europeana
Kultur. Denken

Interoperability Challenges in Digital Libraries

(Mostly) conceptual, (some) technical and (a few) political aspects of a core notion about to degenerate into a buzzword

Prof. Dr. Stefan Gradmann
Humboldt-Universität zu Berlin / School of Library and Information Science
stefan.gradmann@ibi.hu-berlin.de

europeana
Kultur. Denken

The Funny Talk You Won't Get ...

Overview

- **Intro**
  - Some politics (EC FP7, Europeana)
  - Definitions, Motivations
- **Technical**
  - Interoperability aspects of **Selected Frameworks** for DL modelling
    - DELOS, DRIVER, OAI-ORE, DCMI abstract, JISC Information Environment, JCR, iRODS
    - Thanks to Donatella Castelli, Wolfram Horstmann, Andy Powell, Herbert van de Sompel, Pete Johnston and a lot more ...
    - Deliberately discarded: DAREnet, aDORe. CORDRA / IMS DRI (CP and ECL), e-Framework, O.K.I. Open Service Interface Definitions (OSIDs), and many more ...
  - Six dimensions of the **interoperability matrix** abstracted from these frameworks plus some thoughts on **abstraction levels**
- Politics revisited: **Interoperability 2.0**

europeana

Background

- Google Books & related political trouble helped to trigger
- **EC i2010 (Lisbon) agenda** with Digital Libraries as one of 3 'flagship initiatives': the setting up of the European Digital Library as a common multilingual access point to Europe's distributed digital cultural heritage including all types of cultural heritage institutions
  - 2008: at least 2 million digital objects; multilingual; searchable and usable; work towards including archives.
  - 2010: at least 6 million digital objects; including also museums and private initiatives.
  - "I am not suggesting that the Commission creates a single library. I envisage a network of many digital libraries – in different institutions, across Europe." V. Reding (29 September 2005)
- => High level group, Expert group, **Interoperability group**
  - Contribute to the **short term DL agenda** => identify areas for short term action and recommend elements of an action plan (**list of prioritised feasible options**)
  - Contribute to the **long term DL agenda** => identify key elements for a **long term strategy**

Definitions, definitions ...

- “Interoperability is the capability to **communicate**, **execute programs**, or **transfer data** among various **functional units** in a manner that requires the user to have **little or no knowledge** of the **unique characteristics** of those units.”
(ISO/IEC 2382 Information Technology Vocabulary)
- „the ability of two or more systems or components to **exchange information** and to **use** the information that has been exchanged.“
(IEEE)
- Interoperability is a property referring to the ability of diverse systems and organizations to work together (inter-operate). The term is often used in a technical systems engineering sense, **or alternatively in a broad sense, taking into account social, political, and organizational factors that impact system to system performance** (WikiPedia).
- => Plethora of definitions should make you suspicious!

Talk Motivation

- “Open and interoperable are two words in the Information Technology world susceptible to **misunderstanding** at best, at worst to **self-serving abuse**. It is important to clarify their accepted meanings, because how they are understood in the market has direct practical consequences for consumers, vendors and regulatory authorities.“ (European Committee for Interoperable Systems/ECIS, http://www.interoperability.eu/)
- Motivation: provide **conceptional** and **political** complements to the **technical**, mostly computer science driven approaches in DL.org.

Interoperability Motivations: Europeana as Example

- Europeana will be federating objects from **distributed sources**
- Europeana will be federating objects from **heterogeneous sources** with **different community background** – e. g. libraries vs. museums vs. archives ... but also scholars vs. policy makers vs. meta users ...
- Europeana will be part of a bigger framework of **interacting global information networks** including e. g. 'Digital libraries', scientific repositories and commercial providers
- Europeana will have to be built with **minimal development efforts** and thus rely as much as possible on **web and internet standards** and **existing building blocks**
- And this is why interoperability figures so prominently place in the name of the “technical” WP of EDLnet: **Interoperability is the heart of the technical vision of Europeana**!

europeana

Inter-what?

