Challenges in Digital Libraries
Key Issues Learned from Metadata-Centric Projects at Tsukuba

Shigeo Sugimoto
Research Center for Knowledge Communities
Graduate School of Library, Information and Media Studies
University of Tsukuba
sugimoto@slis.tsukuba.ac.jp

Outline

- Introduction
- Overview of Digital Libraries
  - Looking Back R&D activities
  - Digital Libraries at Libraries
- Metadata-Centric Projects at Tsukuba and Some Lessons Learned
- Challenges in Digital Libraries
- Summary

Introduction

- Digital Library: a perspective
  - DL: integration of many new information technologies in accordance with types of collections, users and environments
    - Technology × Collection × Environment
  - Technology oriented view vs. Service oriented view
- “Challenges in Digital Libraries”
  - Many challenging issues in digital libraries
  - “Interoperability, Preservation, and Theory” (from conversation with Prof. Edward Fox at ICADL’02 in Singapore)

Looking Back DL Projects and Programs

- IT oriented Programs
  - Digital Library Initiative (Phase 1, 2), USA
    - DLI-1(1994-98): NSF, NASA, DARPA
    - DLI-2(1998-2003): NSF, NASA, DARPA, NLM, LoC, NEH
    - Working Groups for Key Topics in DL (EU-NSF)
      - Information Discovery, Metadata, Preservation, IPR issues, etc.
- Digital Libraries for Education
  - National Science Digital Library (NSDL)

Looking Back DL Projects and Programs (cont'd)

- Development of Digital Collections
  - Historical and Rare Materials
    - American Memory by Library of Congress
    - Making of America by Cornell, Michigan, UC Berkeley
    - Digital Library from the Meiji Era by NDL, Japan,

      etc.
  - Scholarly Repositories
    - ePrint Archive, NDLTD, etc.
    - Open Archives Initiative
      - Cooperative activity by repositories

Looking Back DL Projects and Programs (cont'd)

- Metadata for DL and Internet
  - Dublin Core
    - Metadata for Resource Discovery/Description across Domains
  - Domain Specific Schemas
    - Preservation, Education, Government, Accessibility, Geospatial Information, e-Commerce, Rights Management, Video, etc.
  - Semantic Web
    - Metadata Framework
    - Ontology

Overview of Digital Libraries
Fundamental Changes in our Information Environment

- The Internet and WWW
  - The Internet and WWW are important part of information lifeline.
  - Information seeking behavior has been changed by Internet search engines and directory services, e.g. Google and Yahoo!.
- Online Resources for Libraries
  - Scholarly Resources
    - Electronic Journals
    - Scholarly Repositories
    - Digitized Collections
  - Learning Resources
  - Governmental Resources

Overview of Digital Libraries
Fundamental Changes in our Information Environment (cont'd)

- Statistics from 2005 White Paper Information and Communications in Japan, Ministry of Internal Affairs and Communications
  - Internet Users: 80M
  - PC: 61M, Mobile Phone: 58M
  - PC only: 21M, Mobile Phone only: 15M

Overview of Digital Libraries
DL Functions in Library Services

- Building Collections of Primary Resources
  - Electronically Published Resources
    - E-Journals
    - Repositories, etc.
  - Digitized Resources
- Navigational Services (Secondary/Meta Resource Collection)
  - Subject Gateways, Portals
  - Reference Services
- Preservation of Digital Resources
- User Environment Building
  - Personalization
  - Hybrid Library

Metadata-Centric Projects at Tsukuba Outline

- Subject Gateway Projects
  - ULIS-DL
    - Gateway for Library and LIS Resources
    - Simple Dublin Core
  - IPL-Asia
    - Resources for Public Library Users in Chinese, Japanese and Korean (CJK) languages
    - Metadata in CJK
  - Collaboration with Okayama Prefectural Library
    - Digital Okayama Dai-Hyakka
      - Regional Portal
  - On Going Activity
    - Gateway to Digital Collections in Public Libraries in Japan
      - Created approximately 5000 records for resources published in 170 library Web sites

Metadata-Centric Projects at Tsukuba

- Metadata Schema Projects
  - Metadata Schema Registry
    - Collaboration with DCMI
  - Layered Model for Metadata Interoperability
  - On Going Activity
    - Metadata Framework for Context Sensitive Resource Selection and Adoption
- Why Metadata-Centric Projects?
  - Value addition by metadata
  - Semantic Web Technologies, i.e., RDF/XML, OWL, etc.

ULIS-DL
Building a Core Subject Vocabulary

- Outline of ULIS-DL
  - Subject gateway for resources published by libraries and LIS institutions.
  - Metadata records created based on Simple Dublin Core.
  - ULIS-DL has a retrieval function but no directory style interface to browse and navigate the contents.
  - Subject terms are given as free terms, i.e. no controlled vocabulary.
- Goal of our research
  - To create a subject vocabulary for directory style interface of ULIS-DL

ULIS-DL
Building a Core Subject Vocabulary

- Status as of 2003
  - 26,000 metadata records
  - 16,000 distinct text strings In the *Subject* element of the raw metadata records, including typos and synonyms
- Issue: How to choose appropriate subject terms
- Methodology to build a Core Subject Vocabulary
  - Extract terms of Subject elements which appear more than once.
  - Measure coverage/uncoverage ratio of the term sets (Candidate Term Set-n).
- Result: approximately 90% of collected sites is covered by 1025 subject terms.

ULIS-DL
Building a Core Subject Vocabulary

	Subject terms	Excluded records	Uncoverage ratio
CTS-2	3979	1519	6%
CTS-3	2045	2083	8%
CTS-4	1366	2590	10%
CTS-5	1025	2801	11%

- Findings
  - 90% of the whole resources can be covered by CTS-5.
  - “1000 words” is a manageable size for manual encoding in XML.
  - Includes NDC terms, proper nouns, ULIS-DL specific terms, etc.

