Approaches for Large Scale Digital Library

ISTITUTO DI SCIENZA E TECNOLOGIE DELL'INFORMAZIONE "A. FAEDO"

UNI BASEL

Introduction & Motivation

Tutorial at RCDL 2007
October, 15th 2007

Thomas Risse
L3S Research Center, Hannover, Germany
Email: risse@L3S.de

In cooperation with Claudia Niederee (L3S), Carlo Meghini (CNR-ISTI), Heiko Schuldt (Uni Basel)

Agenda

Information Society Technologies

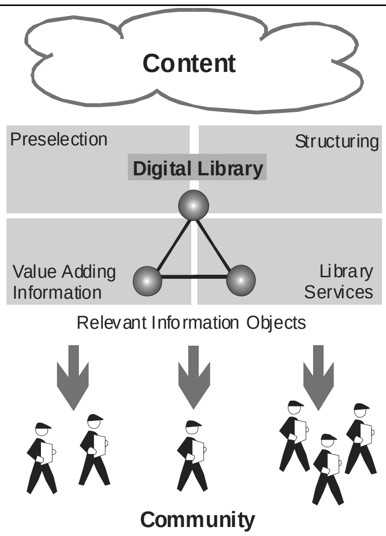
1. Introduction & Motivation
2. Challenges of bringing DL to distributed Infrastructures
3. Underlying Technologies and their promises (SOA, P2P)
4. Solutions for decentralized DL infrastructures (with BRICKS Demos)
5. Conclusions and future directions

BRICKS Building Resources for Integrated Cultural Knowledge Services

Page 2

Introduction

Digital Libraries:
Content-to-Community
Mediation

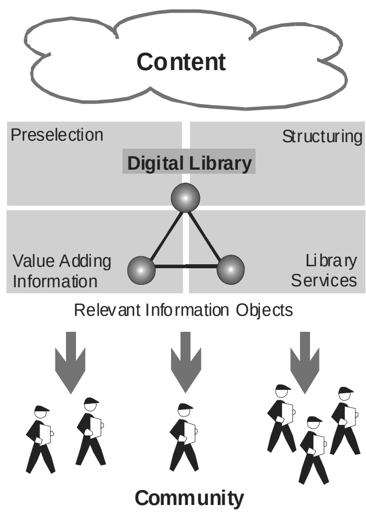


BRICKS
Building Resources for Integrated Cultural Knowledge Services

Page 3

Digital Libaries: That‘s where we are

- Core functionality of a DL is well-understood (Content-to-Community Mediation)
- Prototypical as well as operational Digital Libraries for various kinds of content have been build (Maps, Music, Historical Artifacts, …)
- Standards for metadata, search, metadata harvesting, etc. have been established and are (more or less consequently) used fostering interoperation and integration
- Underlying technologies like information retrieval, have been advanced



BRICKS
Building Resources for Integrated Cultural Knowledge Services

Page 4

Digital Libaries: That‘s where we are (2)

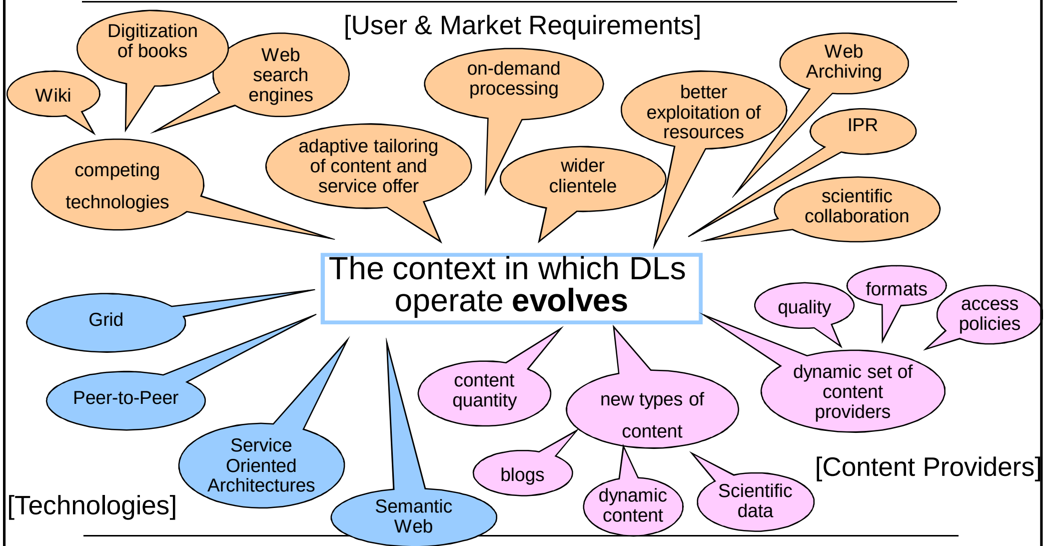
- Digital Library Management Systems in operation (e.g. Greenstone, DSpace, OpenDLib, FEDORA)

Example: DSpace

- Open source system developed by MIT Libraries and Hewlett-Packard Labs
- Repository for the digital research and educational material produced by members of a research university or organization
- Addresses the need of organizations to collect, preserve, index and distribute their material
- Is designed to make participation by depositors easy
- Information model is built around the idea of organizational "Communities“

BRICKS
Building Resources for Integrated Cultural Knowledge Services

However …



BRICKS
Building Resources for Integrated Cultural Knowledge Services

Page 6

Scenario 1 – Finds[1] identification today

- The public is an important source for understanding the past and archeological sites
- Thousands of objects are discovered by people during walking, traveling, gardening, etc.
- One day-to-day job of Archeological Museum Curators is to identify these objects
- Traditional way of working
  - Examining objects
  - Preliminary identification
  - Comparison to specialist reference collections, e.g. roman coins
    - Sometime object descriptions are enough
    - Sometime it is the starting point for a scientific analysis process
  - Comparison to objects in the museum collection
  - Record object in detail

[1] With *Finds* we mean *archeological finds*

BRICKS
Building Resources for Integrated Cultural Knowledge Services

Page 7

Scenario – Finds identification today

Disadvantage of the traditional way

- The reference collections have to be known by the curator
- Curator needs information about new collections, e.g. when a museum opens its archive
- Curators need to access them one by one
- Collections are heterogeneous from point of view of data, search facilities, user interface, languages, etc.

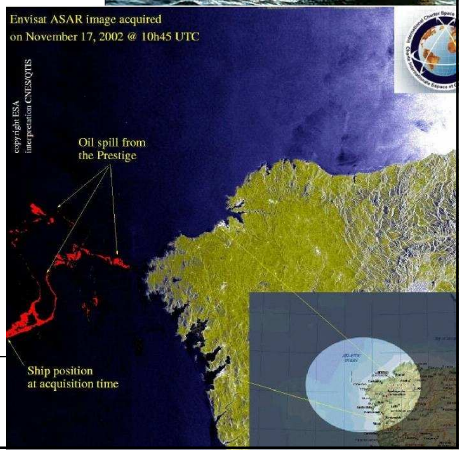
As a consequence curators invest a lot of time in the identification process

BRICKS
Building Resources for Integrated Cultural Knowledge Services

Page 8

Scenario 2 – Environmental Incidents (ImpECt[1])

- ESA ENVISAT satellite monitored an oil spill caused by a tanker off the Galician coast
- An expert group from different domains has to be coordinated and work on the problem (e.g. oil-fighting bodies, ministries of the involved countries, …)
- different resources tailored to situation are required (relevant ecological and environmental data, historical dossiers on tankers accidents;
- On demand simulations on oil distributions have to be computed based on weather forecast, wind, temperature, currents, …
- Postprocessing: reports have to be generated, environmental effects have to be analysed



BRICKS Building Resources for Integrated Cultural Knowledge Services

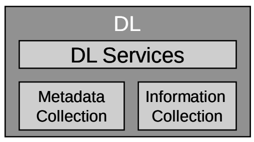
[1]ImpECt= Implementation of Environmental Conventions

Next Generation Digital Libraries (NGDL)

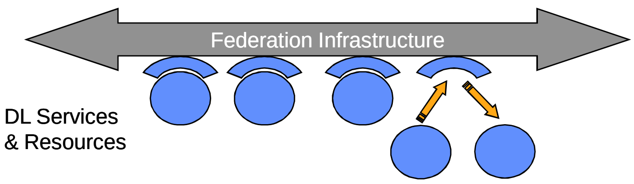
Information Society Technologies

Current plans for next generation DL architectures are aiming for a **transition**

- from the DL as an integrated, centrally controlled system



- to a **dynamic configurable federation** of DL services and information collections for enabling large-scale DLs.



BRICKS Building Resources for Integrated Cultural Knowledge Services

Page 10

In the Tutorial …

Closer look

- on these challenges and the resulting requirements
- Selected solutions for these challenges developed in the European Project BRICKS
- technologies for implementing next generation digital libraries (SOA, Peer-to-Peer)
- lessons learned from implementing large-scale digital libraries on top of distributed infrastructures
- small demos (if time is available)

BRICKS
Building Resources for Integrated Cultural Knowledge Services

Page 11

Approaches for Large Scale Digital Library

Challenges of bringing DL to distributed Infrastructures

Tutorial at RCDL 2007
October, 15th 2007

Thomas Risse
L3S Research Center, Hannover, Germany
Email: risse@L3S.de

In cooperation with Claudia Niederee (L3S), Carlo Meghini (CNR-ISTI), Heiko Schuldt (Uni Basel)

Forschungszentrum · Research Center L3S

ISTI
ISTITUTO DI SCIENZA E TECNOLOGIE DELL'INFORMAZIONE "A. FAEDO"

UNI BASEL

Agenda

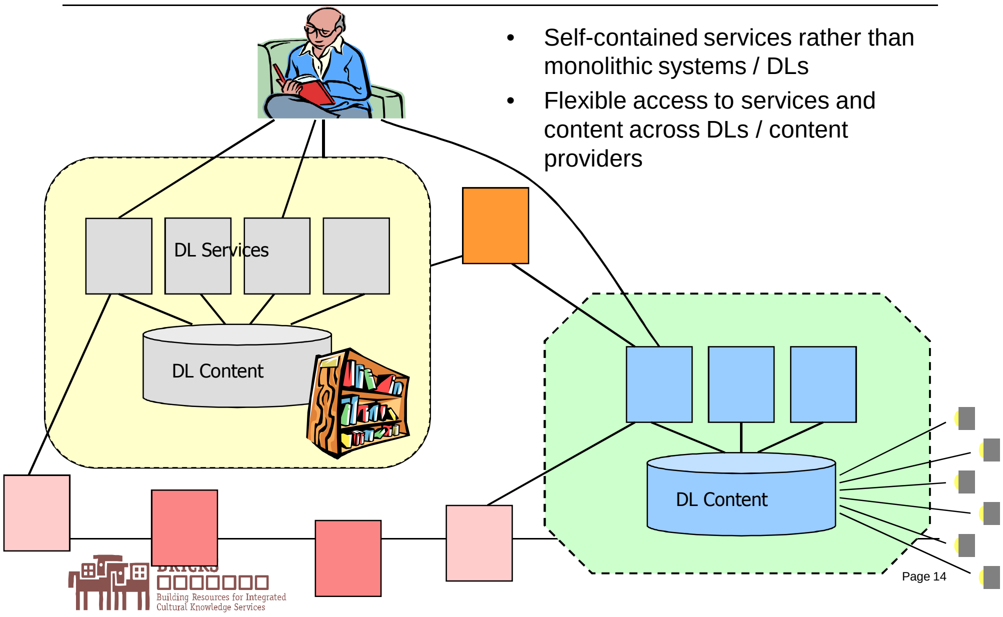
1. Introduction & Motivation
2. Challenges of bringing DL to distributed Infrastructures
3. Underlying Technologies and their promises (SOA, P2P)
4. Solutions for decentralized DL infrastructures (with BRICKS Demos)
5. Conclusions and future directions