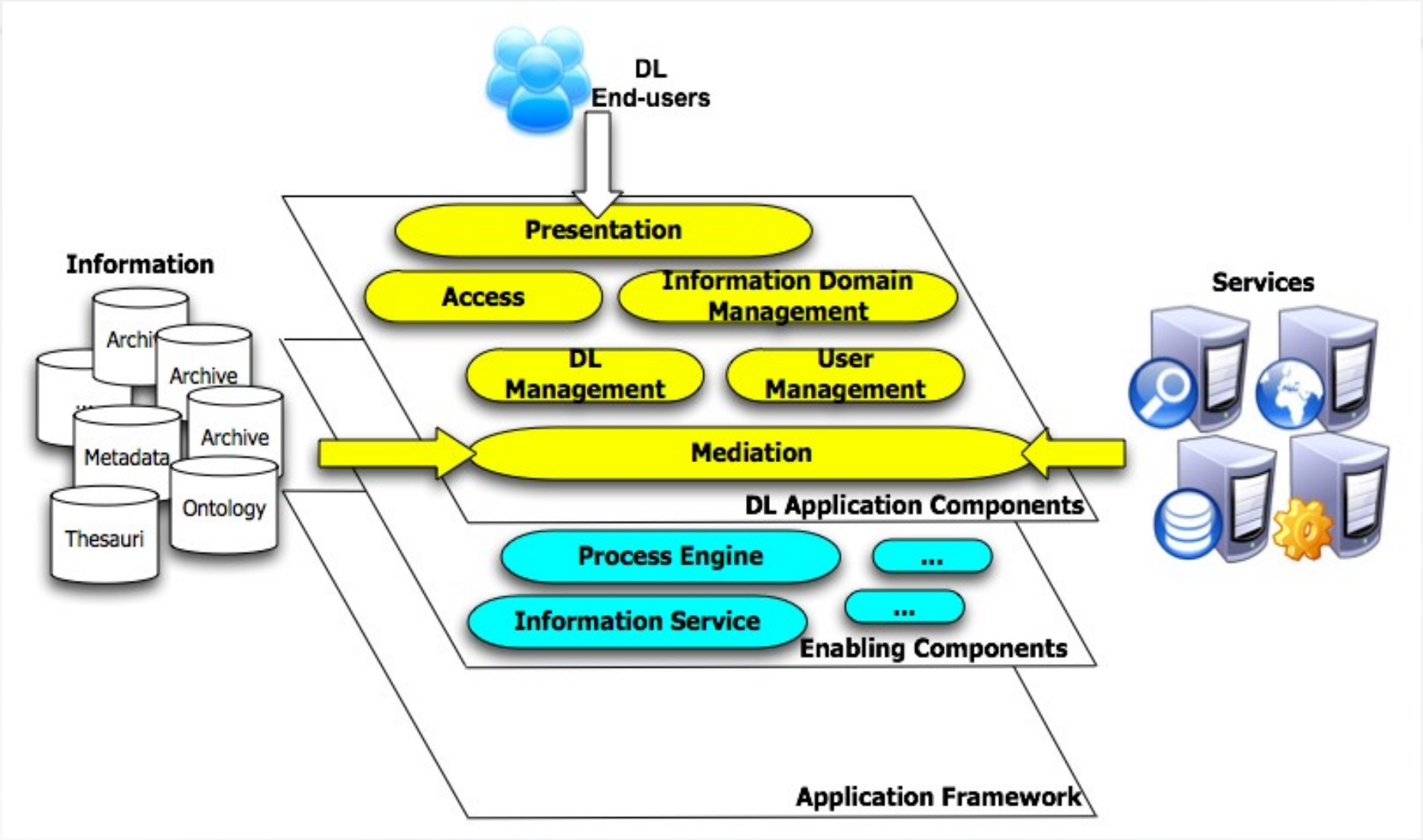
Selected Frameworks of Information Systems Architecture and Interoperation