IPL-Asia Metadata (Chinese, Korean, Japanese, English)

Element	Korean	Japanese	Chinese	English
TITLE				
MAIN-TITLE	Tour2 Korea.com	Tour2 Korea.com	Tour2 Korea.com	Tour2 Korea.com
SUB-TITLE	한국관광공사	韓国観光公社	韩国国际旅行社	Korea National Tourism Organization
PUBLISHER				
CURRENT	한국관공공사	韓国観光公社	韩国国际旅行社	Korea National Tourism Organization
IDENTIFIER	http://www.knto.or.kr/Korean/index.html	http://www.knto.or.kr/Korean/index.html	http://www.knto.or.kr/Korean/index.html	http://www.knto.or.kr/Korean/index.html
DESCRIPTION				
LONG	한국 관광 공사의 홈페이지로, 구성은 크게 원천 서비스, 멀티 미디어 서비스, 여행 안내, 공사 신문 청사초롱, 추천 코너로 되어져 있다. 원천서비스 한국 관광 공사가 보유하고 있는 관광 정보 데이터 베이스를 대외에 개방, 공사 관광 정보 이용을 원하는 법인을 대상으로 관광정보 이용 회원을 모집하여 정보를 제공하는 서비스이다. 멀티 미디어 서비스에서는 사진 공모전의 수상작들과 공사의 홍보 비디오, 포스터, 홍보 광고, 관광 달력등을 이용할 수 있다. 여행 안내에서는 관광지, 문화시설, 스포츠 시설, 레저, 음식적, 쇼핑, 교통, 관광 상품, 주요 연락처 별로 검색을 할 수 있고, 그 외에도 지역별 검색과, 행사안내, 관광지도 정보 등도 제공하고 있다. 그리고, 추천할 만한 여행지, 지하철 관광 코스등 한국을 여행하기 전 꼭 한번 체크해 볼 만한 페이지다. 한국어 페이지 외에도 영어, 일본어, 독일어, 에스파냐어, 프랑스어, 중국어의 페이지가 준비되어져 있다.	韓国観光公社のホームページで、構成は大まかに源泉情報サービス、マルチメディアサービス、旅行案内、公社の新聞、お勧めコナーから成っている。源泉サービスは韓国観光公社が所有している観光情報のデータベースを、公社の観光情報を必要とする法人を対象として会員を募集して情報を提供するサービスである。マルチメディアサービスでは、写真公募展の受賞作と公社の広報ビデオ・CM、観光カレンダー等が利用できる。旅行案内では、観光地、文化施設、スポーツ施設、レジャー、飲食店、ショッピング、交通、観光商品、重要連絡先別に検索でき、その他にも地域毎の検索、行事案内、観光地図の情報等も提供している。それからお勧めの観光地、地下鉄観光コースなど、韓国を旅行する前に一度チェックしてみる価値があるページである。韓国語のページ以外にも英語、日本語、ドイツ語、エスパニャ語、中国語のページが用意されていて、その内容は異なっている。	韩国国际旅行社的主页是公司由经营状况信息服务、多媒体服务、旅游向导、公司新闻、热力推荐等部分组成的。经营状况信息服务是韩国国际旅行社为募集需要公司旅游信息的会员合作者所提供的公司所有的旅游信息数据库服务。多媒体服务有照片公开募集展览的获奖作品、公司的广报录像、广告、旅游日历等。旅游向导可对观光地、文化设施、运动设施、休闲地、饮食店、购物、交通、观光纪念品、重要联络地址等分别进行检索，并且还提供通过地域的检索、行事向导、观光地图等。除此之外，还有推荐观光地、地下铁观光路线等。在出游韩国之前，此网页值得事先浏览。除了提供韩国语的主页之外，还提供有英语、日语、德语、西班牙语、中国语的主页，但内容有所不同。	The homepage of the Korean National Tourism Organization is broadly composed of the following categories: tourism information service, multimedia service, travel guide, newsletter and recomm- endations corner. The tourism information service, mainly used by its corporate members, distributes tourism information from the organizations's database. Multimedia service users can access a gallery of prize-winning photos from a tourism photo contest, the organization's promotional video and tourism advertisments, as well as a calendar of events. Users of the travel guide can search for information under the major subheadings of travel destinations, culture facilities, venues for leisure and sporting events, restaurants, shopping, tranportation, and souvenirs. Users can also search for event guides and tourist maps based on geographic region. Also displayed are major tourist sights and subway tours; these pages are definitely worth visiting at least once before visiting Korea. In addition to