BRICKS
Building Resources for Integrated Cultural Knowledge Services

Page 13

Digital Libraries – The Future

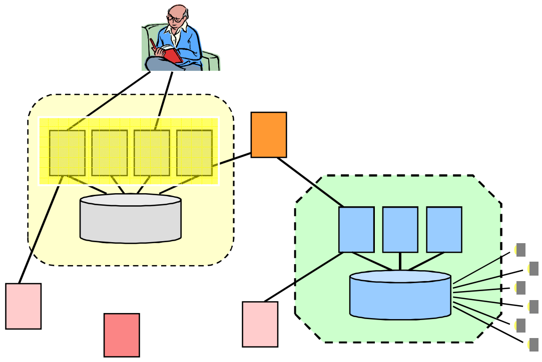
- Self-contained services rather than monolithic systems / DLs
- Flexible access to services and content across DLs / content providers



BRICKS
Building Resources for Integrated Cultural Knowledge Services

Page 14

Requirements …

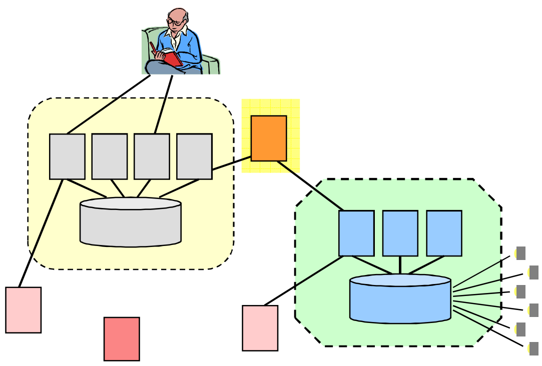


- Specialized services, **local** to a content provider
  - Search
    - Different media types
    - Content-based
    - Multi-object, multi-feature
    - Multilingual Access
    - Relevance feedback
    - …
  - Indexing
  - Annotation
  - Metadata management
  - Content management
  - Resource Management
  - …

BRICKS
Building Resources for Integrated Cultural Knowledge Services

Page 15

… Requirements …

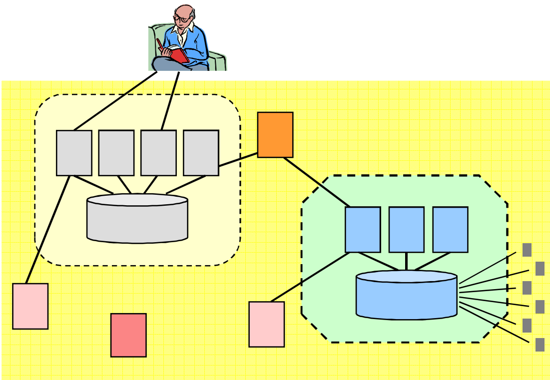


- Specialized services, **across** different providers
  - Search
    - Different media types
    - Content-based
    - Multi-object, multi-feature
    - Multilingual Access
    - Relevance feedback
    - …
  - Metadata management
  - Content management
  - Resource Management
  - Indexing
- without central control (no censorship)

BRICKS
Building Resources for Integrated Cultural Knowledge Services

Page 16

… Requirements …



Virtual DL (virtual collection)

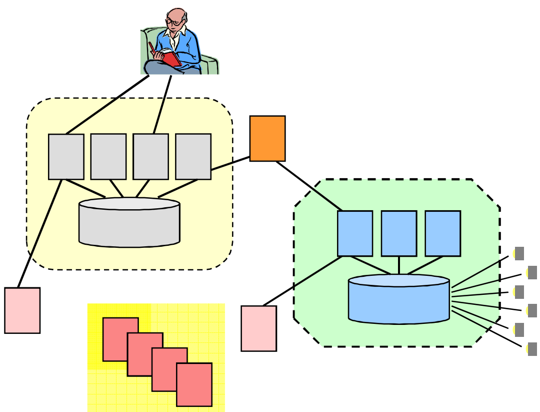
- Easy to extend by integrating new content providers, services, etc.
- Example: collaboration in eScience applications

Management of services which are

- Distributed
- Heterogeneous
- Autonomous

Page 17

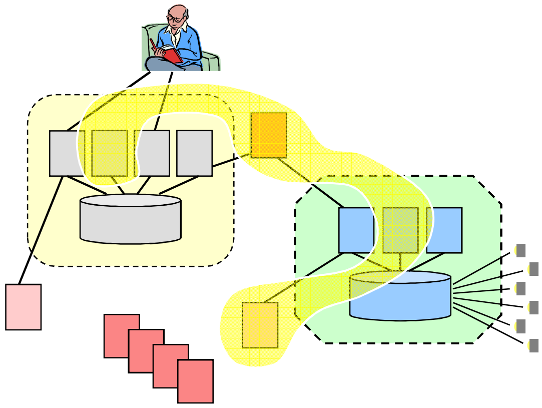
… Requirements …



- Management of computationally intensive services
  - Scale-out
  - Load balancing
- Example: services for extracting characteristic features from image/audio/video documents

Page 18

… Requirements …

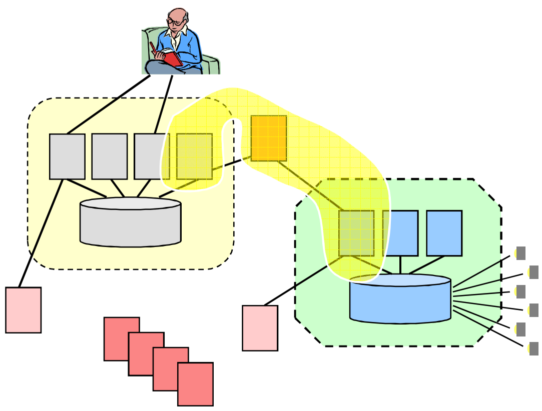


Composition of services

- Defining complex services / processes / workflows on the basis of existing services
- Flexibility: automatic adaptation
- Example: complex processes for automated storage and replication of data, generation of meta data (content features), …

BRICKS
Building Resources for Integrated Cultural Knowledge Services

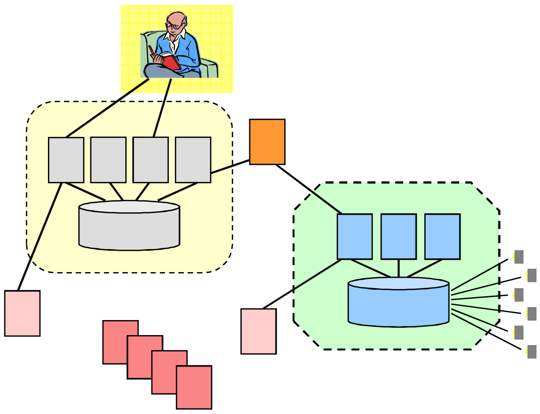
… Requirements …



- Notification of changes
- Guaranteed consistency of derived data
- Example: automated updates of complex reports

BRICKS
Building Resources for Integrated Cultural Knowledge Services

… Requirements …

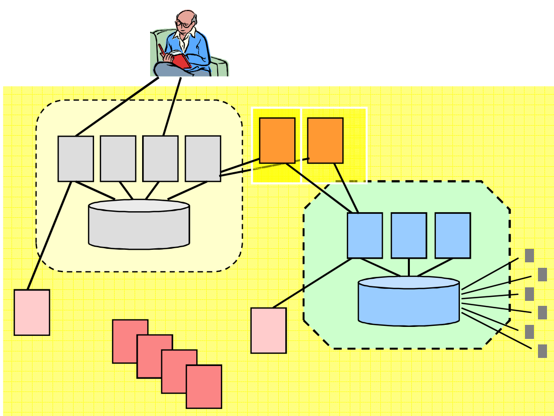


- Personalization
- Visualization
- Access anywhere
  - Mobile devices
- Context- and location-aware services
- Authentication & authorization

BRICKS
Building Resources for Integrated Cultural Knowledge Services

Page 21

… Requirements …

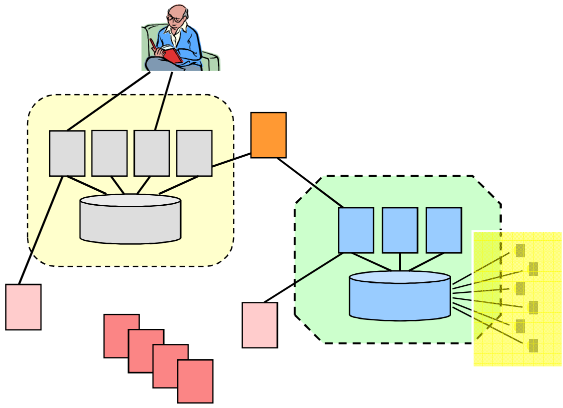


- High degree of availability: Access anytime
  - Replication
- High degree of dependability / reliability
  - Systems their users can count on
- High degree of scalability

BRICKS
Building Resources for Integrated Cultural Knowledge Services

Page 22

… Requirements



Continuously generated data

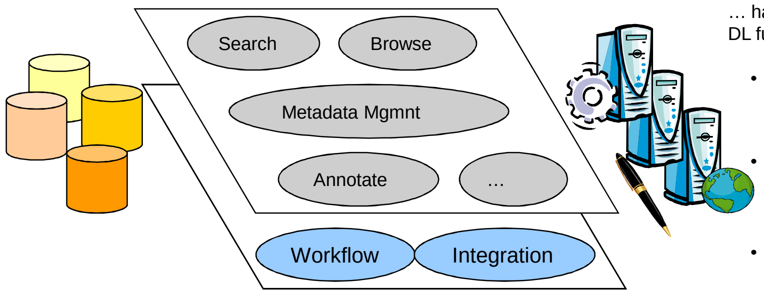
- Sensor networks
- Sensor data streams (hardware / software sensors)
- Example: data created by certain instruments in eScience

BRICKS
Building Resources for Integrated Cultural Knowledge Services

Page 23

DL Architecture – an Abstract View

DL users …



… have access to DL functionality that …

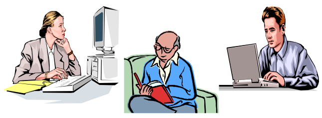
- operates on distributed content
- is implemented by distributed services
- is supported by system services

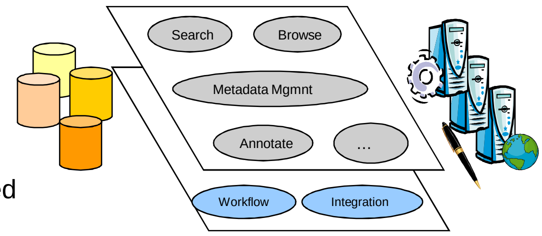
BRICKS
Building Resources for Integrated Cultural Knowledge Services

Page 24

Summary

- The challenges and requirements of new DLs and their applications cannot be met by current DL Systems





- New technologies are needed that allow to bring DLs to **Distributed Infrastructures**

Questions & Discussions

Approaches for Large Scale Digital Library

ISTITUTO DI SCIENZA E TECNOLOGIE DELL'INFORMAZIONE "A. FAEDO"

Underlying Technologies

Tutorial at RCDL 2007
October, 15th 2007

Thomas Risse
L3S Research Center, Hannover, Germany
Email: risse@L3S.de

In cooperation with Claudia Niederee (L3S), Carlo Meghini (CNR-ISTI), Heiko Schuldt (Uni Basel)

UNI BASEL

Agenda

1. Introduction & Motivation
2. Challenges of bringing DL to distributed Infrastructures
3. Underlying Technologies and their promises (SOA, P2P)
4. Solutions for decentralized DL infrastructures (with BRICKS Demos)
5. Conclusions and future directions

BRICKS Building Resources for Integrated Cultural Knowledge Services

Page 28

Overview

- Service-oriented Architectures
  - Definition
  - Service Model
  - Web Service Stack
  - Example: Supply Chain
- Peer to Peer Infrastructures
  - The nature of peer-to-peer systems
  - Some history
  - Application domains
  - The Data access challenge
- Summary

BRICKS Building Resources for Integrated Cultural Knowledge Services

Page 29

Service Oriented Architectures (SOA)

BRICKS Building Resources for Integrated Cultural Knowledge Services

Page 30

Definitions

Gartner Group (technical definition)

- Web Services are loosely coupled software components that interact with one another dynamically via standard Internet technologies.

Forrester Research (business definition)

- Web Services are automated connections between people, systems and applications that expose elements of business functionality as a software service and create new business value.