DELOS Reference Model

A Computer Science Based Framework

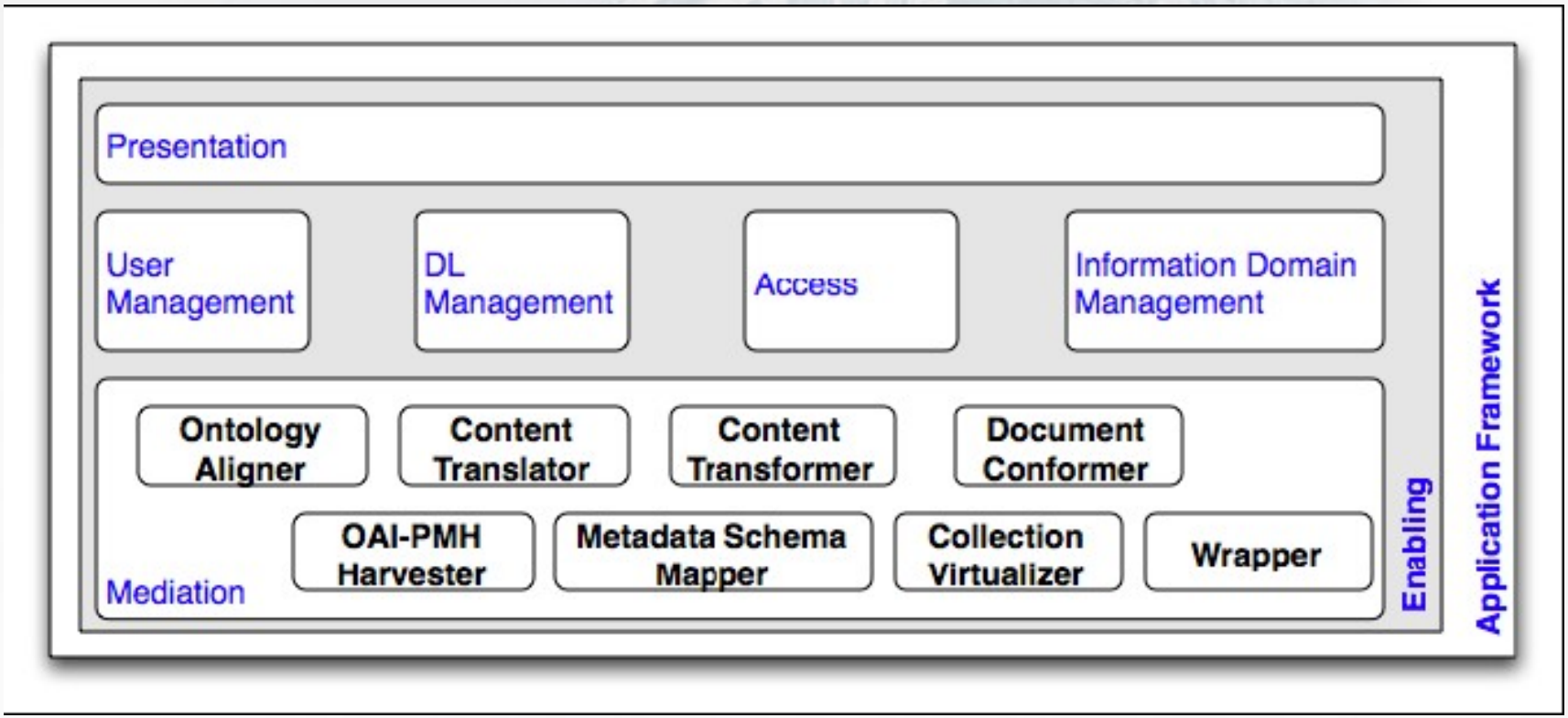
europeana

DL Reference Architecture: Functional Areas



DL Reference Architecture: Mediation Area

„Re-use, integration, interoperability are key requirements in the DL application area. In the current situation where no established rules nor principles exist for the development of digital library systems (DLS) the satisfaction of these requirements is difficult and has to be done on a case-by-case basis. In order to start overcoming this lack some of the DELOS partners active on the design of DL architectures decided to specify a *Reference Architecture* for component-based DLSs. These systems are particularly suitable to support one of the most important class of DLs: the *federated digital libraries*.“



DELOS Reference Framework: Characteristics

- Very **abstract model** rooted in Computer Science and only loosely related to cultural institutions' reality
- Although intended to create and enhance interoperability, the **reference architecture** still remains too abstract to really help
- It is unclear when work on the reference architecture will be taken up again and by whom
- The reference model is a **very good starting point for conceptual work**, even though it is not yet entirely mature and stable (and probably never will be, anyway!)