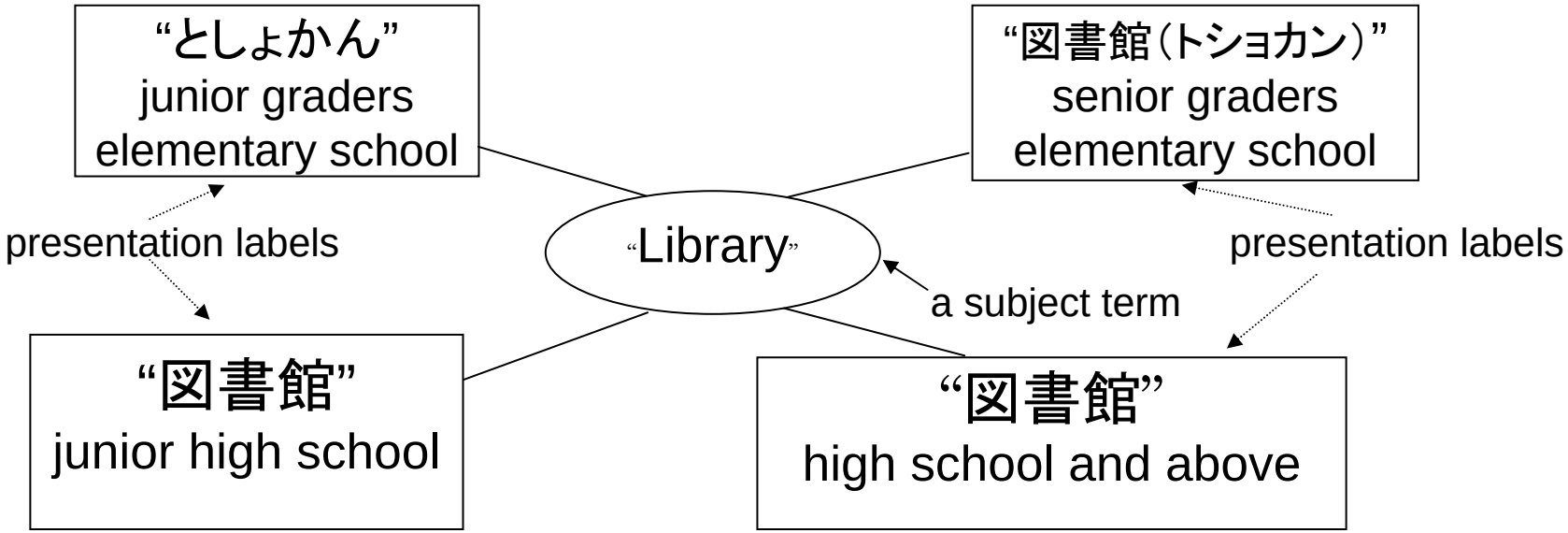
IPL-Asia

- IPL-Asia
  - Provides information about CJK resources useful as a public library resources.
  - Provides resource information in CJK languages.
- Lessons Learned
  - Domain oriented subject vocabularies need not be large but need community-specific terms, e.g., school activities and regional activities.
  - User interface for children requires to represent subject terms in accordance with their ages.
  - Costs

Digital Okayama Dai-Hyakka (DODH)

- Regional Portal by the Library of Okayama Prefecture
- Metadata Creation by Librarians and Non-professionals, e.g. School Teachers, Students, and Volunteers.
- Three Subject Vocabularies
  - NDC (Nippon Decimal Classification)
  - Kid's Vocabulary
    - Multiple labels in accordance with user ages
  - Prefectural Resource Vocabulary
- Small Set of Subject Terms Usable for the Non-professionals designed in accordance with Regional Needs.

Multiple Labels for a Single Term



- Multiple labels for a single concept in accordance with type of audience to improve accessibility
  - Encoding in XML is straightforward, e.g. OWL, XML TopicMaps.

Comparison of Subject Vocabularies - Okayama's Case -

- Comprehensive and conventional subject vocabularies are not always useful for domain-specific resources.
- Comparison between Subject Vocabularies used in Okayama
  - Prefecture government's subject vocabulary for governmental resources (Prefecture Vocabulary)
  - A subject vocabulary for children (Kids Vocabulary)
  - Three vocabularies used in DODH
    - NDC, PV, and KV
    - Mappings between all pairs of these three vocabularies

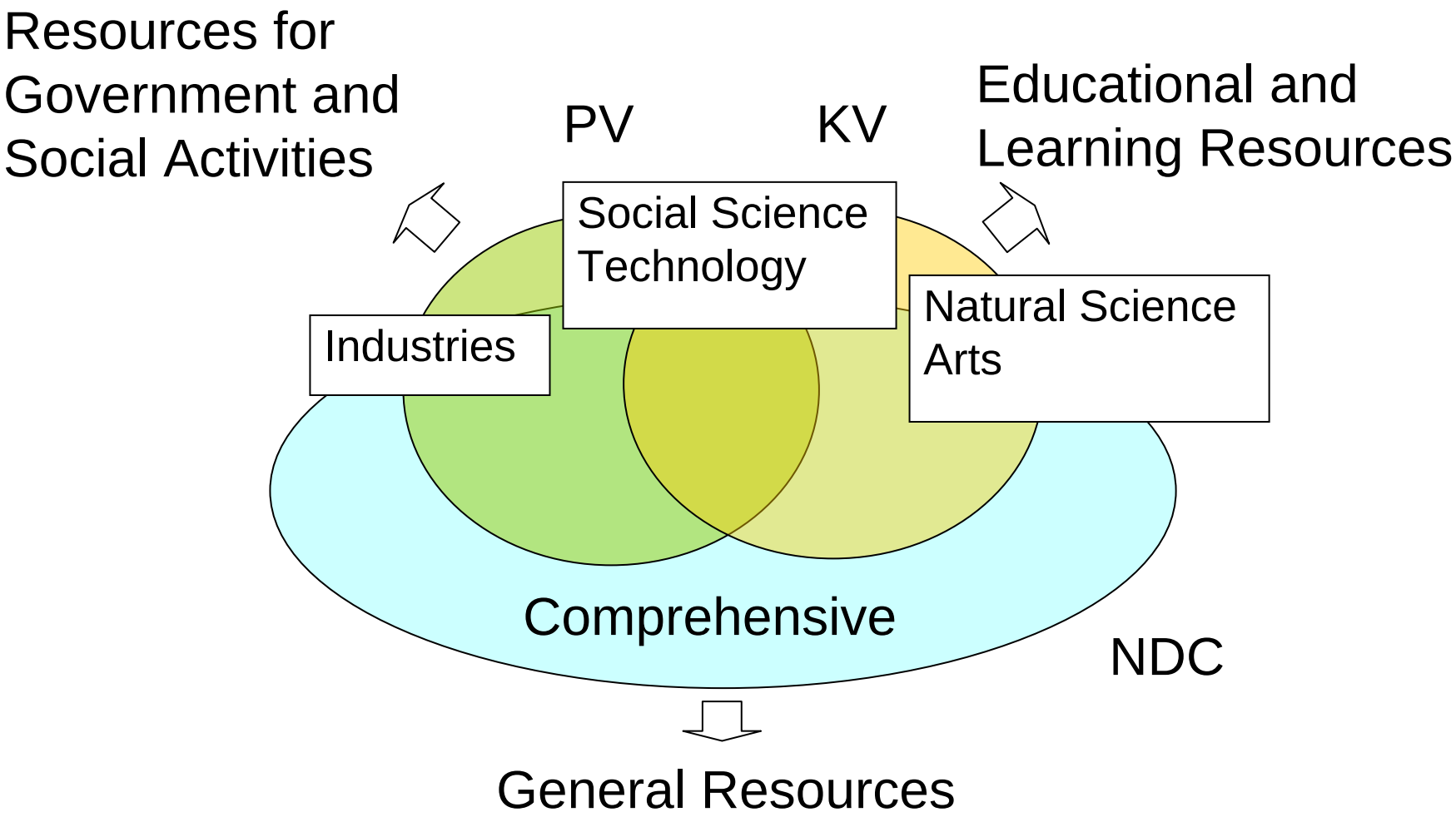
Distribution of Terms in the NDC term space - Okayama's Case -