BRICKS
Building Resources for Integrated Cultural Knowledge Services

Page 31

Service Oriented Computing

Goals of Service Oriented Computing

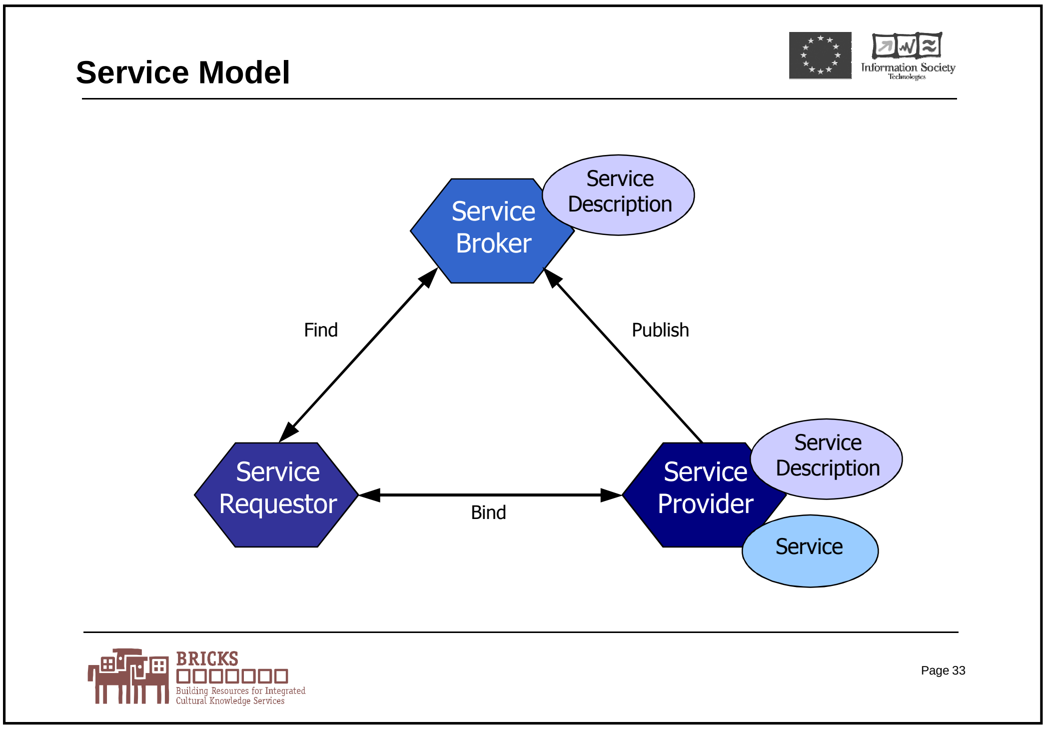
- Distributed interoperable Systems
- Cost reduction due to reuse of services
- Flexibility
- Easy adaptation to future developments

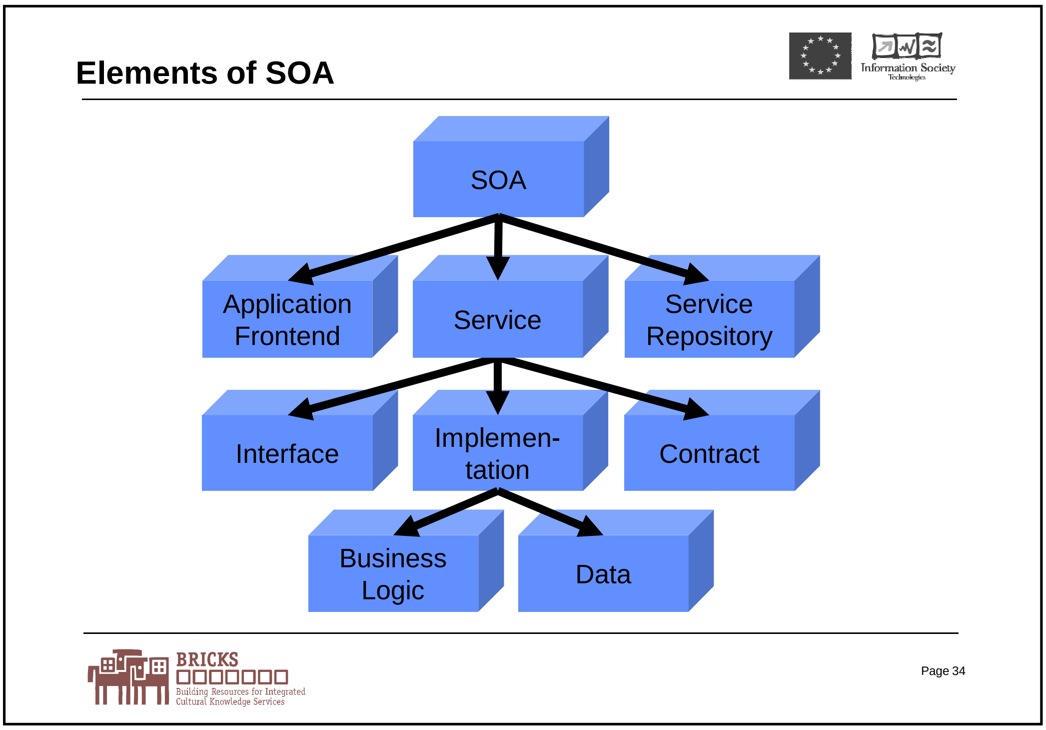
Properties of services

- Self describing
- Accessed using a well-defined interface
- Autonomous
- Loosely coupled
- Independent from the Platform / Operating System
- Independent from the Programming language

BRICKS
Building Resources for Integrated Cultural Knowledge Services

Page 32


Service Model
Information Society Technologies
Service Broker
Service Description
Find
Publish
Service Requestor
Bind
Service Provider
Service Description
Service
BRICKS
Building Resources for Integrated Cultural Knowledge Services
Page 33


Elements of SOA
Information Society Technologies
SOA
Application Frontend
Service
Service Repository
Interface
Implemen-tation
Contract
Business Logic
Data
BRICKS
Building Resources for Integrated Cultural Knowledge Services
Page 34

Web Services Stack

Contracting (Trading Partner Agreement)	
Management (WSDM)	Administration
Security (WS-Security, WS-Trust, …)	Security
Coordination (WS-Coordination,WS-AtomicTransaction)	Synchronization of processes
Messaging (WS-Notification, …)	
Workflow (**BPEL4WS**, WS-CDL)	Process specification
Publish, discovery (**UDDI**, WSIL)	Searching & finding services
Description (**WSDL**), Policy (WS-Policy)	Interface definition
Interaction (**SOAP**)	Protocol for distribution
Syntax (**XML, XML Schema, XML-Namespace**)	Model for data integration
Wire (**HTTP**)	Transportation

BRICKS
Building Resources for Integrated Cultural Knowledge Services

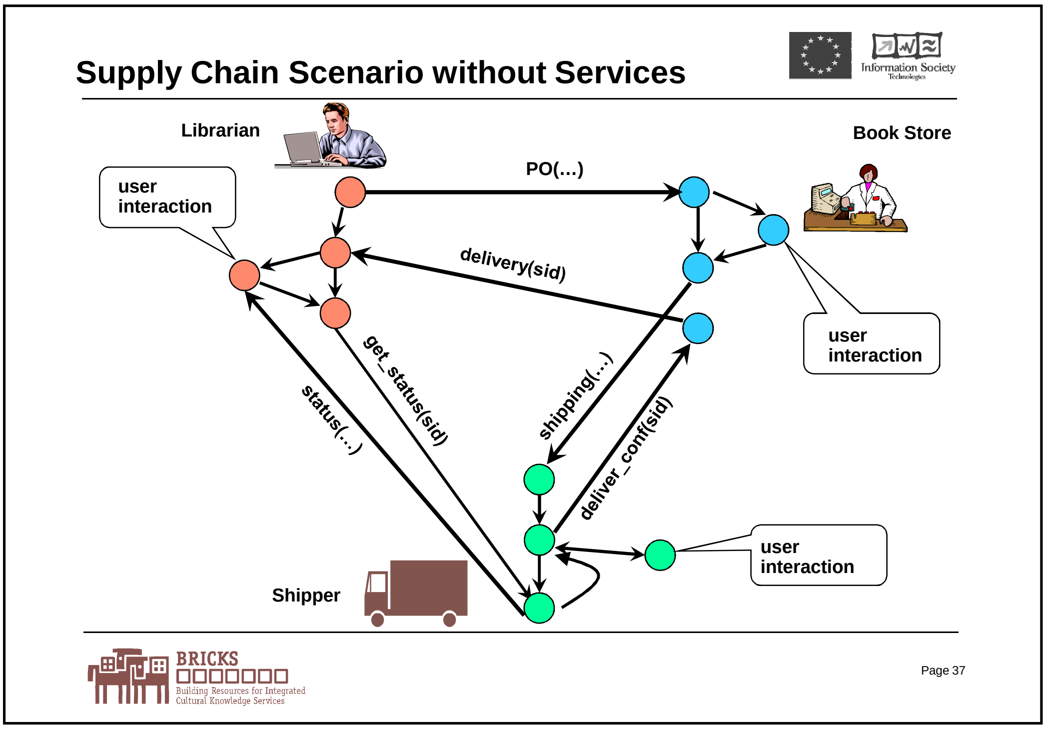
Page 35

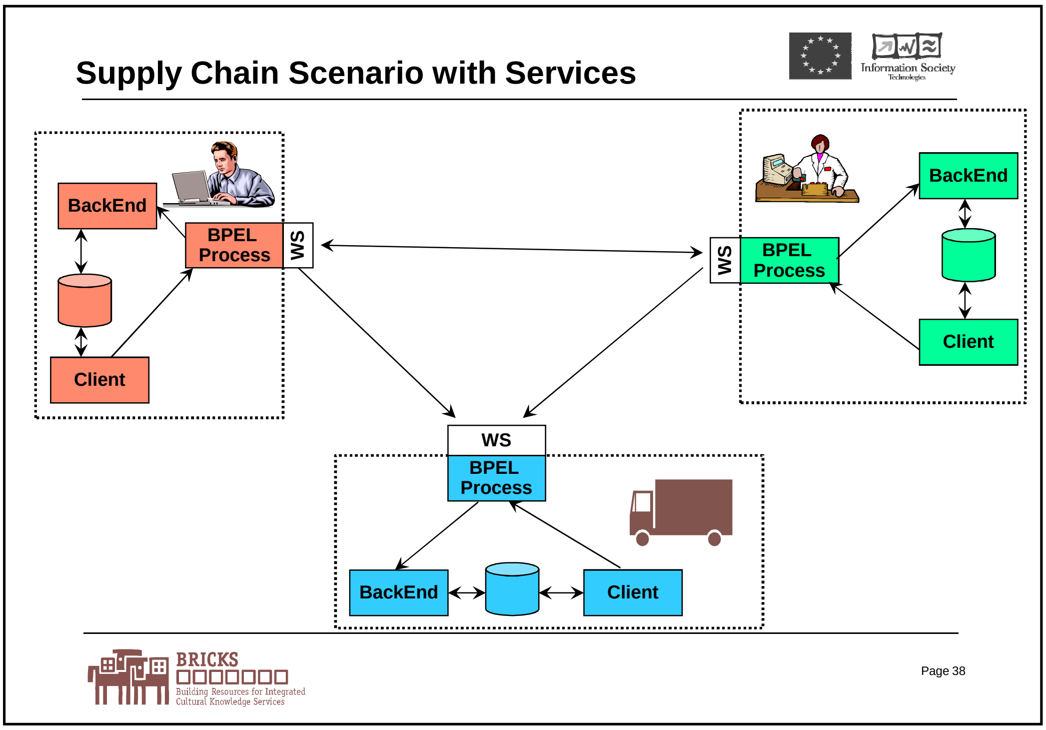
Relevant Standards Overview

XML	A universal model for data exchange and data integration
XML Schema	Defines the schema of a XML document, makes syntactical restrictions, defines structural patterns, defines data types
XML Namespace	Provide means to avoid naming conflicts in XML documents
SOAP	Simple Object Access Protocol - A universal communication protocol
WSDL	WS Description Language - Description of the WS interfaces, parameters, etc.
WS-Policy	General-purpose model to describe the policies of a Web service
UDDI	Universal Description, Discovery and Integration A standard for publication and discovery of information
BPEL4WS	Business Process Execution Language Specification of workflow for the composition of services
WS-Notification	Defines the publish/subscribe pattern for message oriented systems
WS-Coordination	Coordination of distributed actions. Includes transaction management.
WS-Security	Secure SOAP messages
WS-Trust	Management of trust relationships
WSDM	Distributed Management of Web Services

