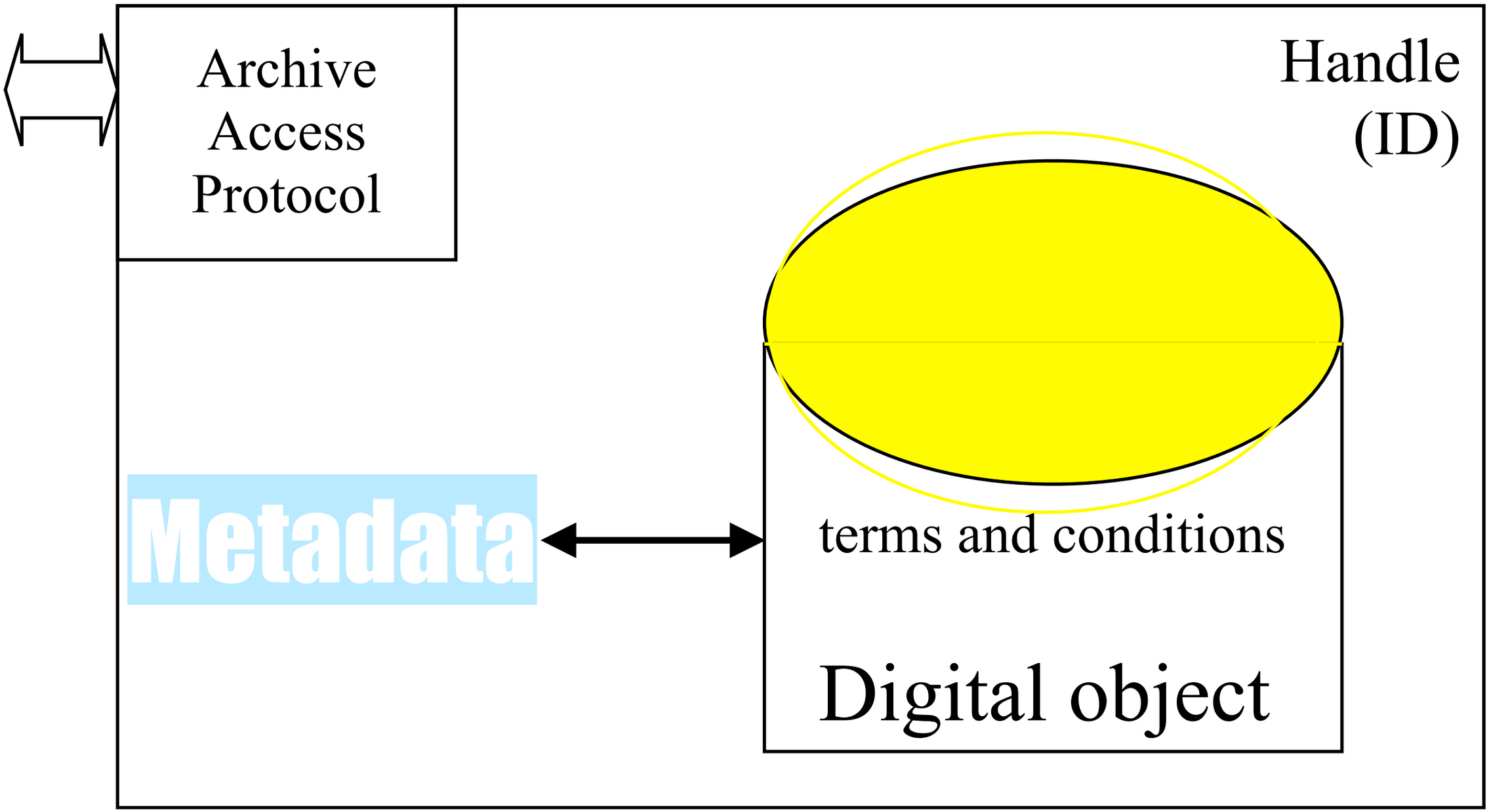
Metadata harvesting

Document models

OAI Philosophy

- Self-archiving = submission mechanism
- Long-term storage system = archive
- Open interface = harvesting mechanism + Data provider + service provider
- Start with "gray literature"
  - e-prints/pre-prints, reports, dissertations, …