NDC	000	100	200	300	400	500	600	700	800	900	total
#Kids terms	17	8	8	196	58	54	28	62	6	6	443
#NDC in KV	7	7	3	44	27	27	20	26	4	4	169
#Pref terms	15	2	12	171	30	56	44	17	1	1	349
#NDC in PV	4	2	5	34	11	18	25	15	1	1	116

NDC: 000=Generalities, 100=Philosophy, 200=History, 300=Social Sciences, 400=Natural Sciences, 500=Technology, 600=Industry, 700=The Arts, 800=Language, 900=Literature

#NDC in KV/PV: the number of NDC terms in x00 used in the KV/PV mapping

Subject Vocabularies



Some Lessons Learned from Subject Gateway Projects

- Subject vocabularies were always the central issue for these projects.
  - Conventional and comprehensive vocabularies are not always useful because of the narrow domain.
  - (Reasonably) small subject vocabulary seems useful for end-users.
  - Domain-specific/regional vocabularies are useful, however vocabulary maintenance technology should be investigated.
    - Semantic Web technologies
- Issues
  - Interoperability vs. Domain Specificity
  - Maintenance
  - Reusability and Customizability of Metadata Vocabularies

A Layered Model of Metadata Schema Model

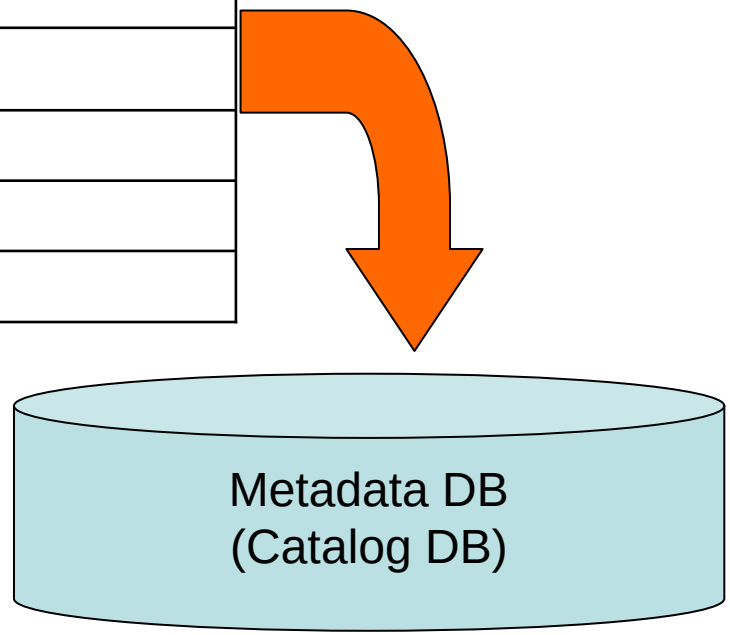
- Metadata Schema
  - Element Set, Application Profiles, XML Bindings, etc.
- Splitting metadata schema into Three Layers in Metadata Schema
  - Semantics
    - Metadata Vocabularies
  - Abstract syntax
    - syntax definition neutral to implementation technology
    - Application Profile
  - Concrete syntax
    - syntax definition to encode metadata instances

A Concrete Syntax: a Catalog Card

Title	Challenges in Digital Libraries
Creator	Sugimoto, Shigeo
Creator	杉本, 重雄／すぎもと, しげお
Subject	Metadata Schema, Interoperability
Publisher	Nagoya University Library
Publishing Location	Nagoya, Japan
Date	2005-8-25
Language	English

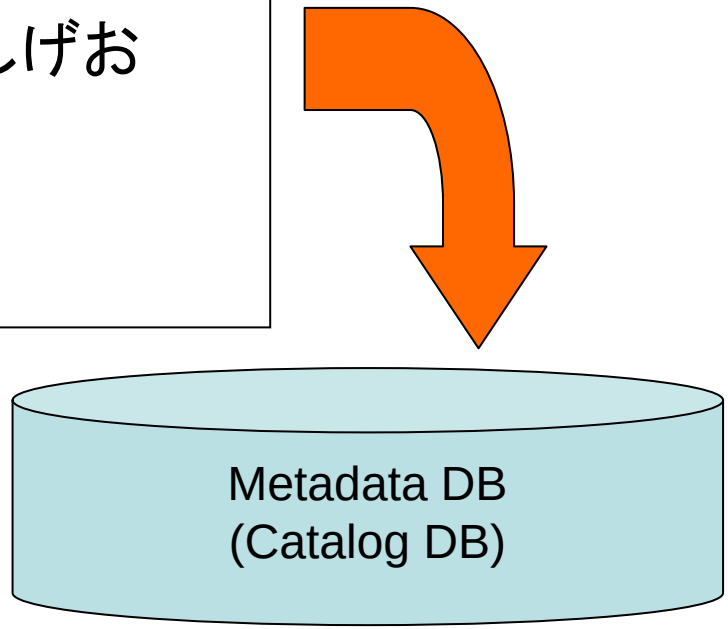
A Concrete Syntax: a DB record

Title (en)	Challenges in Digital Libraries
Creator (en)	Sugimoto, Shigeo
Creator (ja)	杉本, 重雄　（pronunciation: すぎもと, しげお）
Subject (en)	Metadata Schema
Subject (en)	Interoperability
Publisher (en)	Nagoya University Library
Publishing Location (en)	Nagoya, Japan
Date (ISO-8601)	2005-8-25
Language (ISO-639)	en



A Concrete Syntax: an XML instance

```
<dc:title> Challenges in Digital Libraries
</dc:title>
<dc:creator> Sugimoto, Shigeo</dc:creator>
<dc:creator xml:lang="ja">
    <rdf:value>杉本, 重雄 </rdf:value>
    <uliscore:pronunciation>すぎもと, しげお
    </uliscore:pronunciation>
</dc:creator>
 …
```