BRICKS
Building Resources for Integrated Cultural Knowledge Services

Page 36


Supply Chain Scenario without Services
Information Society Technologies
Librarian
Book Store
PO(...)
user interaction
delivery(sid)
get_status(sid)
status(...)
shipping(...)
deliver_conf(sid)
user interaction
user interaction
Shipper
BRICKS
Building Resources for Integrated Cultural Knowledge Services
Page 37


Supply Chain Scenario with Services
Information Society Technologies
BackEnd
BPEL Process
WS
Client
WS
BPEL Process
BackEnd
Client
WS
BPEL Process
BackEnd
Client
BRICKS
Building Resources for Integrated Cultural Knowledge Services
Page 38

Advantages and Disadvantages

Advantages
- Clear Description of Services and Interfaces
- Transparent access to Legacy Systems
- Higher Flexibility and Dynamics
- Widely accepted Web Service standards
- Various software is available

Disadvantages
- Semantic standards are still under development
- Several complementing standards are in development
- No Resource Sharing
- Depends partly on central management services

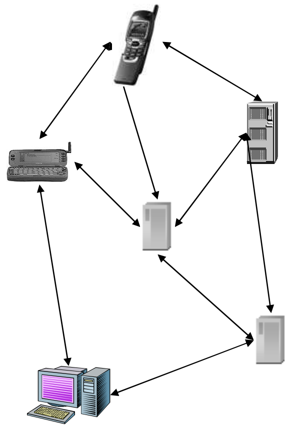
Page 39

Peer-to-Peer Computing

Page 40

The Nature of P2P

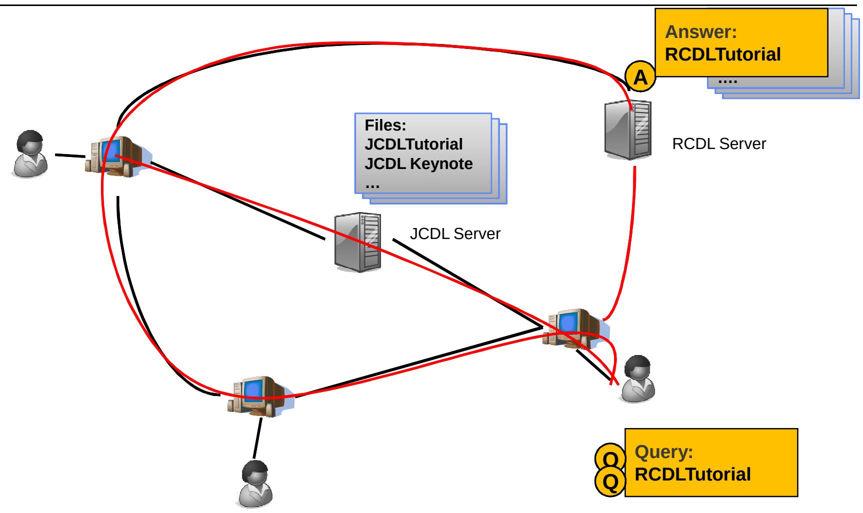
- Peers act as client and server
  - Peers communicate directly
  - Both provide services, e.g. storage, calculations
- Peers are autonomous
  - Owner decide e.g. when a service is active
- No central administration
- No global system view
  - Peers have only some knowledge about their neighborhood
- Highly dynamic
  - Peers appear and disappear whenever they like
- Unreliable connections
- Mostly self organizing
  - Services and clients are actively looking for appropriate partners



BRICKS
Building Resources for Integrated Cultural Knowledge Services

Page 41

Simple P2P File Sharing Example



BRICKS
Building Resources for Integrated Cultural Knowledge Services

Page 42

Some History

P2P receives a lot of attention ...
- File sharing systems
  - Gnutella, Freenet, Napster, ...

... but P2P is nothing new
- Telephone systems
- Federated databases
- ARPAnet
  - First networks were P2P systems
- Usenet, Domain Name Service
  - Servers communicate in a P2P manner
- Service Oriented Architectures

gnutella.com

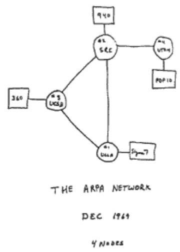


FIGURE 6.2 Drawing of 4 Node Network (Courtesy of Alex McKenzie)

BRICKS
Building Resources for Integrated Cultural Knowledge Services

Page 43

Advantages of Peer-to-Peer

- Scalability
  - P2P solutions should be scalable by definition
- Availability of information, services, …
  - Redundancy increases the availability
- Flexibility
  - Flexible reaction on changes, e.g. leaving of peers, changing from wire to wireless communication
- Low administration costs
  - Self organization requires no central administration

BRICKS
Building Resources for Integrated Cultural Knowledge Services

Page 44

Disadvantages

- Advantages grow with the number of peers
  - To accomplish the advantages many peers are necessary
- Software has to deal with P2P properties
  - Some P2P properties (e.g. dynamicity) are challenging for the development
  - Software components can provide functions to other peers as a service (e.g. Microsoft .Net)

BRICKS
Building Resources for Integrated Cultural Knowledge Services

Page 45

Application Domains

- Resource sharing
  - Joined usage of unutilized resources (e.g. CPU, storage)
- Content sharing
  - Sharing of files (music, video), calendar, spam sender data, …
  - News distribution
  - Searching for information
  - Examples: Gnutella, Freenet, Napster, Groove, Hive, Cloudmark SpamNet
- Collaboration
  - Peers work together
    - Collaboration within the project team, e.g. document creation
    - Instant Messaging
  - Benefits
    - Autonomous
    - Offline working
    - No administration
  - Examples: Groove (Collaboration), Skype (Messaging), …

BRICKS
Building Resources for Integrated Cultural Knowledge Services

Page 46

Challenge: Data Access

Data Access Structures
- Methods to query and access „nomadic data"
- Partial knowledge

Simple approaches
- Breadth-first-search (Gnutella)
- Depth-first-search (Freenet)

Disadvantage
- High accumulated bandwidth

More sophisticated approaches
- Use global unique identifier (GUID) of object for routing, (DHT, Plaxton Routing, Chord, Tapestry, Pastry, CAN, P-Grid)

BRICKS
Building Resources for Integrated Cultural Knowledge Services

Page 47

Summary

Web Services
- Well established standard for interoperabel services
- Lightweight communication protocol
- Easy to use
- Depends partly on central management services
  → Fully decentralized Web Services are developed in BRICKS

Peer-to-Peer
- Describes the way of communication between entities
- Self-organizing system of autonomous entities
- Highly flexible and scalable

BRICKS
Building Resources for Integrated Cultural Knowledge Services

Page 48

Questions & Discussions

BRICKS
Building Resources for Integrated Cultural Knowledge Services

Page 49

Approaches for Large Scale Digital Library

Forschungszentrum · Research Center L3S

ISTITUTO DI SCIENZA E TECNOLOGIE DELL'INFORMAZIONE "A. FAEDO"

UNI BASEL

Solutions for decentralized DL infrastructures: issues, solutions and advantages in the framework of BRICKS

Tutorial at RCDL 2007
October, 15th 2007

Thomas Risse
L3S Research Center, Hannover, Germany
Email: risse@L3S.de

In cooperation with Claudia Niederee (L3S), Carlo Meghini (CNR-ISTI), Heiko Schuldt (Uni Basel)

Agenda

1. Introduction & Motivation
2. Challenges of bringing DL to distributed Infrastructures
3. Underlying Technologies and their promises (SOA, P2P)
4. Solutions for decentralized DL infrastructures (with BRICKS Demos)
5. Conclusions and future directions

BRICKS
Building Resources for Integrated Cultural Knowledge Services

Page 51

The BRICKS Project
http://www.brickscommunity.org/

BRICKS
Building Resources for Integrated Cultural Knowledge Services

Page 52

BRICKS - Project Identity Card

- Project Acronym: BRICKS - Building Resources for Integrated Cultural Knowledge Services
- Consortium: 24 organizations from 9 countries
- Thematic area: Digital Libraries Services
- Duration: 42 months
- Started in January 2004
- Budget: 12,2 Mill. Euro
- Homepage: http://www.brickscommunity.org/
- Technical Partners: Engineering, Italy; Fraunhofer IPSI; CNR-ISTI, Italy; Metaware, Italy; ARC Seibersdorf Research GmbH, Austria; EPFL, Switzerland; Canoo, Switzerland; Uni. of Athens, Greece; Uni. of Sheffield, UK; Scuola Normale Superiore di Pisa, Italy; Uni. of Florence, Italy; Oxford ArchDigital, UK; PolyDisplay, Norway
- User Partners: Italian Ministry of Culture, Italy; Vatican Secret Archives; Uffizi Gallery, Italy; Schönbrunn Palace, Austria; Austrian National Library, Austria; European Museum Forum, UK; Museum of Cycladic Art, Greece; Russian Cultural Heritage Network, Russia; Museums, Libraries and Archives Council, UK; Studio Azzurro, Italy

BRICKS
Building Resources for Integrated Cultural Knowledge Services

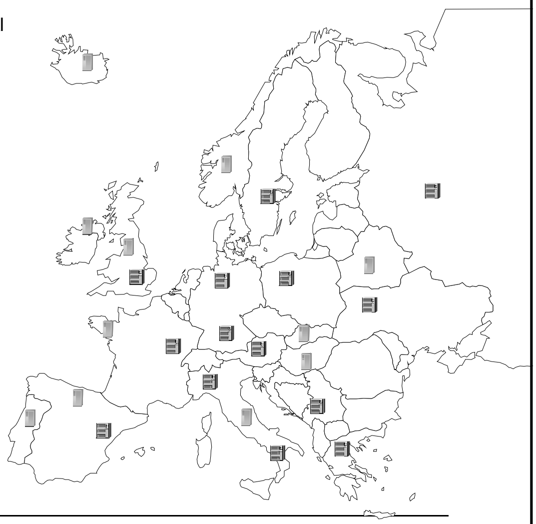
Page 53

Access to professional cultural resources today

Situation
- Large number of independent distributed digital information sources
- Often professional sources are not public
- Still a large number of interesting sources are missing, e.g. small museums

Find and access information sources (examples)
- Search engines, e.g. CiteSeer, DBLP, Google
- Z39.50 for some libraries
- Personal Knowledge
- …

Disadvantages
- Time consuming
- Limited openness
- Seldom collaboration support
- Seldom language support

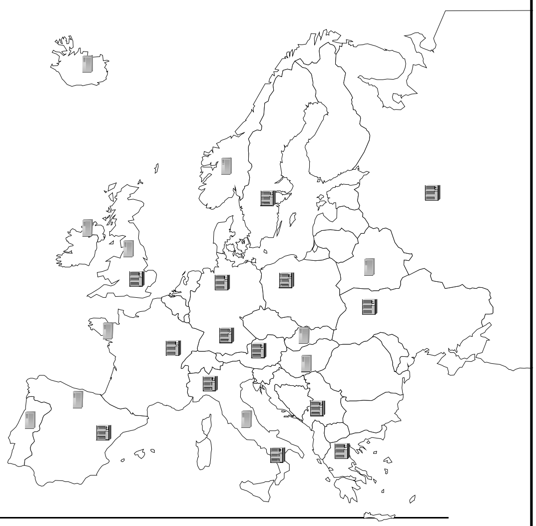


BRICKS
Building Resources for Integrated Cultural Knowledge Services

Page 54

Overall Goals of BRICKS

- Transparent access to distributed available information sources
- Retrieval of information with knowledge support
- Multi-lingual
- Protection of intellectual properties
- Low Cost
- Easy installation and maintenance
- Platform-independent