DCMI Abstract

A Very Abstract Framework

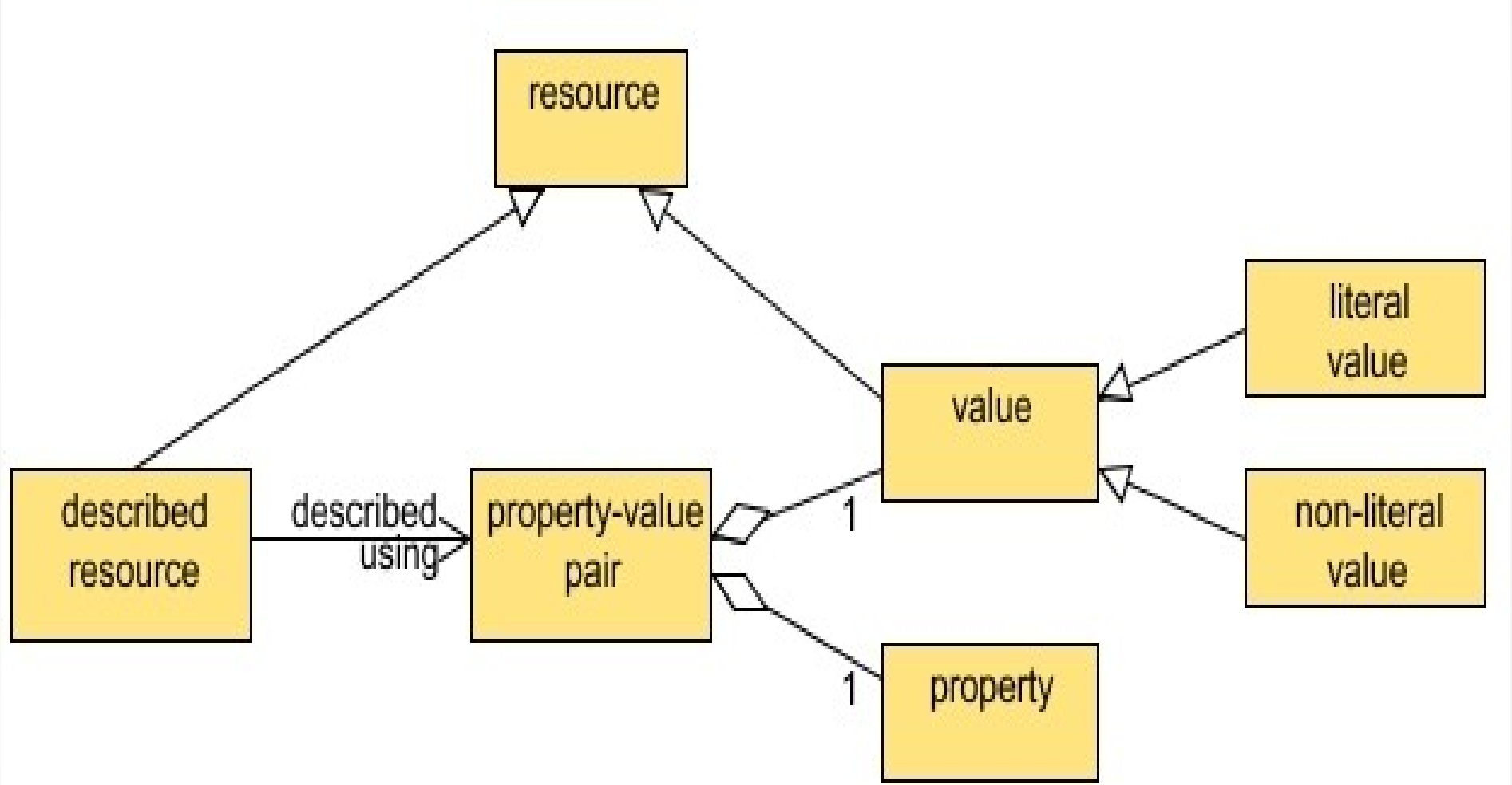
DCAM: Resources

- DCAM concerned with description of **resources**
- DCAM adopts Web Architecture/RFC3986 definition of resource
  - the term "resource" is used in a general sense for **whatever might be identified by a URI**. Familiar examples include an electronic document, an image, a source of information with consistent purpose (e.g., "today's weather report for Los Angeles"), a service (e.g., an HTTP to SMS gateway), a collection of other resources, and so on.
  - A resource is **not necessarily accessible via the Internet**; e.g., human beings, corporations, and bound books in a library can also be resources.