An Abstract Syntax

Syntactic features neutral to implementation syntax

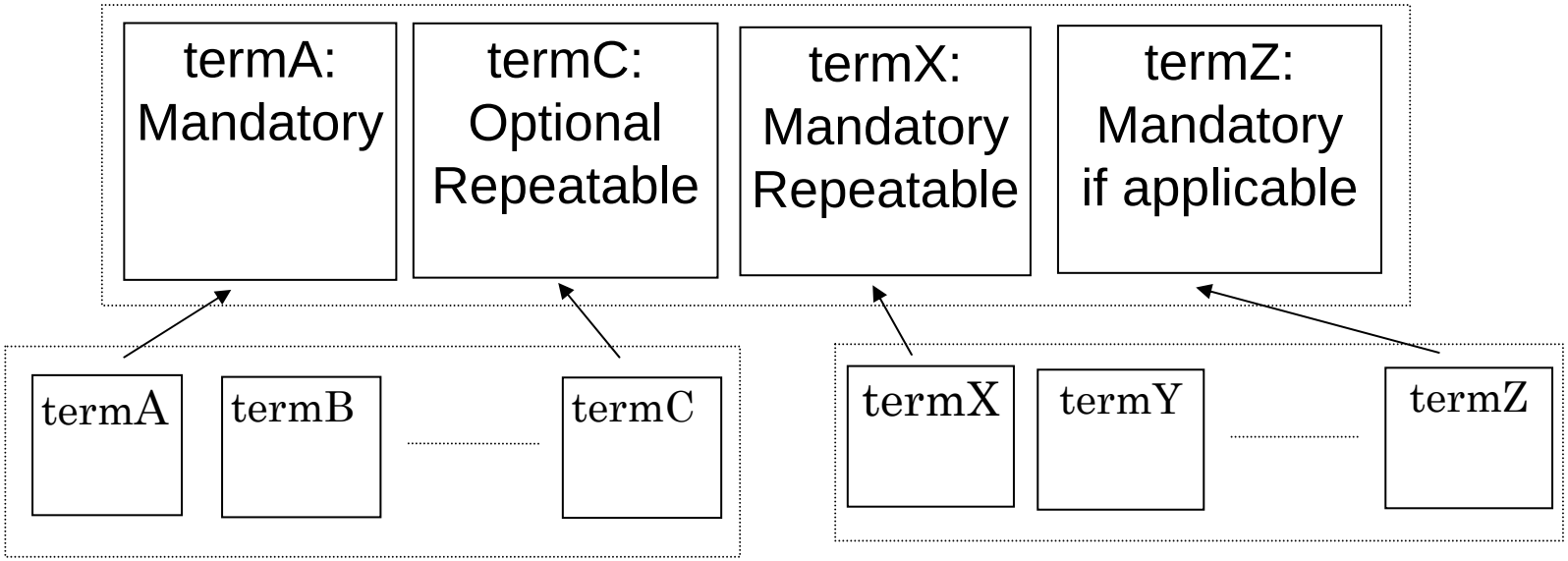
Title: Mandatory
Creator: Mandatory if Applicable, Repeatable
Subject: Optional, Repeatable
etc.

Semantics - Definition of Terms

Title
Element Name: Title
Label: Title
Definition: A name given to the resource.
Comment: Typically, Title will be a name by which the resource is formally known.

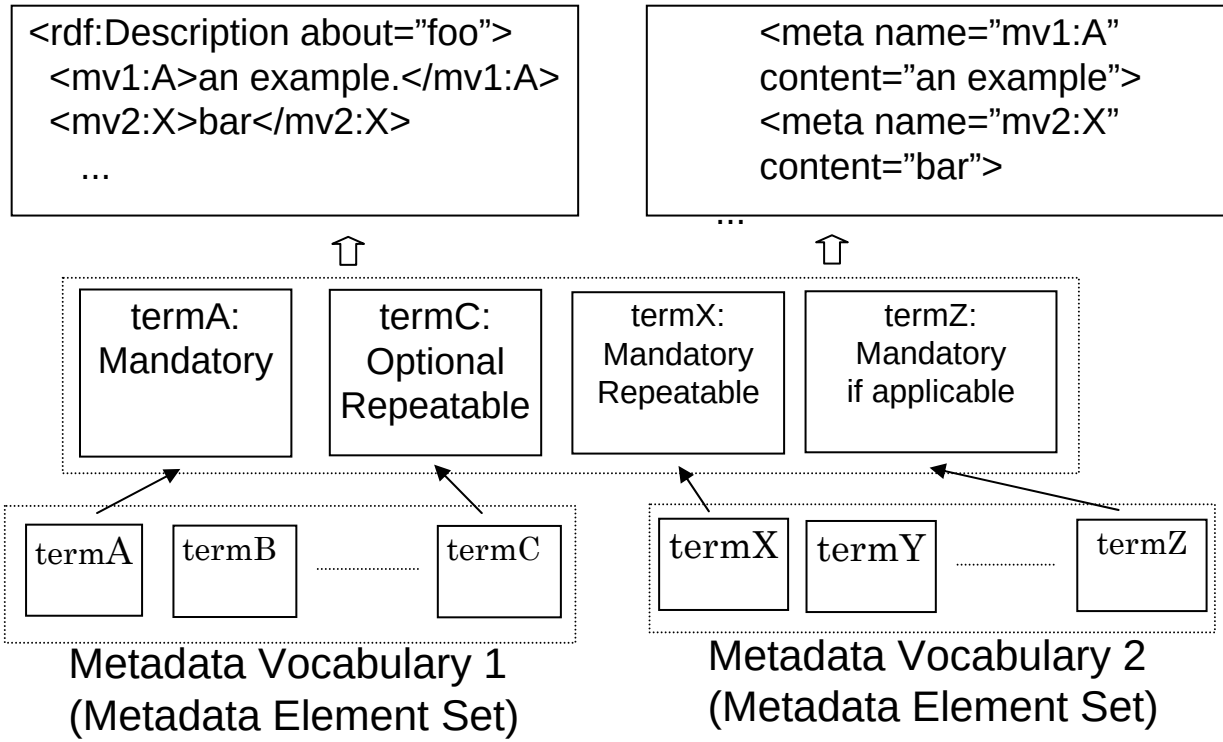
Creator
Element Name: Creator
Label: Creator
Definition: An entity primarily responsible for making the content of the resource.
Comment: Examples of Creator include a person, …