BRICKS
Building Resources for Integrated Cultural Knowledge Services

Page 55

BRICKS Approach

Service oriented

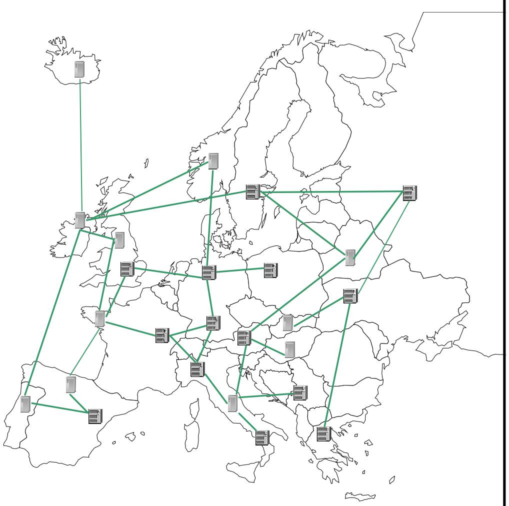
- Standardized interface descriptions based on Web Services
- Platform independent
- Flexible composition of services

Decentralized Organization

- Avoid central coordination
  → No Single Point of Failure
  → No central maintenance organization necessary
- Highly scalable
- Increased reliability
- Minimized maintenance cost

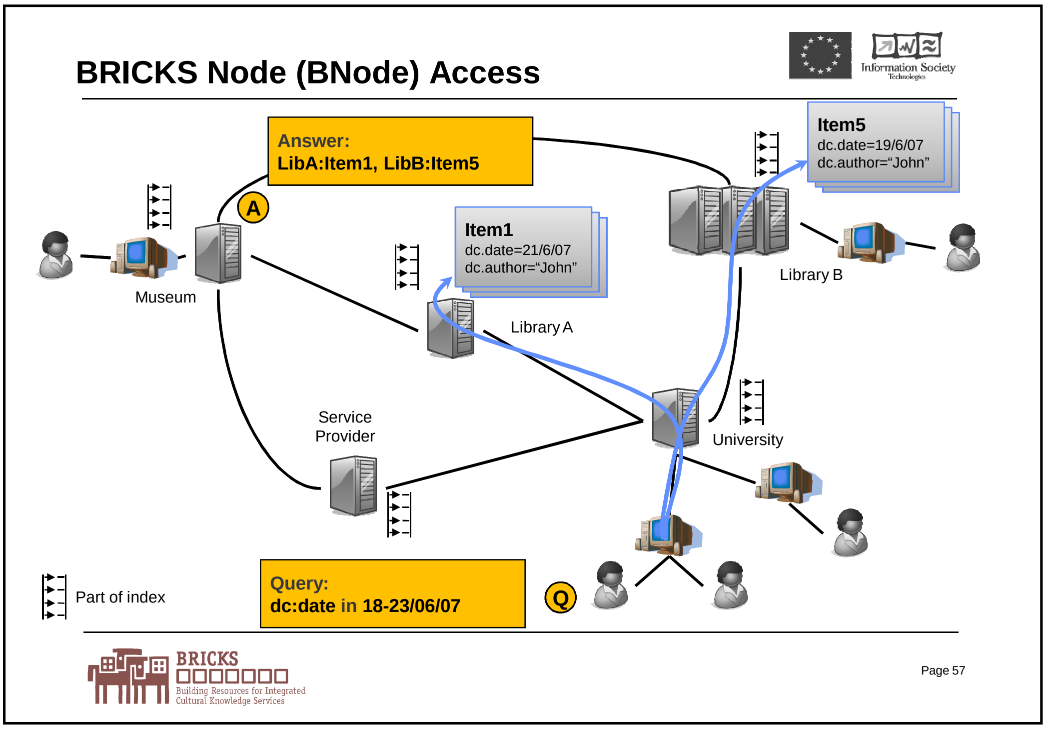
Usage & Costs

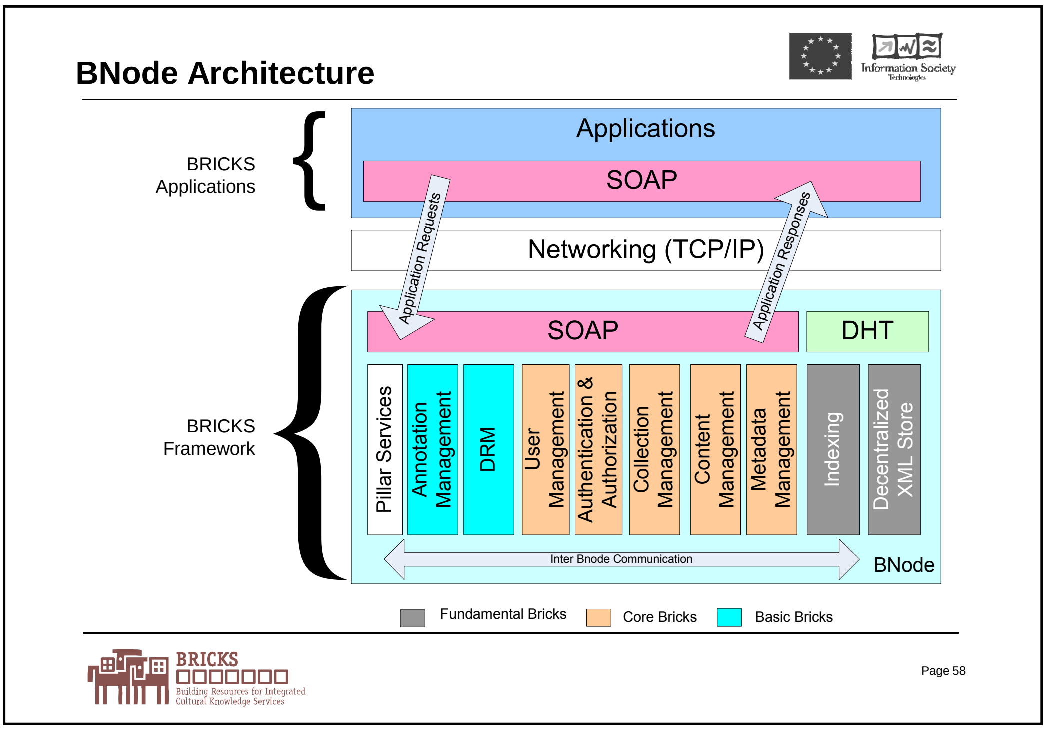
- Easy installation
- Open Source lowers the barrier to use BRICKS
- Allows tailoring to user needs



BRICKS
Building Resources for Integrated Cultural Knowledge Services

Page 56


BRICKS Node (BNode) Access
Information Society Technologies
Answer:
LibA:Item1, LibB:Item5
A
Item5
dc.date=19/6/07
dc.author="John"
Item1
dc.date=21/6/07
dc.author="John"
Museum
Library A
Library B
Service Provider
University
Part of index
Query:
dc:date in 18-23/06/07
Q
BRICKS
Building Resources for Integrated Cultural Knowledge Services
Page 57


BNode Architecture
Information Society Technologies
BRICKS Applications
Applications
SOAP
Application Requests
Application Responses
Networking (TCP/IP)
BRICKS Framework
SOAP
DHT
Pillar Services
Annotation Management
DRM
User Management
Authentication & Authorization
Collection Management
Content Management
Metadata Management
Indexing
Decentralized XML Store
Inter Bnode Communication
BNode
Fundamental Bricks
Core Bricks
Basic Bricks
BRICKS
Building Resources for Integrated Cultural Knowledge Services
Page 58

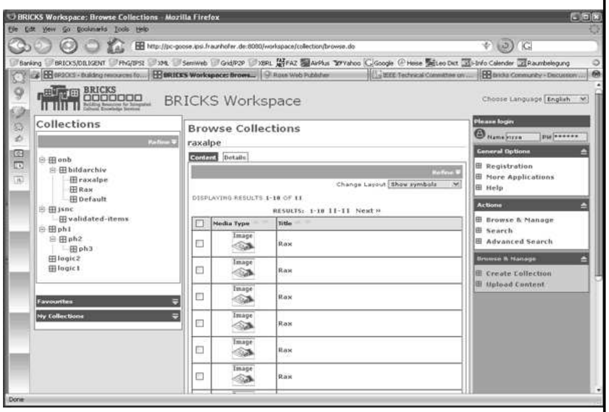
BRICKS Workspace Demo

BRICKS
Building Resources for Integrated Cultural Knowledge Services

Page 59

The BRICKS Workspace

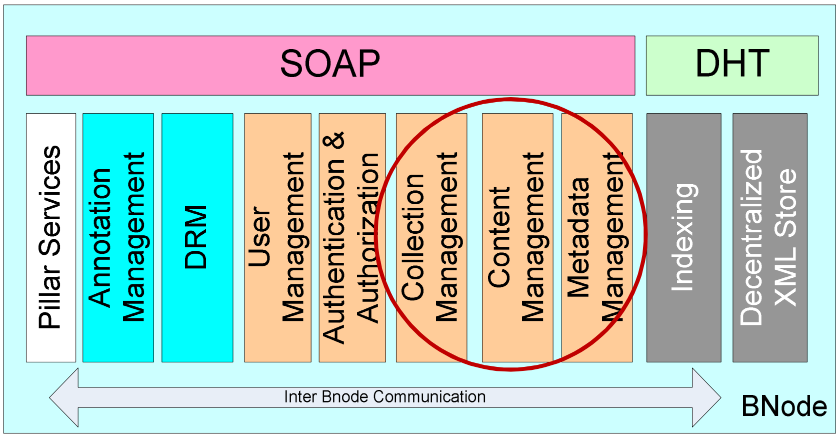
- What does it demonstrate?
  - A web application (thin client) accessing BRICKS Foundation services
  - Navigation through the BRICKS information space
- Target audience
  - General end-users
  - Application developers (as an example)
- Alternatives
  - BRICKS Desktop
    An ECLIPSE based native user interface
  - Domain & Application specific interfaces



BRICKS
Building Resources for Integrated Cultural Knowledge Services

Page 60

Some more BRICKS Framework Details



Fundamental Bricks Core Bricks Basic Bricks

BRICKS
Building Resources for Integrated Cultural Knowledge Services

Page 61

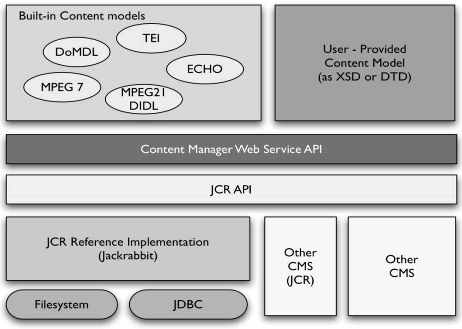
BRICKS
Content Management
Brief Overview

BRICKS
Building Resources for Integrated Cultural Knowledge Services

Page 62

BRICKS Content Management

- BRICKS is based on the Java Content Repository
- Content Repository for Java™ Technology API (JSR-170)
  - Born from the need to standardize the proprietary content repositories
  - Supports a wide range of applications
  - Provides a unified API
    - Creation and access
    - Versioning
    - Access control
    - Full text searching



  - API is independent from the Back-End storage systems, e.g. file system, WebDAV repository, XML database, SQL database
  - JCR based systems: Apache-Jackrabbit, Magnolia, Alfresco, …
- BRICKS extends JCR with a Web Service interface
- Provide some meta-content models, e.g. DoMDL Model, MPEG-21, MPEG-7, TEI-Lite, …

BRICKS Building Resources for Integrated Cultural Knowledge Services

Page 63

BRICKS Metadata Management

BRICKS Building Resources for Integrated Cultural Knowledge Services

Page 64

Metadata

A lot of definitions → Always a good reason for long discussions

Professional Users (e.g. Librarians) have a concrete view

Metadata are value-added information that professional users create to arrange, describe, track and access information objects

Different Types of Metadata

→

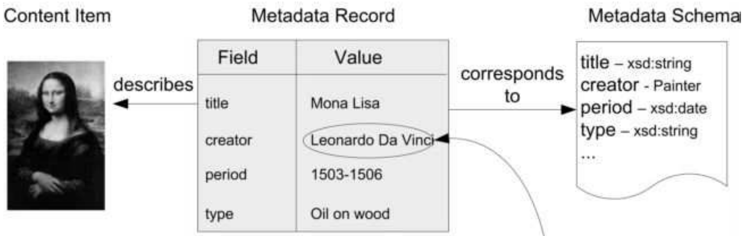
Minerva classification

Typ	Definition
Administrative	Metadata used in managing and administering information resources
Descriptive	*Metadata used to describe or identify information resources*
Preservation	Metadata related to the preservation management of information resources
Technical	Metadata related to how a system functions or metadata behave
Use	Metadata related to the level and type of use of information resources