Archive of Digital Objects



Required: Protocol

Set Structure

URI Scheme

Required: DC

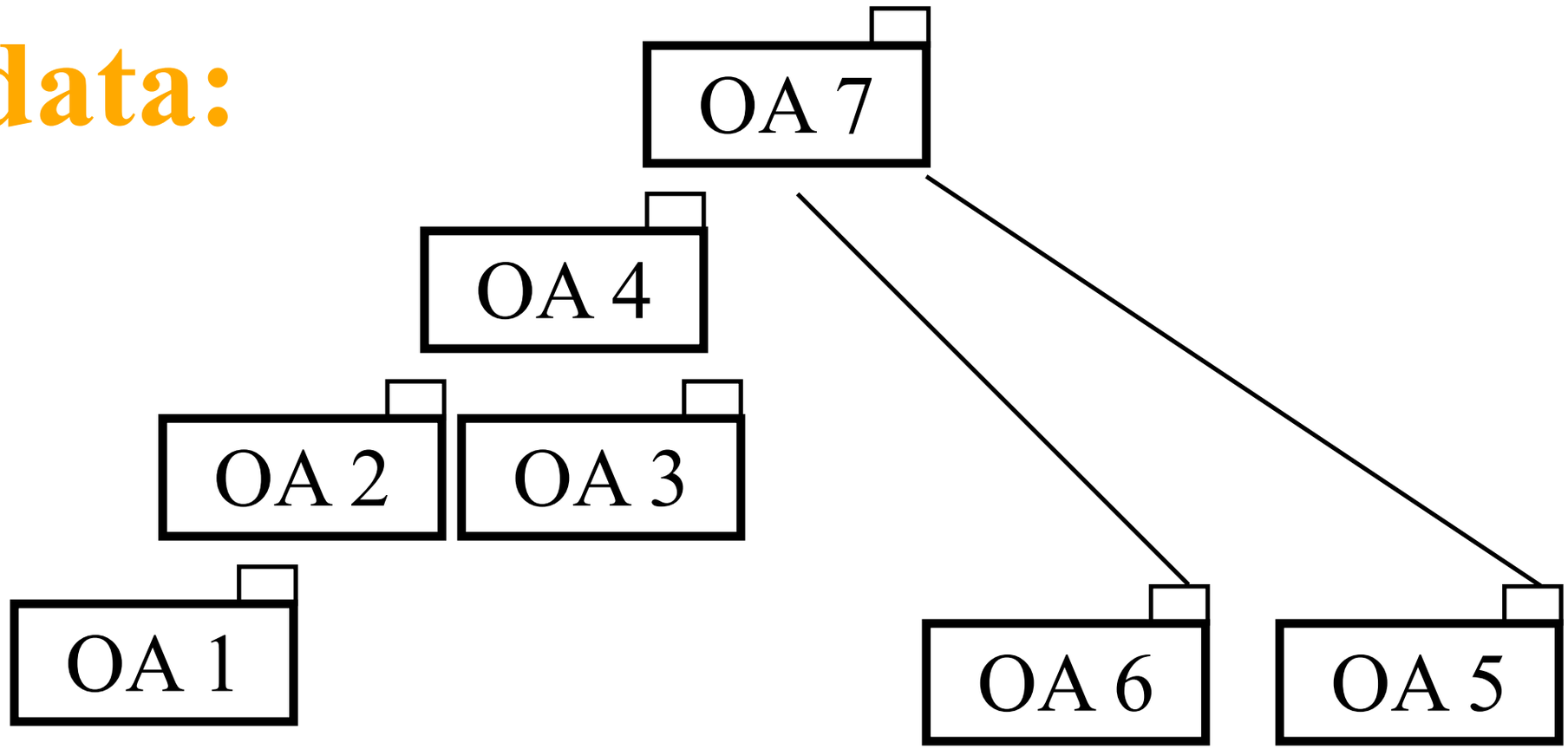
MDO	*MDO*	*MDO*	*MDO*
MDO	*MDO*	*MDO*	*MDO*
DO	DO	DO	DO