Application Profile



A Structural View of an Application Profile

Abstract Syntax and Concrete Syntax

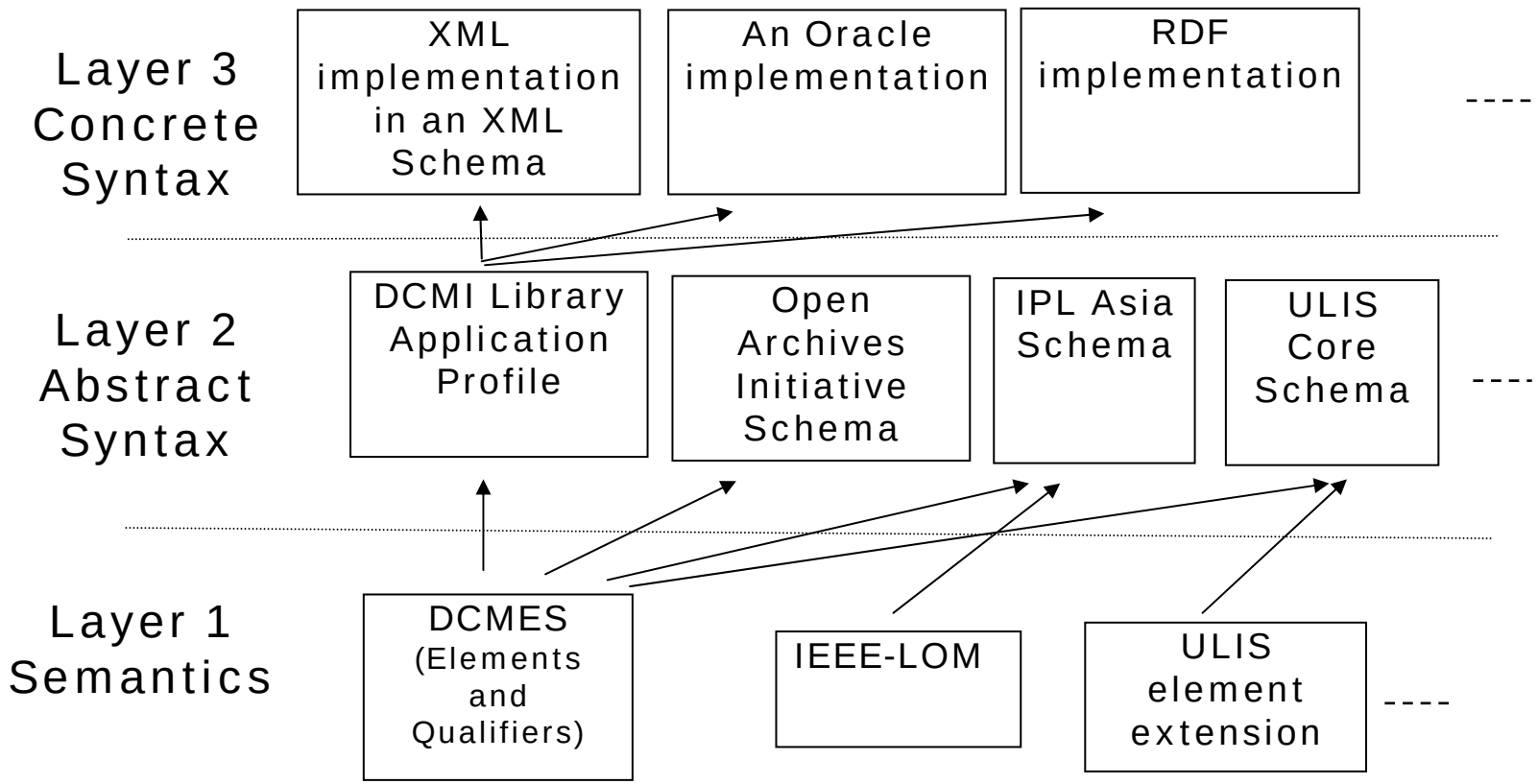


Application Profile

Description in a syntax defined in an application

Application Profile: Terms used in an application and structural constraints

A Layered Model
split semantics and syntax into layers

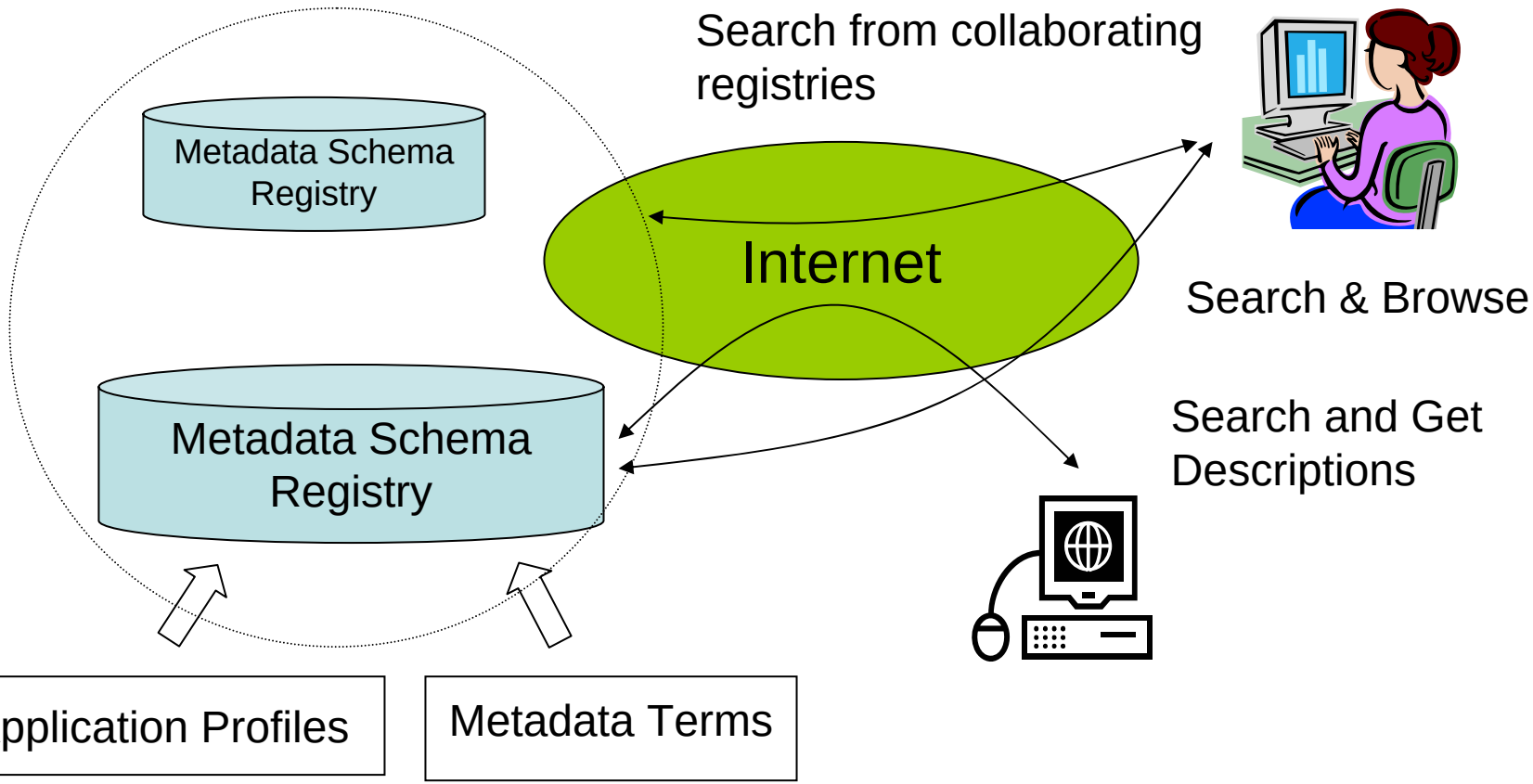


Layered Model of Metadata Schema

Metadata Schema Registry

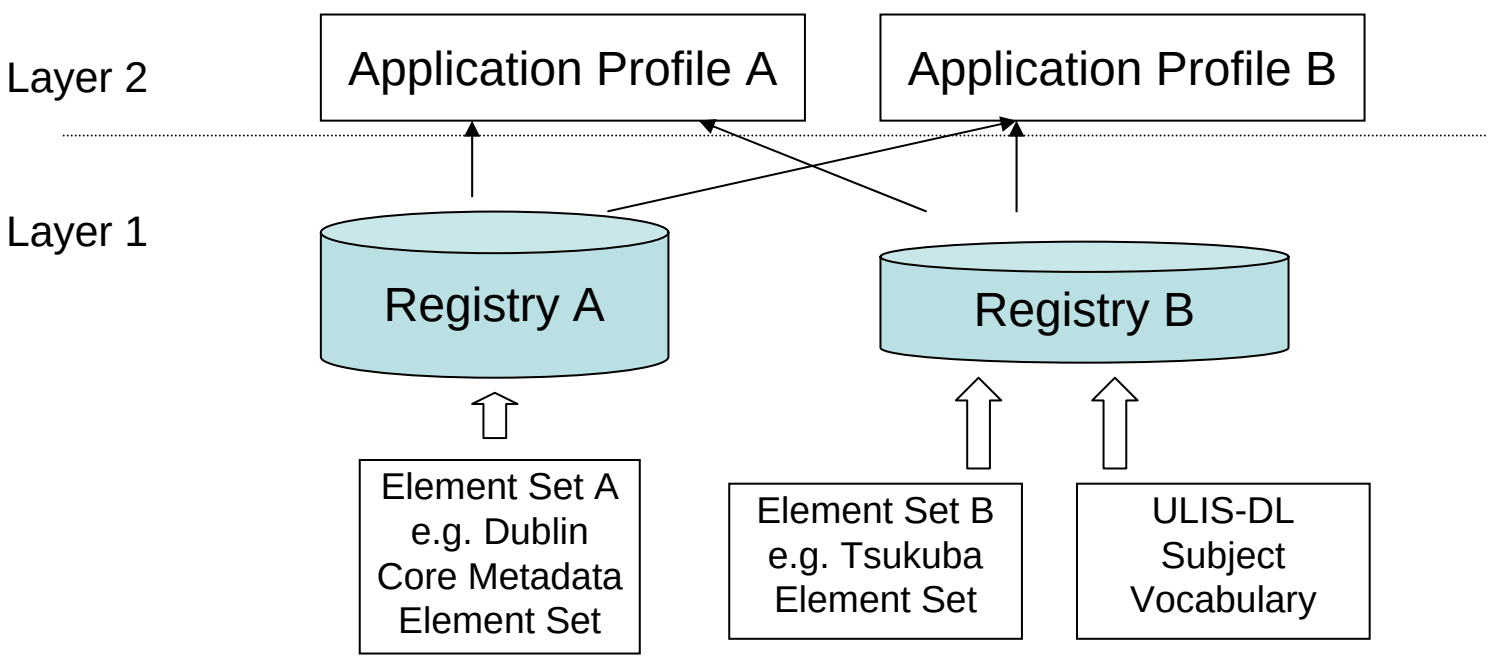
- Primary Function
  - Store reference descriptions of metadata schemas
    - Metadata Terms, i.e. vocabularies, element sets
    - Application Profiles
  - Provide the reference descriptions not only to human users but also to machines via the Internet
    - Searching and browsing functions for human users
    - Application program interfaces (APIs) via the Web for machines
    - Metadata Schema in an Understandable Form both for humans and machines
- Collaborating Distributed Schema Registries
  - Enhance information sharing