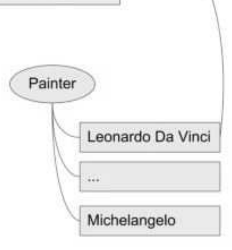
BRICKS Building Resources for Integrated Cultural Knowledge Services

Page 65

Descriptive Metadata



- Several standards exists
  - Dublin Core
  - MARC (MAchine-Readable Cataloging format)
  - CIDOC-CRM (Conceptual Reference Model for CH information)
  - PRISM (Publishing Requirements for Industry Standard Metadata)



Controlled Vocabulary

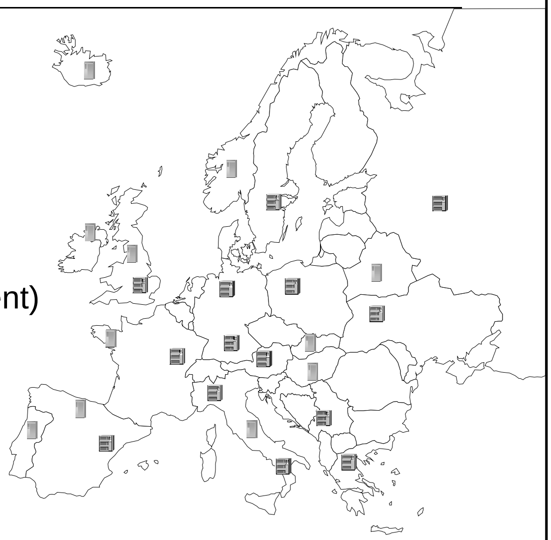
BRICKS Building Resources for Integrated Cultural Knowledge Services

Page 66

Metadata Storage Requirements

Information Society Technologies

- Durability
- Availability / Accessability
- Scalability
- Consistency
  (des. metadata ←→ content)



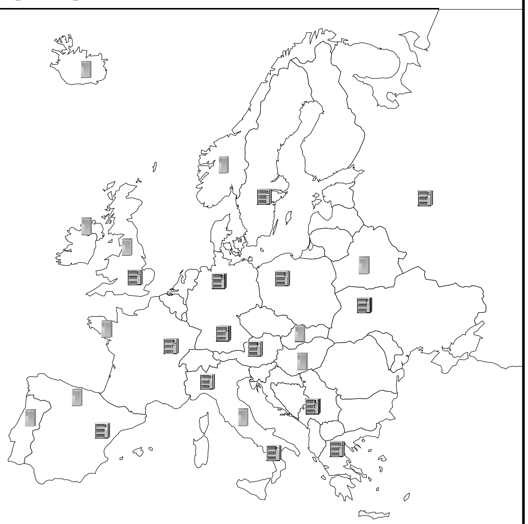
BRICKS
Building Resources for Integrated Cultural Knowledge Services

Page 67

Metadata Storage Modes (1/2)

Information Society Technologies

- Local Storage
  - Advantage: simple, well-known, …
  - Disadvantage: no guarantee of 100% availability

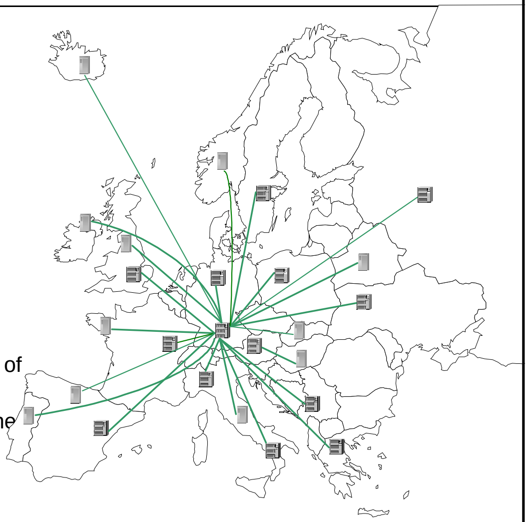


BRICKS
Building Resources for Integrated Cultural Knowledge Services

Page 68

Metadata Storage Modes (1/2)

- Local Storage
  - Advantage: simple, well-known, …
  - Disadvantage: no guarantee of 100% availability
- Central Storage
  - Advantage
    - Well-known
    - Easy to implement transaction guarantees
  - Disadvantage
    - Scalability problems for large volume of data and requests
    - High concentration of resources at one place (bandwidth, space, CPU)
    - High costs
    - Central point of failure



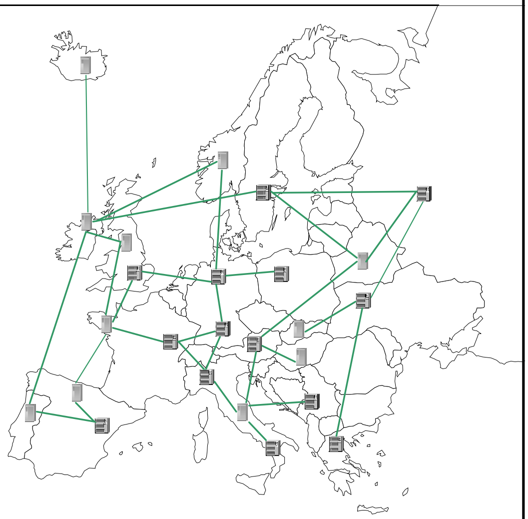
BRICKS
Building Resources for Integrated Cultural Knowledge Services

Page 69

Metadata Storage Modes (2/2)

Decentralized Storage

- Advantage
  - High scalability
  - Fault tolerance, no central point of failure
  - Better usage of available resources
  - Decentralized architecture → decentral storage
  - Transparent access
- Disadvantage
  - Slower access → an index speed it up in many cases
  - Complicated logic behind
  - No transaction guarantees

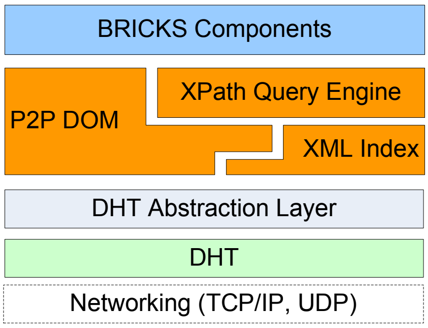


BRICKS
Building Resources for Integrated Cultural Knowledge Services

Page 70

Decentralized Storage in BRICKS

- Used for administrative metadata
- Data spread within BRICKS community, but transparent access to users
- Uses P2P layer for BNode communication
  - Discovering neighboring peers
  - Routing and processing messages within BRICKS community, but without global topology knowledge
- Implements
  - Well-known W3C DOM API for creating and accessing XML documents
  - XPath language for querying XML documents
- Built-in protocols for
  - Maintaining high data availability through replication
  - Concurrency and consistency control



BRICKS
Building Resources for Integrated Cultural Knowledge Services

Page 71

Storage Locations of Metadata Types in BRICKS

- Local storage
  - Descriptive metadata (with decentralized index)
  - Technical (EXIF, etc.)
  - Security metadata (ACL, etc.)
  - Annotations
  - Ontologies (with decentralized discovery)
- Decentralized
  - Service descriptions
  - Collection descriptions

BRICKS
Building Resources for Integrated Cultural Knowledge Services

Page 72

Descriptive Metadata in BRICKS

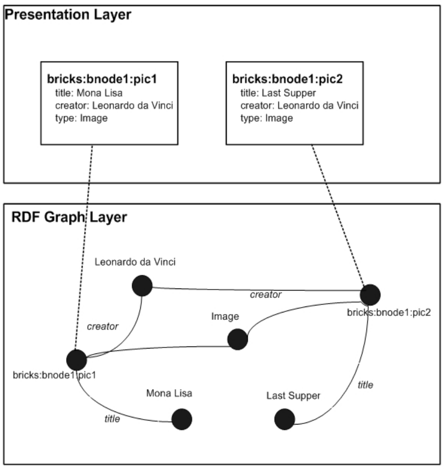
- The BRICKS **Metadata Manager**
  - is responsible for managing cultural assets from various institutions
  - must support arbitrary metadata formats from various institutions
- Schema/Format definition
  - the semantics of records is modeled and exposed in OWL-DL
- Bibliographic records
  - are internally stored and represented in RDF
- Controlled vocabularies, Thesauri, etc.
  - supported if they are represented in RDF, RDFS or OWL

BRICKS
Building Resources for Integrated Cultural Knowledge Services

Page 73

Metadata Manager - Design Decisions

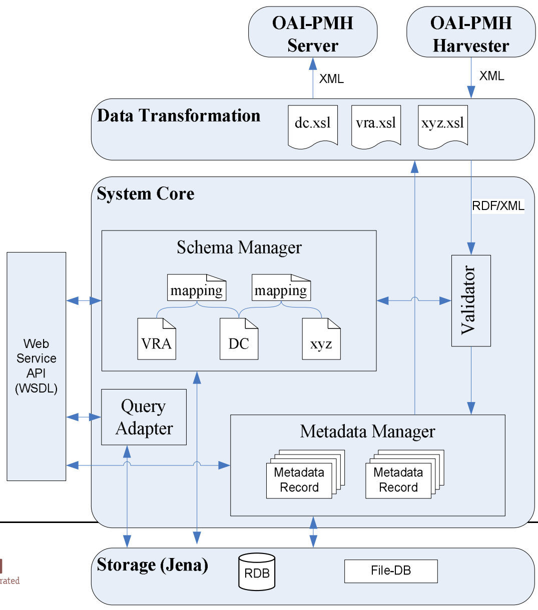
- RDF and OWL
  - are well-suited for handling heterogeneous metadata internally (!)
- But:
  - do NOT show triples or anything else related to RDF or OWL to the user or front end developers!
  - nobody wants to obtain triples in a query result
- In BRICKS we've decided to
  - completely hide RDF and OWL from users and high-level applications
  - expose it only for machines



BRICKS
Building Resources for Integrated Cultural Knowledge Services

Page 74

Architecture of the BRICKS Metadata Manager



BRICKS
Building Resources for Integrated Cultural Knowledge Services

Page 75

BRICKS and Existing Systems

- The BRICKS idea is not to replace but to integrate with existing systems
- Rely on already accepted protocols in the Digital Libraries domain to tap existing metadata and content databases
- Currently available:
  - Import of “existing” metadata via the OAI protocol for metadata harvesting (OAI-PMH)
    - easy to understand
    - easy to use and minimal implementation effort for institutions
    - gives us a minimum level of interoperability (Simple Dublin Core)
  - Wrapping of SRU sources (Search/Retrieval via URL)
    - Integration of SRU sources as collections
    - Providing BRICKS metadata via SRU interface

BRICKS
Building Resources for Integrated Cultural Knowledge Services

Page 76

BRICKS Collection Management
or
How to navigate through the BRICKS information space?

BRICKS
Building Resources for Integrated Cultural Knowledge Services

Page 77

Outline

- What is a collection manager
- Requirements on collections
- Implementing the collection management
- Searching a distributed DL
- Implementing the search mechanism

BRICKS
Building Resources for Integrated Cultural Knowledge Services

Page 78

Outline

- What is a collection manager
- Requirements on collections
- Implementing the collection management
- Searching a distributed DL
- Implementing the search mechanism

BRICKS
Building Resources for Integrated Cultural Knowledge Services

Page 79

What is a Collection Manager

- A Collection Manager is a mediator between:
  - the applications
  - and the DL contents.
- Applications may range from GUIs to arbitrarily complex programs for performing domain specific tasks.
- DL contents are the digital objects stored in the DL and their associated information.
- The job of the CM is:
  - to represent the DL contents via a conceptual model that is understandable and intuitive from an application viewpoint
  - to offer primitives for the manipulation of this model in support of the application tasks.

BRICKS
Building Resources for Integrated Cultural Knowledge Services

Page 80

The DL Conceptual Model

- We can view the contents of a DL as a set of documents, that is multimedia complex objects, with associated information, which supports the basic services offered on these objects.