OAI – Black Box Perspective

Services:

Search	Browse	Summarize	Visualize

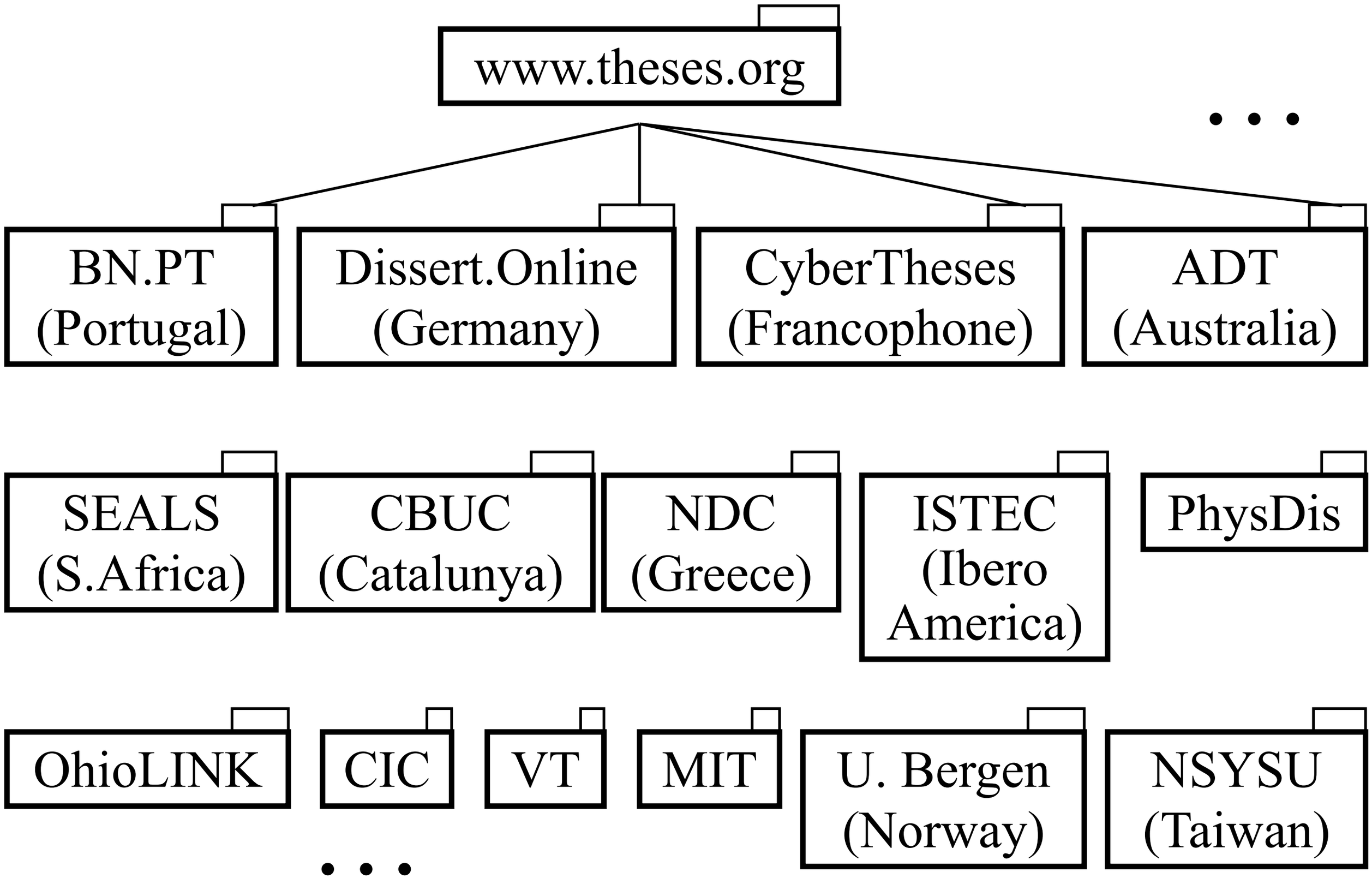
Metadata:



Docs:

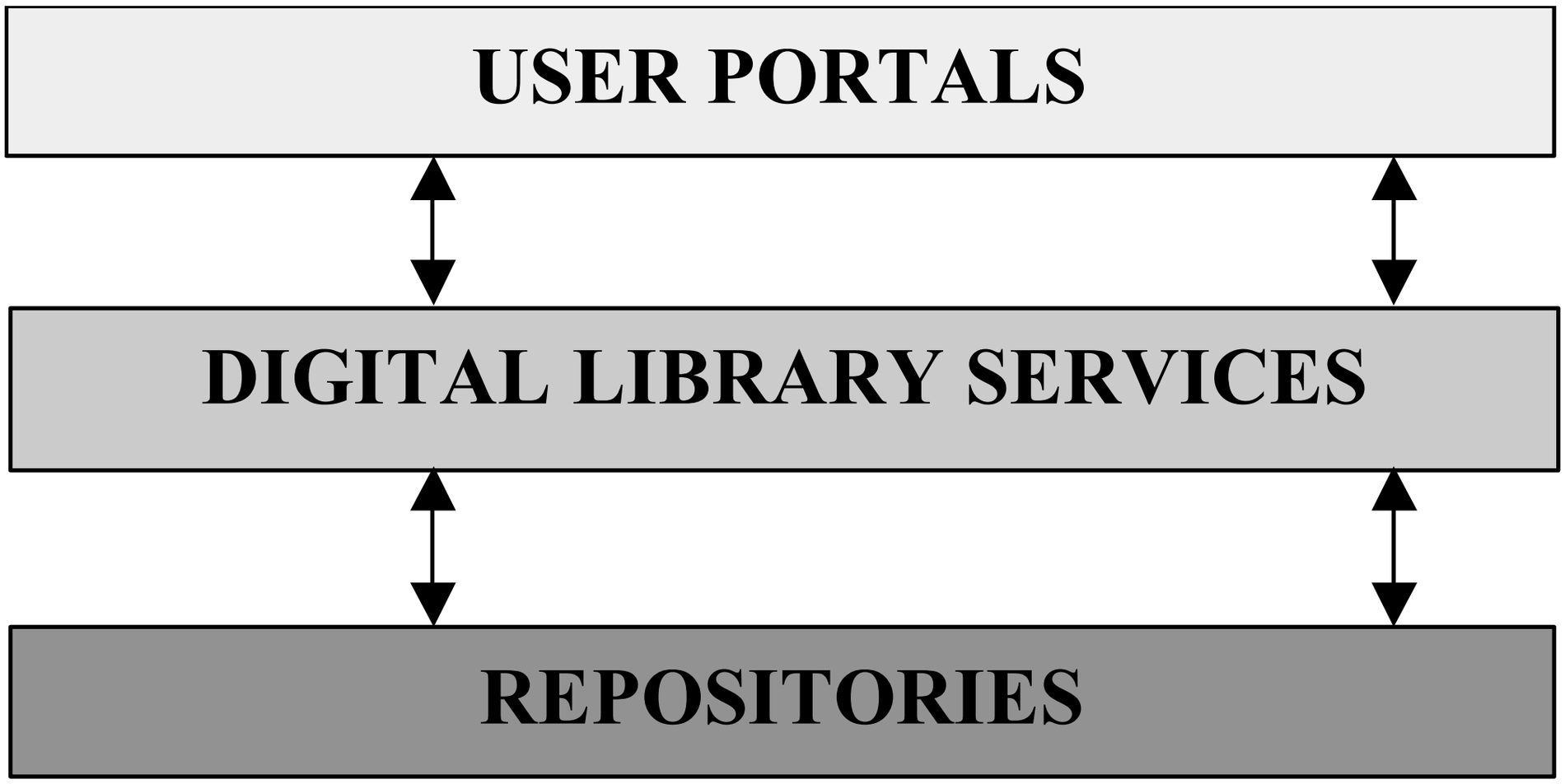
DO DO DO DO DO DO DO

Black Box OAI-ETD Perspective

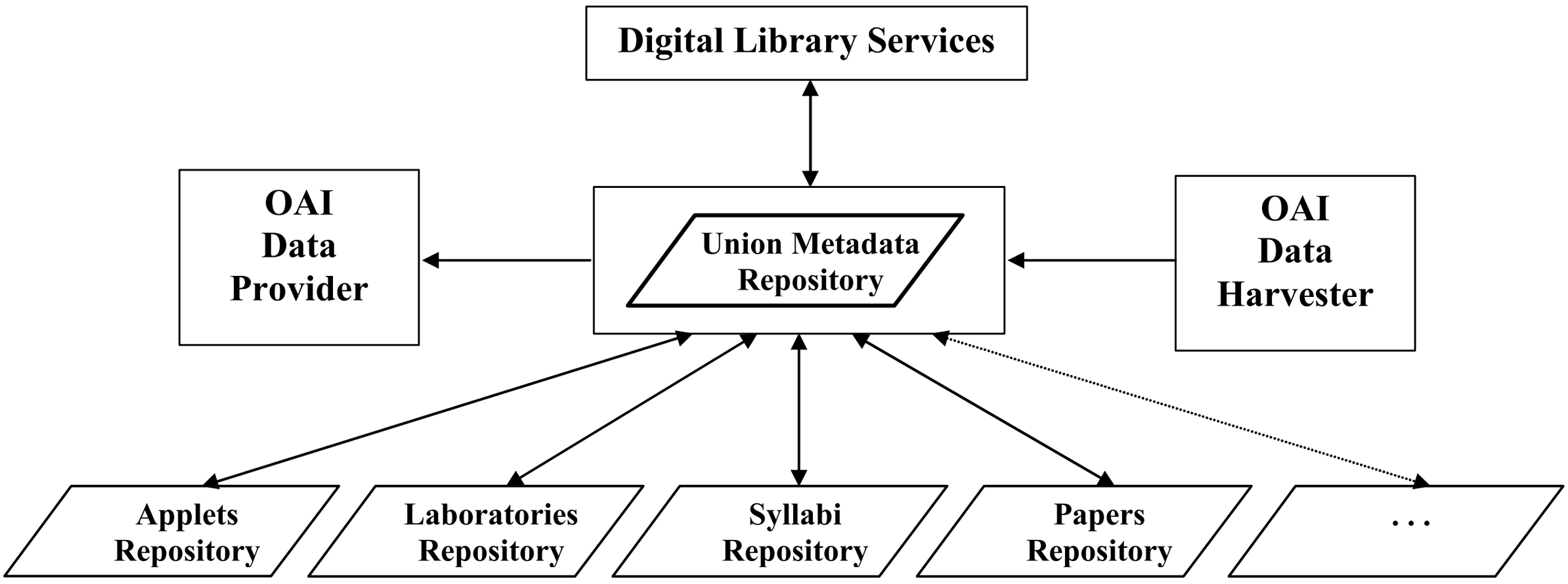


Architecture of Digital Library
3. CITIDEL architecture

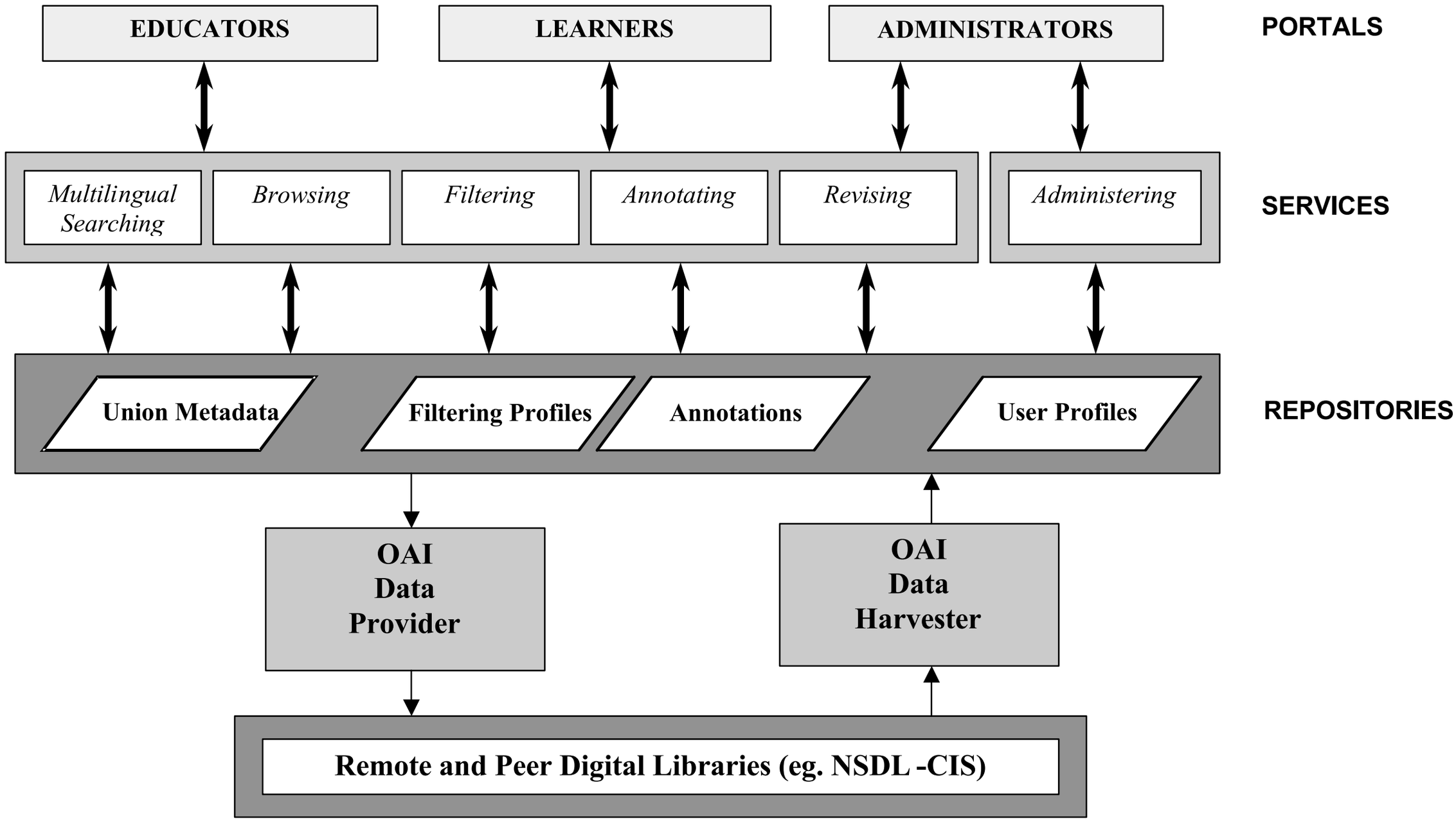
Overview of CITIDEL architecture