Metadata Schema Registry



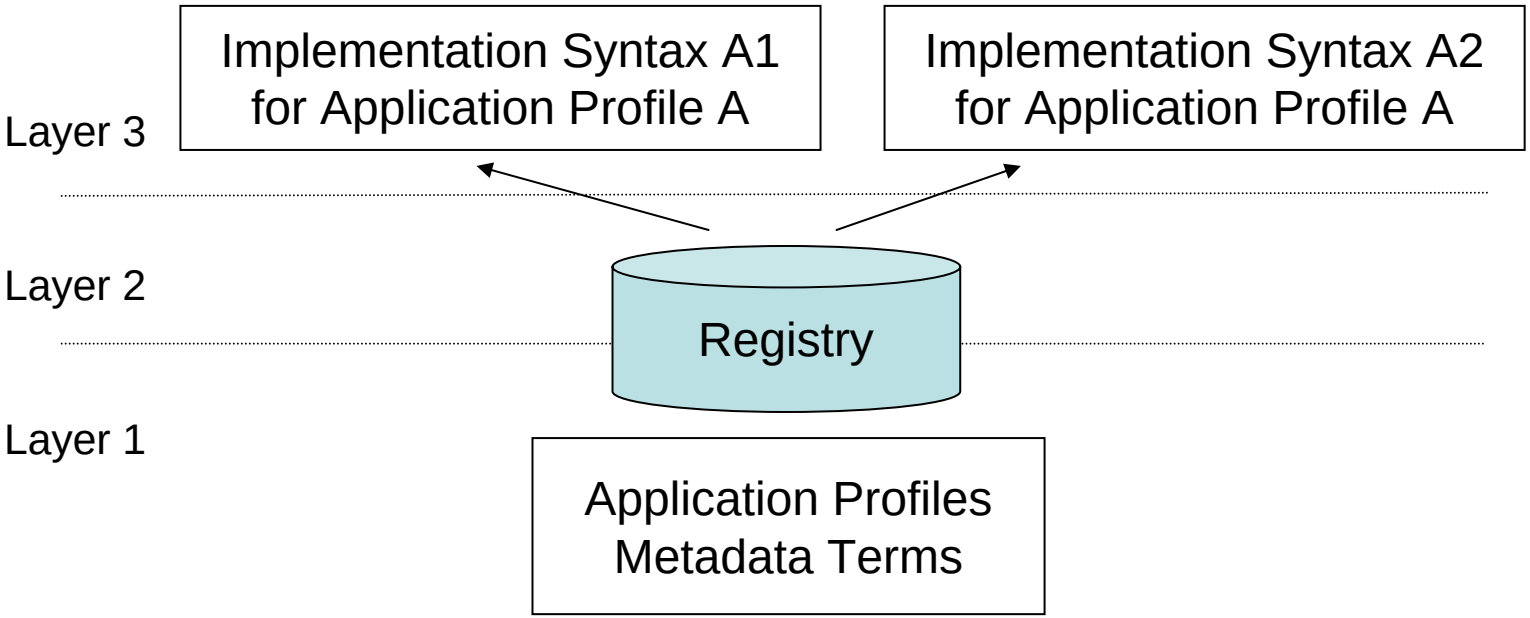
Layered Model and Metadata Schema Registry

Application profile designers get definitions of metadata terms from Registry A and/or B.



Layered Model and Metadata Schema Registry

Application designers get an application profile and define implementation syntax.



A Metadata Framework for Context Sensitive Resource Selection - an on going study -

- Find and access a resource in accordance with user characteristics and user environment
  - Users with Disabilities
    - Identification of appropriate resources in accordance with user characteristics
  - Size of Displays
    - PC, PDA, mobile phone
  - Environment
    - In-door, out-door
    - Band-width of Network Connection
  - etc.

Some Lessons Learned

- Metadata Schema Interoperability
  - A good model is required.
  - The layered model gives a framework. Refinement of the model is required.
- Metadata schema registry
  - Sharing information about metadata schemas
  - Maintenance of metadata vocabularies and application profiles
  - Software tools connected to the registry
    - e.g. Schema editor, Schema maintenance tool, Software tool generator

Challenges in Digital Libraries

- Key Issues based on the experiences in the metadata-centric projects
  - Interoperability
  - Model and Theory
  - Digital Preservation
  - Globalization vs. Localization, Universality vs. Domain-Specificity
  - Accessibility and Adaptability
  - Metadata

Challenges in Digital Libraries

- Interoperability
  - General and basic requirements
    - Single access point for a network of DLs
  - Difficulties in various aspects technologically and non-technologically
- Model and Theory
  - Guidelines for development and maintenance of DLs
  - Guidelines for cooperation between DLs

Challenges in Digital Libraries

- Digital Preservation
  - Temporal Interoperability
  - Digitized Resources and Born Digital Resources
    - Preservation Levels
      - Bit String, Functions, Look & Feel, Technologies
    - OAIS Reference Model
  - Web Archiving
    - Hidden/Deep Web problem
    - Internet Archive / Intranet Archive (Institutional Archive)
      - Negotiation between providers and archives

Challenges in Digital Libraries

- Digital Preservation (cont'd)
  - Maintenance of (Preserved) Resources
    - Migration
    - Broken Link Problem
  - Metadata for Preservation
    - Costs
    - Metadata Schema Issues
  - Resource Identification for Preservation
    - Guidelines to determine to identify resources for preservation
    - Dynamic Resources
      - "single-source multi-use" resources
      - Documents dynamically composed of fragments
    - Preservation of functional features
      - Acceptable degradation of functional features

Challenges in Digital Libraries

- Globalization vs. Localization, Universality vs. Domain-Specificity
  - Metadata Issues
    - Domain specific vocabulary vs. Common vocabulary
    - Dumb-down principle
      - A principle introduced for semantic interoperability by DCMI
      - A framework for semantic refinement of a metadata element without loosing interoperability
    - Layered Model
      - A framework to understand and enhance metadata interoperability
  - User Issue
    - User Interface issues, e.g., language, style

Challenges in Digital Libraries

- Accessibility and Adaptability
  - Guidelines for Web Accessibility
  - Resource Selection by Matching User Characteristics and Resource Characteristics
  - Resource Adaptation in accordance with Users and User Environments
- Metadata
  - The Key and Common Component among the Challenging Issues in DL

Summary

- There are many challenging issues for further development of DLs in addition to the issues listed in this presentation.
- Interoperability is the most fundamental and challenging issue for DL.
  - DLs have to cope with contradictory requirements
    - legacy data vs. new type of resources, media and users
    - Regional/Domain-specific Community vs. Global Community

      etc.
  - Collaboration among libraries is crucial to clarify the issues that should be solved and to promote technology development and knowledge sharing.