- Distinctive features:
  - the size is typically $O(10^3)$ - $O(10^6)$
  - the members are very heterogeneous, *e.g.* in
    - Structure
    - Types
    - Language
    - Format
    - Contents

- *How do we go about taming such a set?*

BRICKS
Building Resources for Integrated Cultural Knowledge Services

Page 81

The DL Conceptual Model

- The problem is not new: organization!

> ***"Hence we view the organization of a** semantic network digital library **as basically an** abstraction mechanism **in terms of which details from a lower level of representation are suppressed. This is a crucial issue when dealing with large** semantic networks digital libraries **-- just as structuring techniques are important in the development of programs"***

(Mylopoulos & Levesque, 1979)

- Digital Library Collections are an abstraction mechanism!

BRICKS
Building Resources for Integrated Cultural Knowledge Services

Page 82

Outline

- What is a collection manager
- Requirements on collections
- Implementing the collection management
- Searching a distributed DL
- Implementing the search mechanism

BRICKS
Building Resources for Integrated Cultural Knowledge Services

Page 83

Requirements: Content Providers

Content providers structure its object space in sets of items termed "collections“
→ Natural to mirror real-world collections and to hold the primary content of the DL.

BRICKS

- These containers are called Physical Collections
- A physical collection is a set of content items which belong to the same content provider and are homogeneous *from the provider point of view:*
  - items are of the same kind
  - items are described by the same metadata format(s)
  - Items have same digital rights
  - …
- In BRICKS the contribution of a content provider to the DL is always defined in terms of one or more physical collections, thus when content items are added to (removed from) physical collections, they are added to (removed from) the digital library.

BRICKS
Building Resources for Integrated Cultural Knowledge Services

Page 84

Requirements: Content Providers

- Conversely: **a content item exists in the digital library only if there is a physical collection holding it**. Physical collections partition the DL information space: they are pair wise disjoint and their union makes up the content of the DL.
- **Physical collections are structured in (physical) sub-collections**, a notion that has been found a useful organizational mechanism by content providers.
- **A physical collection can have an arbitrary number of sub-collections** but only one parent collection. The graph of the sub-collection relation is a forest, each tree of which has a physical collection at the root. Within this tree, a content item of the physical collection belongs to exactly one sub-collection.
- **Physical collections are a central notion in the BRICKS content model**: not only content is organized by physical collection, but so is the discovery of resources and the definition of logical collections.

BRICKS Building Resources for Integrated Cultural Knowledge Services

Page 85

Requirements: Content Consumers

- People go to libraries to acquire knowledge for carrying out their own tasks.
- Typically, they search the whole library and end up with a subset of the library contents, consisting of the items that are relevant to them.
- This subset may be regarded as the *consumer view of the library*.
- The view is never static:
  - Consumers may evolve it manually, by adding newly discovered items or removing no longer useful ones
  - The view may evolve automatically: consumers describe their needs in some language and the items that satisfy these needs are added to the view:
    - *pull mode*: the consumer initiates the process (BRICKS)
    - *push mode*: someone else initiates the process (*e.g.* publish-subscribe)

BRICKS Building Resources for Integrated Cultural Knowledge Services

Page 86

Requirements: Content Consumers

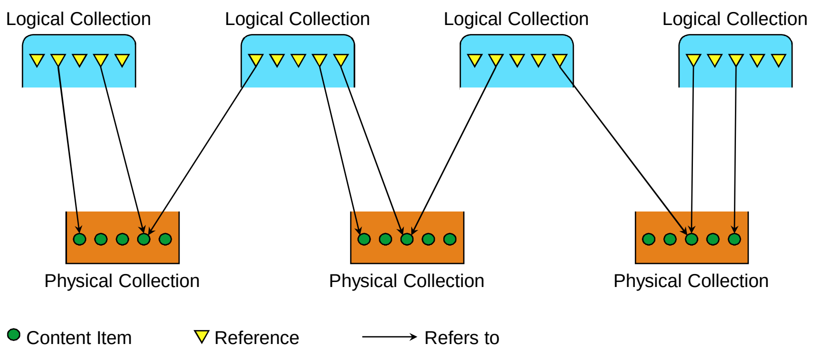
- This notion of *view* is captured in BRICKS by Logical Collections.
- A Logical Collection in BRICKS is a set of *references* to DL items, with identity.
- Any registered BRICKS user can create and operate upon Logical Collections.
- By creating logical collections, BRICKS users organize the information space and at the same time enrich the Digital Library content: once a logical collection is created, it becomes part of the digital library information space and can be *e.g.* searched by other users as any other digital library resource.
- Operations on Logical Collections affect *references* not content items!

BRICKS Building Resources for Integrated Cultural Knowledge Services

Page 87

Requirements: Content Consumers

- A BRICKS content item may be referenced in many Logical Collections.
- Conversely, a Logical Collection may contain references to many items, coming from many different Physical Collections.



BRICKS Building Resources for Integrated Cultural Knowledge Services

Page 88

Static Collections

- BRICKS logical collections come in two flavors:
  - Static Collections
  - Dynamic Collections (a.k.a. Stored Queries)

- A static collection can be evolved manually by its owner, by inserting or removing references.
  - If the collection has been created with the intent of holding, for instance, all works of Maurolico on conics (intention), there is no guarantee that at any moment the collection indeed contains (extension) references to all Maurolico works on conics available in the digital library.
  - A static collection is a static container which communicates with the rest of the digital library only via the intervention of the users authorized to modify its extension.

BRICKS
Building Resources for Integrated Cultural Knowledge Services

Page 89

Dynamic Collections

- A dynamic collection is defined by a query over the DL content, including other collections.
- Any time the dynamic collection is accessed, either in browsing or via a query, the defining query is evaluated.
- Dynamic collections evolve automatically as the DL content evolves.
- Dynamic collections evolve **only** automatically: no operation is offered to add or remove references from a dynamic collection.

- Static collections are extension-driven:
  - for consumers who cannot describe their needs other than extensionally, i.e. by pointing at the relevant items
- Dynamic collections are intension-driven:
  - for consumers who can describe their needs as a BRICKS query
    - BRICKS supports only the pull-mode (for now)

BRICKS
Building Resources for Integrated Cultural Knowledge Services

Page 90

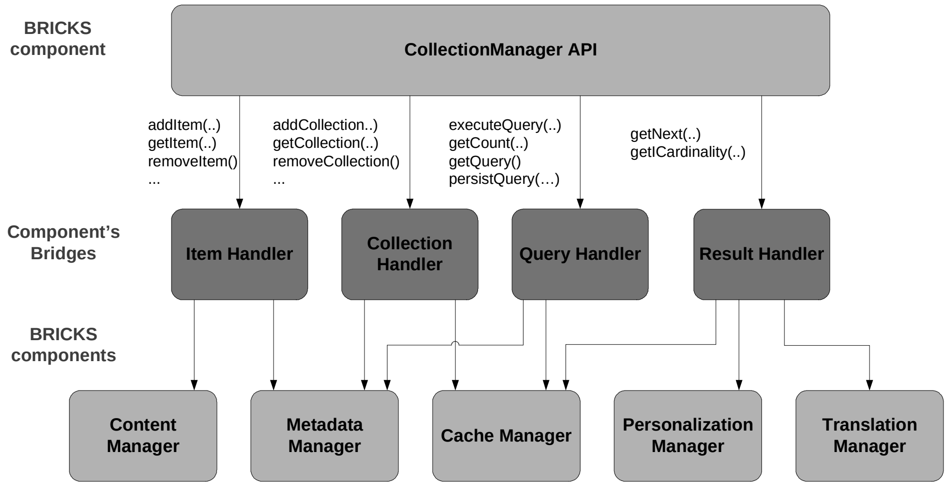
Outline

- What is a collection manager
- Requirements on collections
- Implementing the collection management
- Searching a distributed DL
- Implementing the search mechanism

BRICKS
Building Resources for Integrated Cultural Knowledge Services

Page 91

Architecture



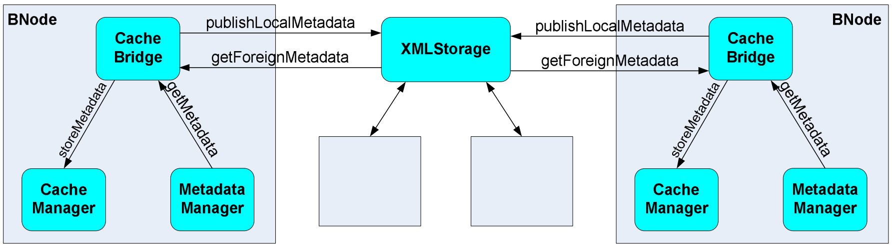
BRICKS
Building Resources for Integrated Cultural Knowledge Services

Page 92

Cache Bridge: Start-up

- In every BNode, the metadata of the locally defined collections are stored in the local Metadata Manager
- At BNode start-up, the Cache Bridge:

1. Publishes the local collection metadata from the local Metadata Manager
    - into the local Cache and
    - into the decentralized XML Storage (to be retrieved by foreign BNodes)
2. gets foreign collection metadata from the XML Storage into the local Cache



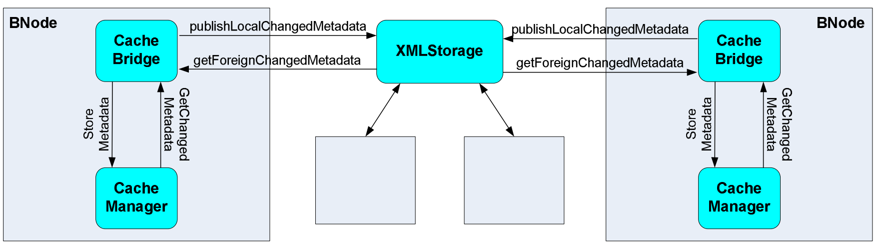
BRICKS Building Resources for Integrated Cultural Knowledge Services

Page 93

Cache Bridge : Runtime

At runtime, whenever the local collections metadata are modified (by some operation)

1. the Cache Bridge gets the the modified local metadata from the Cache Manager and publishes them to the XML Storage for the foreign BNodes to retrieve them upon synchronization
2. Periodically synchronizes the foreign collections metadata with the XML Storage



BRICKS Building Resources for Integrated Cultural Knowledge Services

Page 94

Outline

- What is a collection manager
- Requirements on collections
- Implementing the collection management
- Searching a distributed DL
- Implementing the search mechanism

BRICKS
Building Resources for Integrated Cultural Knowledge Services

Page 95

Searching a distributed DL

- The users of a DL form a very heterogeneous class, ranging from casual, computer-illiterate people to highly specialized scholars.
- All these users expect the DL system help them discover the DL objects of interest, independently from:
  - Location
  - Language
  - Type
  - Format
  - Structure
- in a way that reflects their preferences, if any, amongst which language plays a fundamental role.

BRICKS
Building Resources for Integrated Cultural Knowledge Services

Page 96

Requirements

- An important distinction: application-dependent *vs* application-independent discovery
- Application-dependent: based on criteria highly specialized with respect to a class of applications
  - x-ray images similar to a given one
  - brilliant executions of a piece of music
  - video sequences giving a feeling of anguish
  - spectacular takeovers in Formula 1 races
- Application independent: objective criteria
  - movies directed by Woody Allen
  - recent books on bio-informatics
  - Chopin's preludes played by Pollini

BRICKS
Building Resources for Integrated Cultural Knowledge Services

Page 97

Requirements

The query facility of a DL must be:

- extensible: it must be possible to plug-in it specialized search engines, devoted to capture the semantics of application-dependent search criteria
- flexible: it must be possible to express different kinds of queries, each addressing a different level of skill or knowledge of the DL contents
- user-sensitive: it must adapt to the preferences of the user
- efficient: it must respond as fast as possible

BRICKS
Building Resources for Integrated Cultural Knowledge Services

Page 98

Outline

- What is a collection manager
- Requirements on collections
- Implementing the collection management
- Searching a distributed DL
- Implementing the search mechanism

BRICKS
Building Resources for Integrated Cultural Knowledge Services

Page 99

The BRICKS query language

Types of queries:

- *Collection queries*: allow to discover collections by stating a Boolean condition on the collection metadata
- *DL objects queries*:
  - allow to discover content objects
  - may be contextualized to certain BNodes or collections
  - may be personalized
  - come in 3 flavors:
    - Simple
    - Advanced
    - Ontology

BRICKS
Building Resources for Integrated Cultural Knowledge Services

Page 100

Simple queries

- A simple query is the simplest form of query and most popular
- It is meant to serve casual users or users having a rather vague idea of the desired resources.
- These users typically do not want to address their search to specific metadata attributes, or operators.
- The language for simple queries allows to express sequence of unconstrained terms, very much in the style of e.g. Web search engines. In their search, users will be able to use:
  - wild cards: ca?lo databas*
  - phrases: “jakarta apache”
  - proximity operators: "jakarta apache"~10
  - Boolean operators: "jakarta apache" AND "jakarta lucene“
- In a simple search, metadata records are seen as texts whose words are the metadata attribute values.