Distributed repository structure

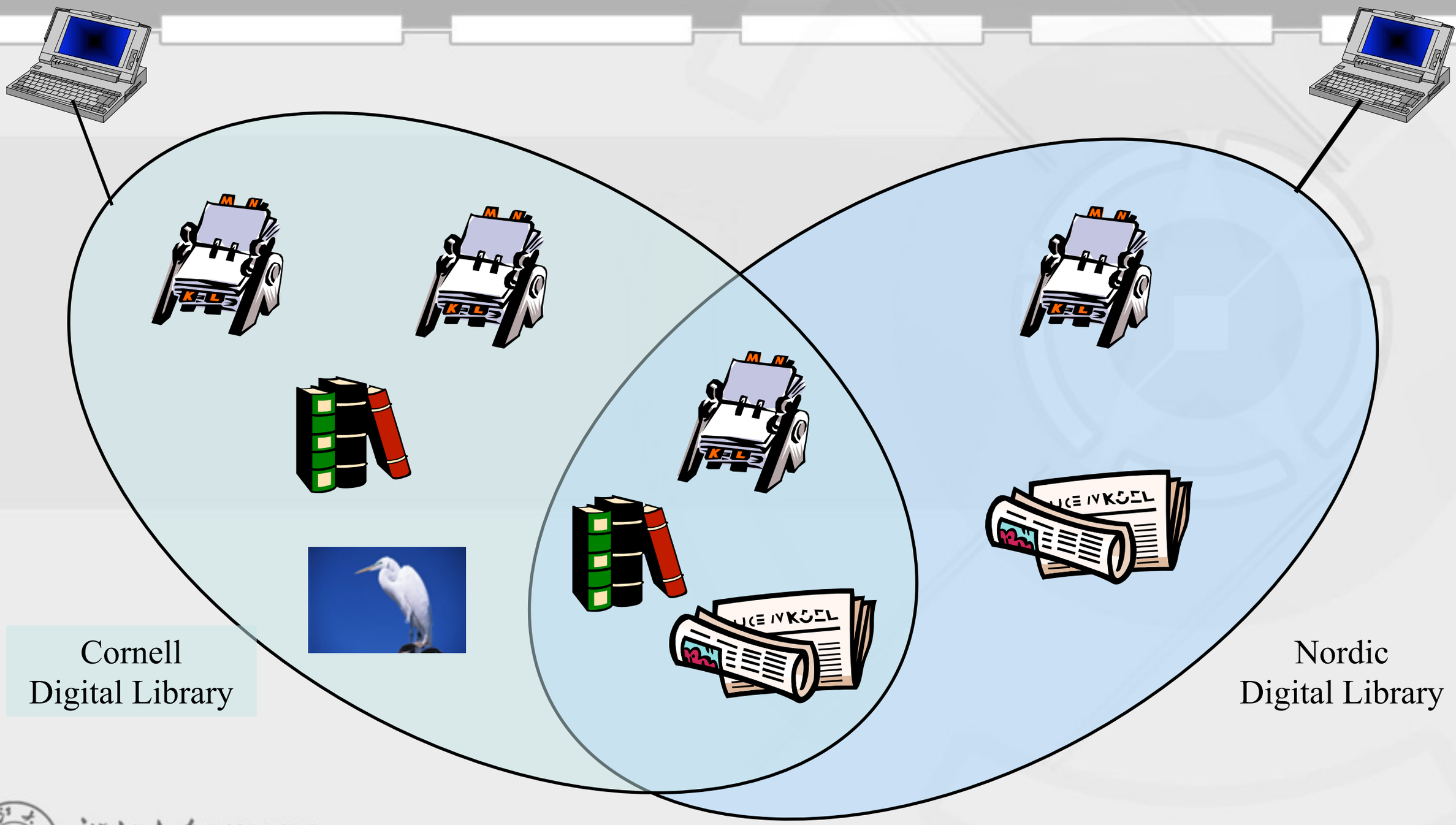


Digital library architecture for local and interoperable CITIDEL services



Architecture of Digital Library
4. New Digital Library Architecture

Digital Library Interoperability



Digital Library Architecture

- Open Architecture
  - functionality partitioned into set of **well-defined services**
  - services accessible via **well-defined protocol**
- Modularization
  - promotes interoperability
  - scalable to different clientele (research library, informal web)
- Federation
  - enable aggregations into logical collections
- Distribution
  - of content (collections) and services
  - of administration and management of DL

Component-Ware Digital Libraries

Collection Services

Name Service

Persistent NAMES

Index Services

User Interface Gateway

Digital Objects

Repository Services

Q&A