  - Likewise, **abstract concepts can be resources**, such as the operators and operands of a mathematical equation, the types of a relationship (e.g., "parent" or "employee"), or numeric values (e.g., zero, one, and infinity).

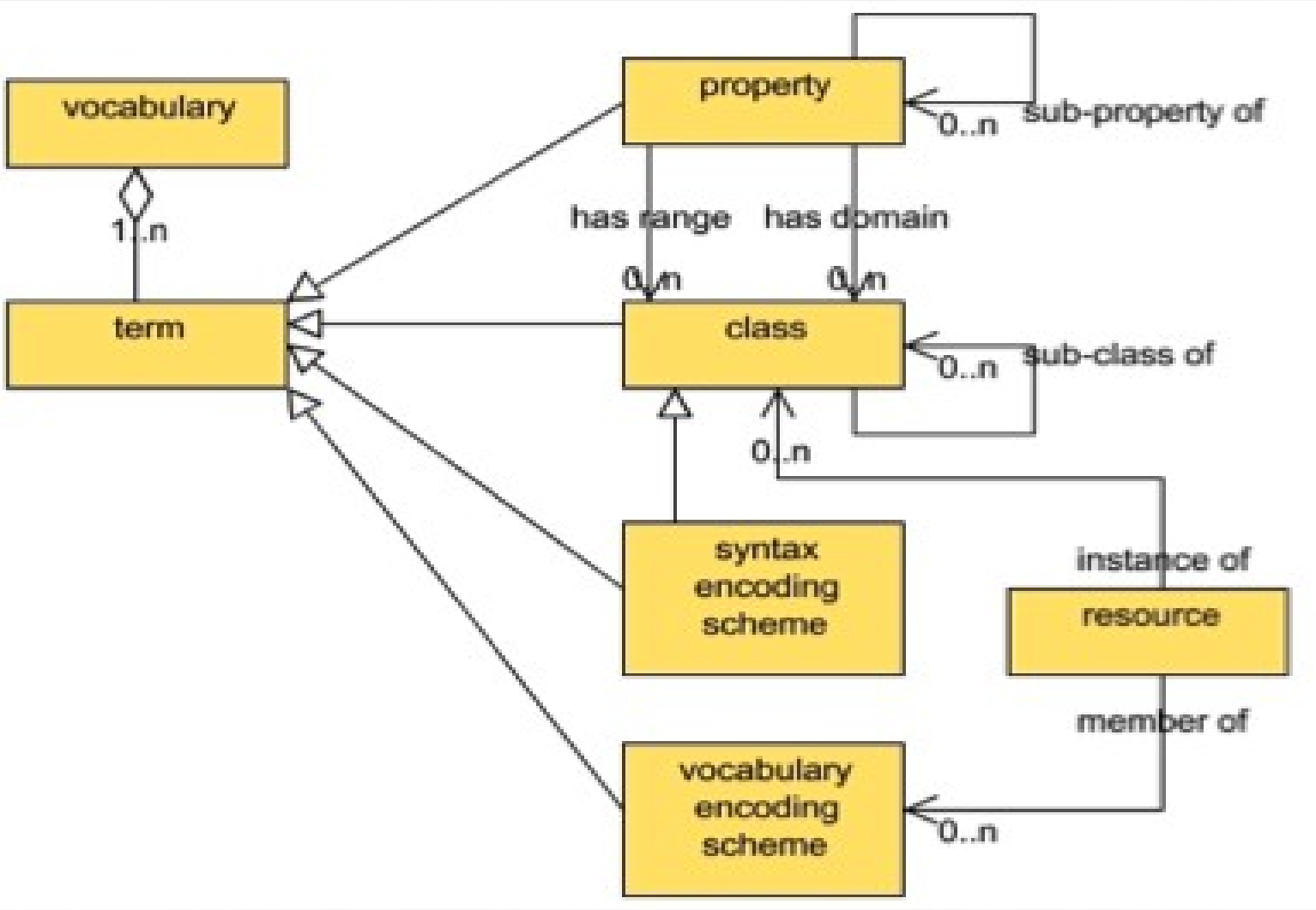
DCAM: Basics

- DCAM describes
  - **Components** and **constructs** that make up an information structure ("DC description set")
  - How that information structure is to be **interpreted**
- DCAM does not describe how to represent DC description set in concrete form
- DCAM describes various **types** of metadata terms, but does not specify the **use of any fixed set of terms**
- Made up of three related "information models"
  - **Resource** model
  - **Description** set model
  - **Vocabulary** model