BRICKS
Building Resources for Integrated Cultural Knowledge Services

Page 101

Advanced queries

- For users who can characterize their information needs very precisely in terms of a metadata schema, such as librarians or expert library users.
- An advanced query is a
  - Boolean combination of simple conditions
  - on metadata fields, possibly coming from different schemas:

DC.creator = “Bob” AND XY.date > 01.01.2000

- In an advanced search, metadata records are seen as sets of (attribute value) pairs, exactly like database records, which may or may not satisfy a query stating simple conditions on such attributes.

BRICKS
Building Resources for Integrated Cultural Knowledge Services

Page 102

Ontology queries

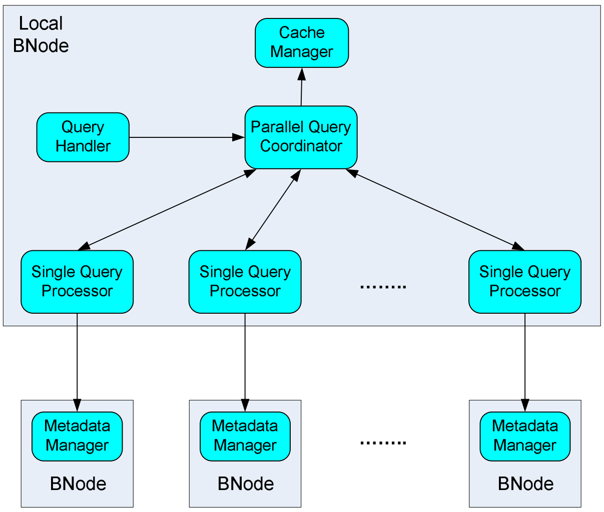
- For highly specialized users, who are familiar with an ontology formally conceptualizing their domain of interest, and want to retrieve documents annotated by descriptions derived from that ontology.
- An ontology query is a SPARQL expression:

```
PREFIX table: <http://example.org/book/book1#>
SELECT ?title, ?author
FROM <http://example.org/book/book1.owl>
WHERE {
    ?element table:title ?title
    ?element table:author ?author
}
```

BRICKS Building Resources for Integrated Cultural Knowledge Services

Page 103

Query Handler: Overall Architecture



BRICKS Building Resources for Integrated Cultural Knowledge Services

Page 104

Archaeological Sites Application Prototype

BRICKS
Building Resources for Integrated Cultural Knowledge Services

Page 105

Finds[1] identification today

- The public is an important source for understanding the past and archeological sites
- Thousands of objects are discovered by people during walking, traveling, gardening, etc.
- Museum curators, archaeologists, students, amateurs all need to identify these finds
- Traditional way of working
  - Examining objects
  - Preliminary identification
  - Comparison to specialist reference collections, e.g. roman coins
    - Sometime object descriptions are enough
    - Sometime it is the starting point for a scientific analysis process
  - Comparison to objects in the museum collection
  - Record object in detail

[1] With *Finds* we mean *archeological finds*

BRICKS
Building Resources for Integrated Cultural Knowledge Services

106

Page 106

Finds identification today

- Disadvantage of the traditional way
  - The reference collections have to be known by the curator
  - Curator needs information about new collections, e.g. when a museum opens its archive
  - Curators need to access them one by one
  - Collections are heterogeneous from point of view of data, search facilities, user interface, languages, etc.
- As a consequence curators invest a lot of time in the identification process

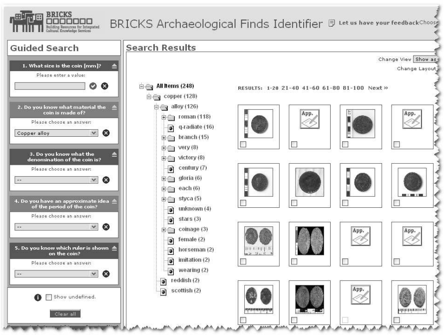
BRICKS
Building Resources for Integrated Cultural Knowledge Services

107

Page 107

Finds identification with BRICKS

- Examining objects
- Preliminary identification
- Comparison to specialist reference collections
  - Easy access to all relevant archeological collections (Even to unknown collection)
  - Advanced search facilities
    Support of ontologies, translation services, different types of search
  - One application to access all collections
  - Collaborative identification processes
- Comparison to objects in the museum collection
- Record object in detail



BRICKS
Building Resources for Integrated Cultural Knowledge Services

108

Page 108

Archaeological Sites Application Prototype

DEMO

Page 109

Lessons Learned

- Developers and system designers have a longer learning curve about service oriented technologies
- A service oriented design is different from traditional design, e.g. the number of function should be limited and well considered
- Communication costs between services are often underestimated
- The communication of the concept of distributed architectures to the end user is a hard process
- An early prototype is helpful for the communication between users and developers

Page 110

Questions & Discussions

Information Society Technologies

BRICKS
Building Resources for Integrated Cultural Knowledge Services

Page 111

Approaches for Large Scale Digital Library

Forschungszentrum · Research Center L3S

ISTI
ISTITUTO DI SCIENZA E TECNOLOGIE DELL'INFORMAZIONE "A. FAEDO"

Conclusions & Future Directions

Tutorial at RCDL 2007
October, 15th 2007

Thomas Risse
L3S Research Center, Hannover, Germany
Email: risse@L3S.de

In cooperation with Claudia Niederee (L3S), Carlo Meghini (CNR-ISTI), Heiko Schuldt (Uni Basel)

UNI BASEL

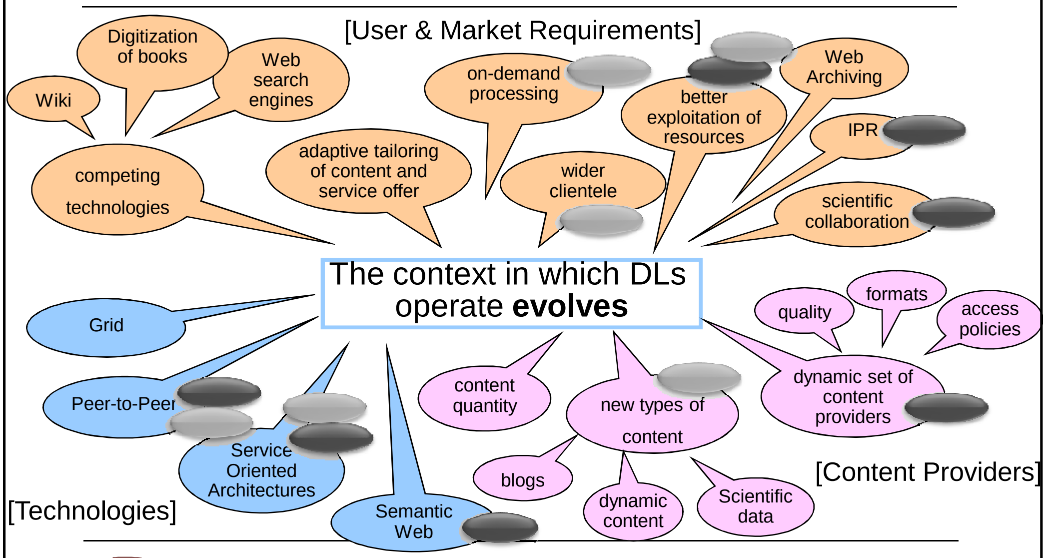
Agenda

1. Introduction & Motivation
2. Challenges of bringing DL to distributed Infrastructures
3. Underlying Technologies and their promises (SOA, P2P)
4. Solutions for decentralized DL infrastructures (with BRICKS Demos)
5. Conclusions and future directions

BRICKS Building Resources for Integrated Cultural Knowledge Services

Page 113

Summary



BRICKS (http://www.brickscommunity.org/)
DELOS (http://www.delos.info/)

Page 114

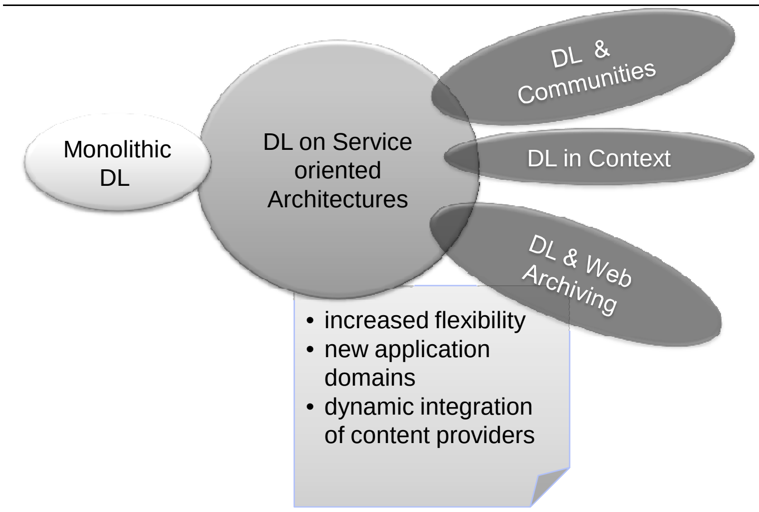
Next Steps for Establishment & Take Up

- Consolidation & Standarization
  - compatibility
  - exchangable modules
- Attraction of a wider community
  - cultural heritage related
  - new application areas (e.g. e-Science)
- Solving of open challenges
  - IPR
  - quality assurance
  - long-term preservation
  - information integration
  - interoperability

BRICKS Building Resources for Integrated Cultural Knowledge Services

Page 115

Future Directions: DL in Context

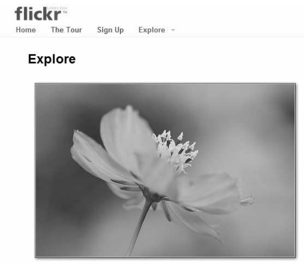


BRICKS Building Resources for Integrated Cultural Knowledge Services

Page 116

DL & Communities

- Strong trend: Involvement of communities
  - community based content creation (e.g. Wikipedia)
  - community-based content collection (e.g. YouTube, Flickr)
  - community-based recommendation + rating (e.g. online Bookstores)
  - community based tagging (folksonomies, ...
- DL Opportunities
  - use the intelligence of communities to improve DL content and service offer (e.g. content enrichment and interlinking)
  - exploit DL technologies in community-based systems
  - create even richer content by combining community created content with DL content in seamless ways.



YouTube
Broadcast Yourself™

BRICKS
Building Resources for Integrated Cultural Knowledge Services

Page 117

DL in Context

Support for the individual so far:

- standard query based access
- Personalization support (e.g. considering interests and preferences)
- Supprt for personal digital libraries (e.g. Daffodil) and virtual libraries (DILIGENT)

DL Opportunities

- Improved personalization support
- integration of DL functionality in working processes
- targeted pro-active information provision
- requires process- and context modelling
- application example: e-Science

BRICKS
Building Resources for Integrated Cultural Knowledge Services

Page 118

DL and Web Archiving

- Increased role of Web in information access, content creation, reflecting opinions, performing transactions etc. in daily life activities, society, culture, politics, education, etc.
- → increasing importance of sound Web archiving
- first generation of Web archiving technology in place
- but: poor access methods for Web archives e.g. archive.org
- little integration with other information sources
- special challenges due to inherent structure w.r.t access

- DL opportunities:
  - apply experience of DL community in metadata, search, etc.
  - integrate Web archive collection with DL collection for more comprehensive information provision
  - develop new types of access methods (time, evolution) and application for targeted communities

BRICKS
Building Resources for Integrated Cultural Knowledge Services

Page 119

Thank you for your attention

BRICKS
Building Resources for Integrated Cultural Knowledge Services

Page 120