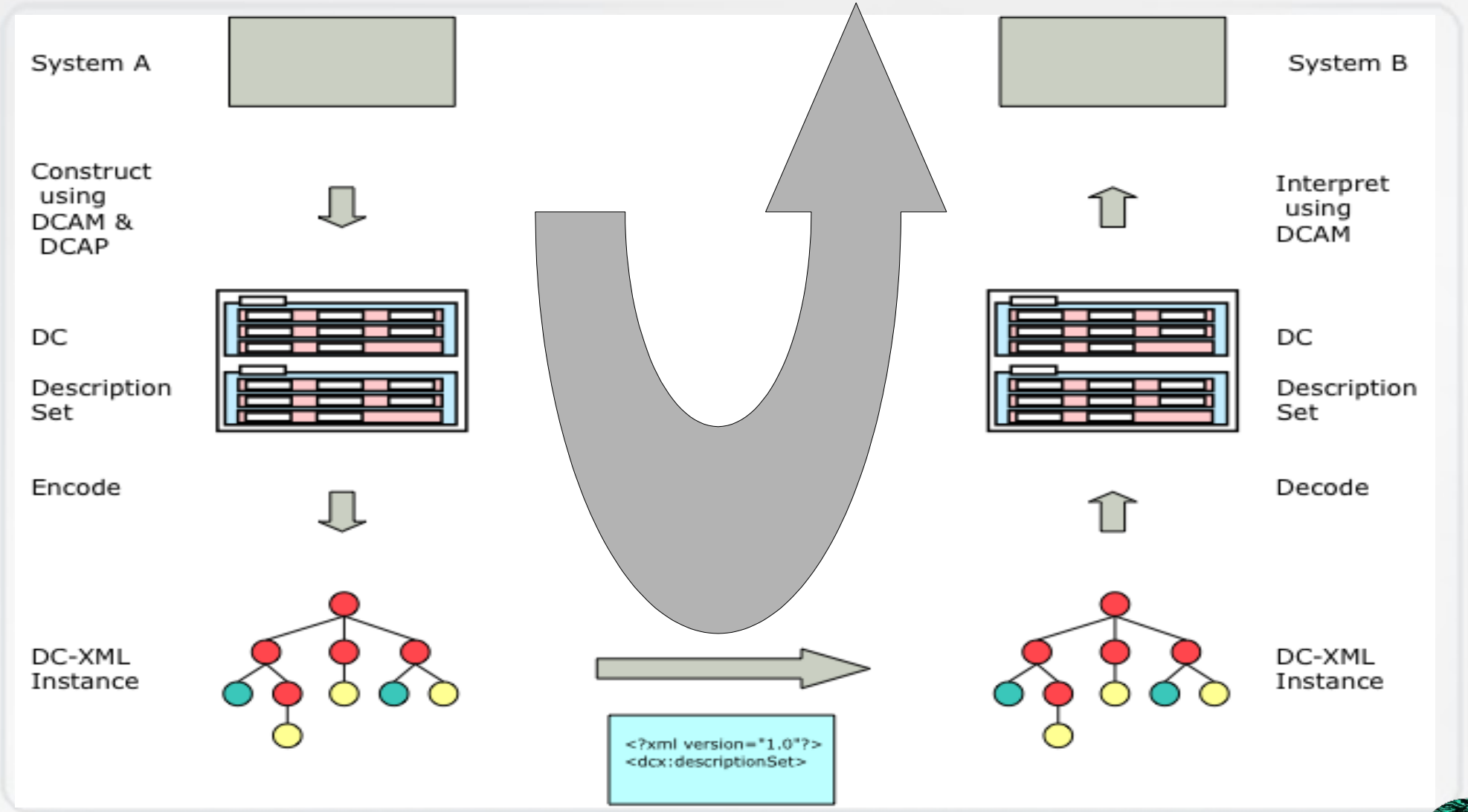
DCAM: Resource Model



DCAM: Vocabulary (1)



DCMI Abstract and Interoperability



DCMI Abstract: Characteristics

- Conceptually useful approach
- Relatively weak acceptance
- Probably too abstract to be considered in operational settings
- Combine with DC:Terms to create strong metadata interoperation framework

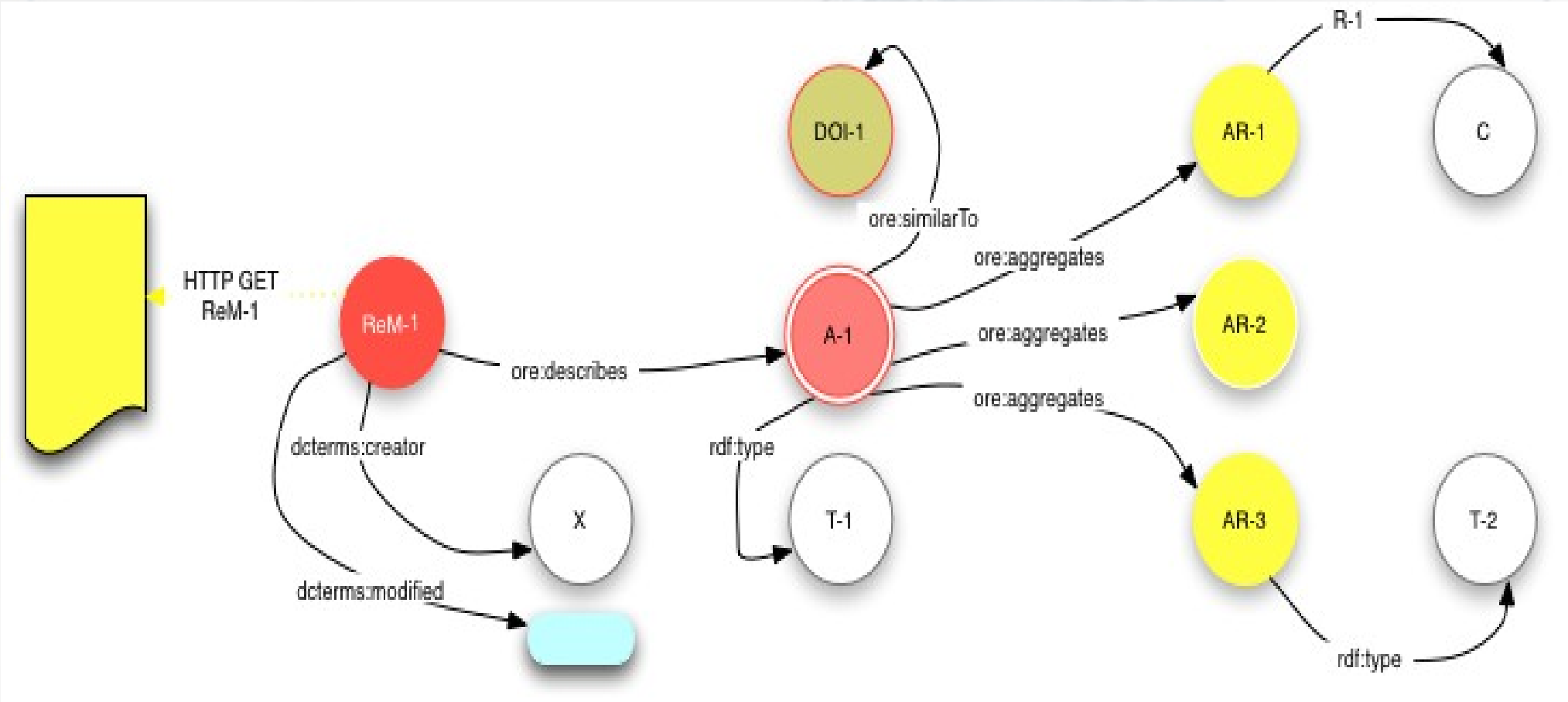
OAI-ORE

Generic W3C Methodology

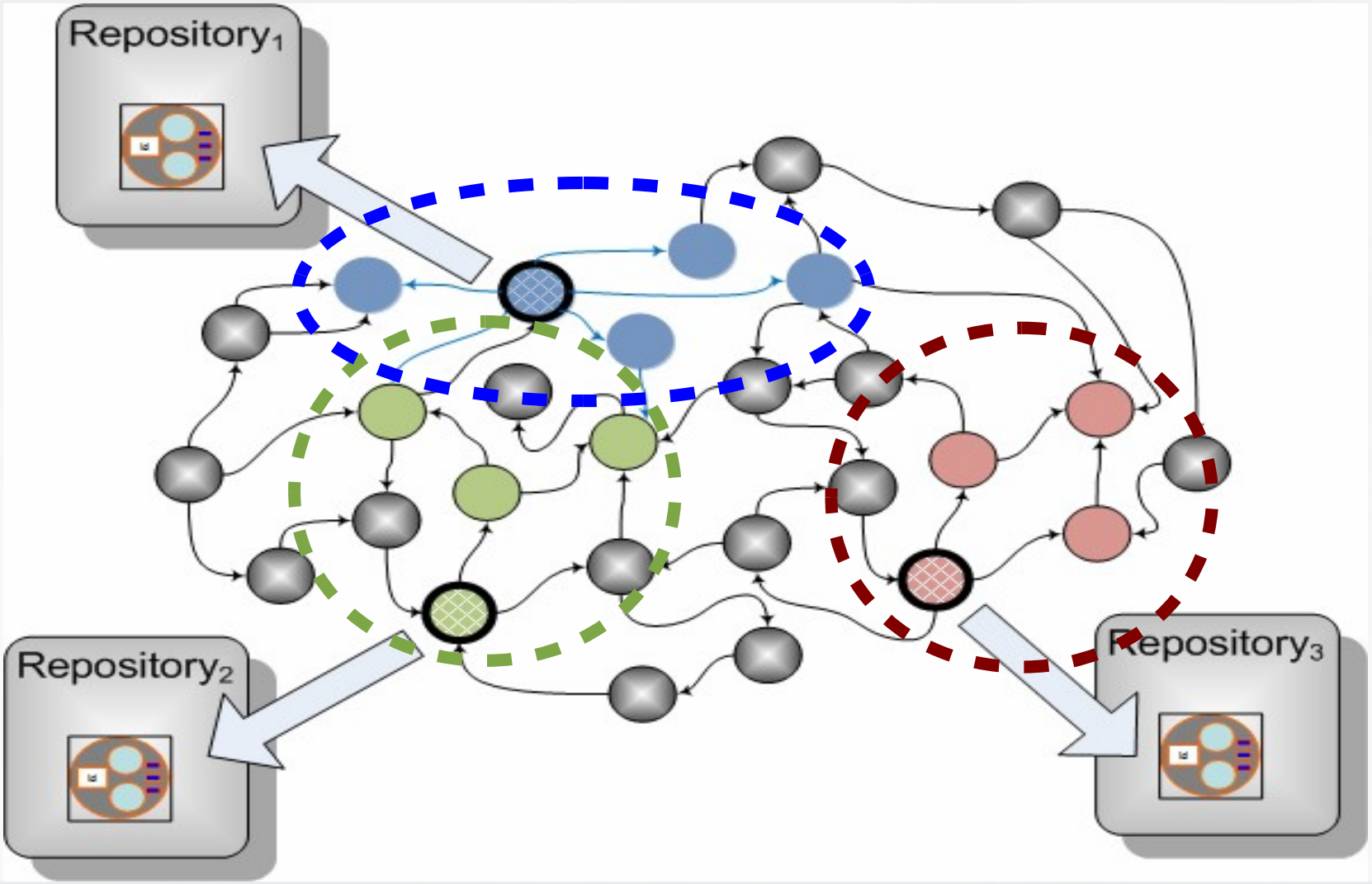
OAI ORE: Principles

- Goal: „Facilitate Use and Re-Use of Compound Information Objects (and of their component parts)“
- „How to deal with compound information objects in a manner that is in sync with the Web architecture?“
- By enriching the web graph with boundary information.

An Example Resource Map



... web graph with boundary information



ORE: Characteristics

- Limited (?) to exchange of web resource aggregations!
- But very useful withinr this limited (?) scope
- 100% based on standard, generic WWW technology
- Does not address/answer some fundamental issues such as boundary constitution
- => Relatively sure bet for persistent web based object modelling

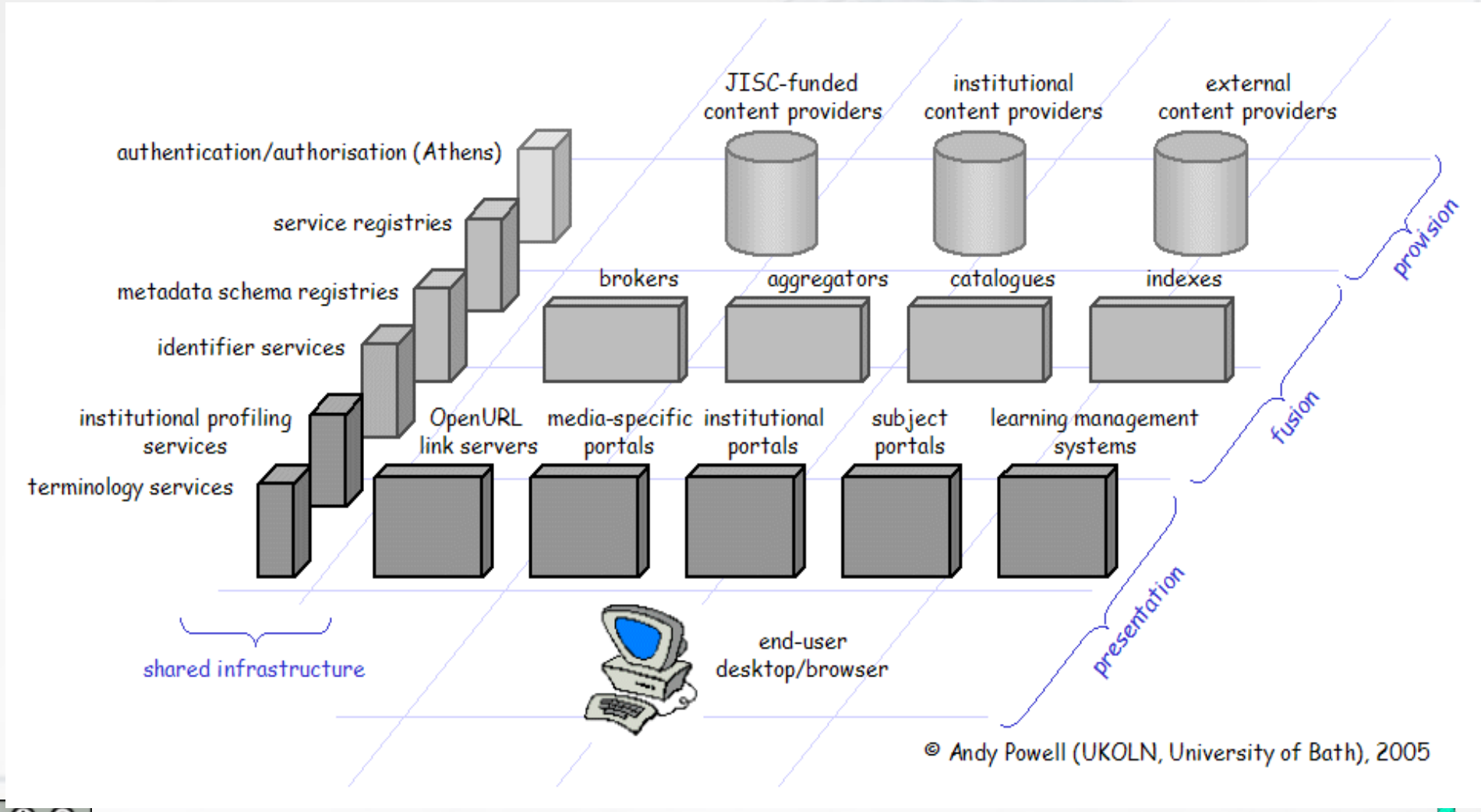
JISC Information Environment

Service Oriented Architecture

europeana

JISC Information Environment

- http://www.ukoln.ac.uk/distributed-systems/jisc-ie/arch/



JISC IE: Service Model for Linking Repositories

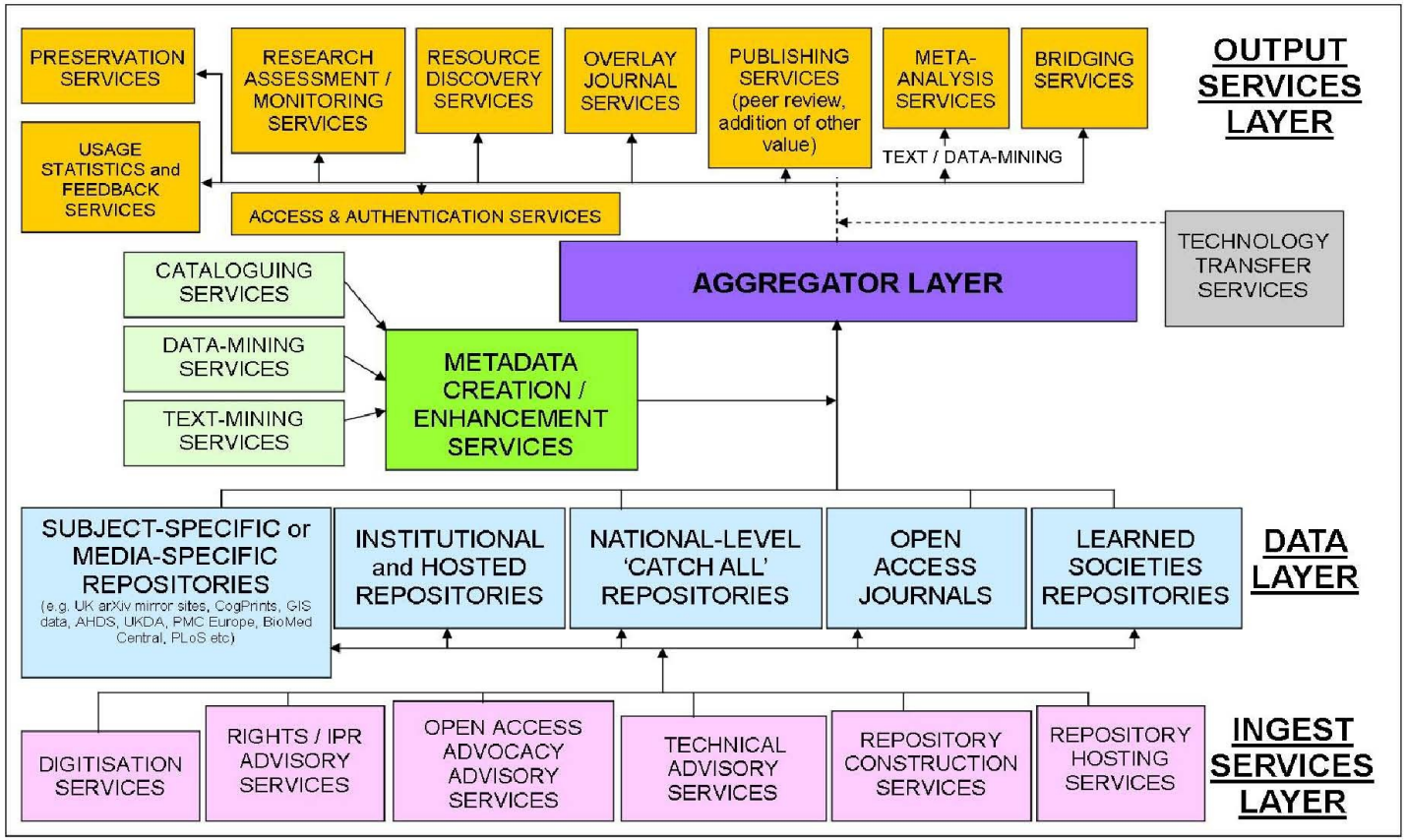


Chart A: Overall model for repositories and the services built across them

JISC IE: Characteristics

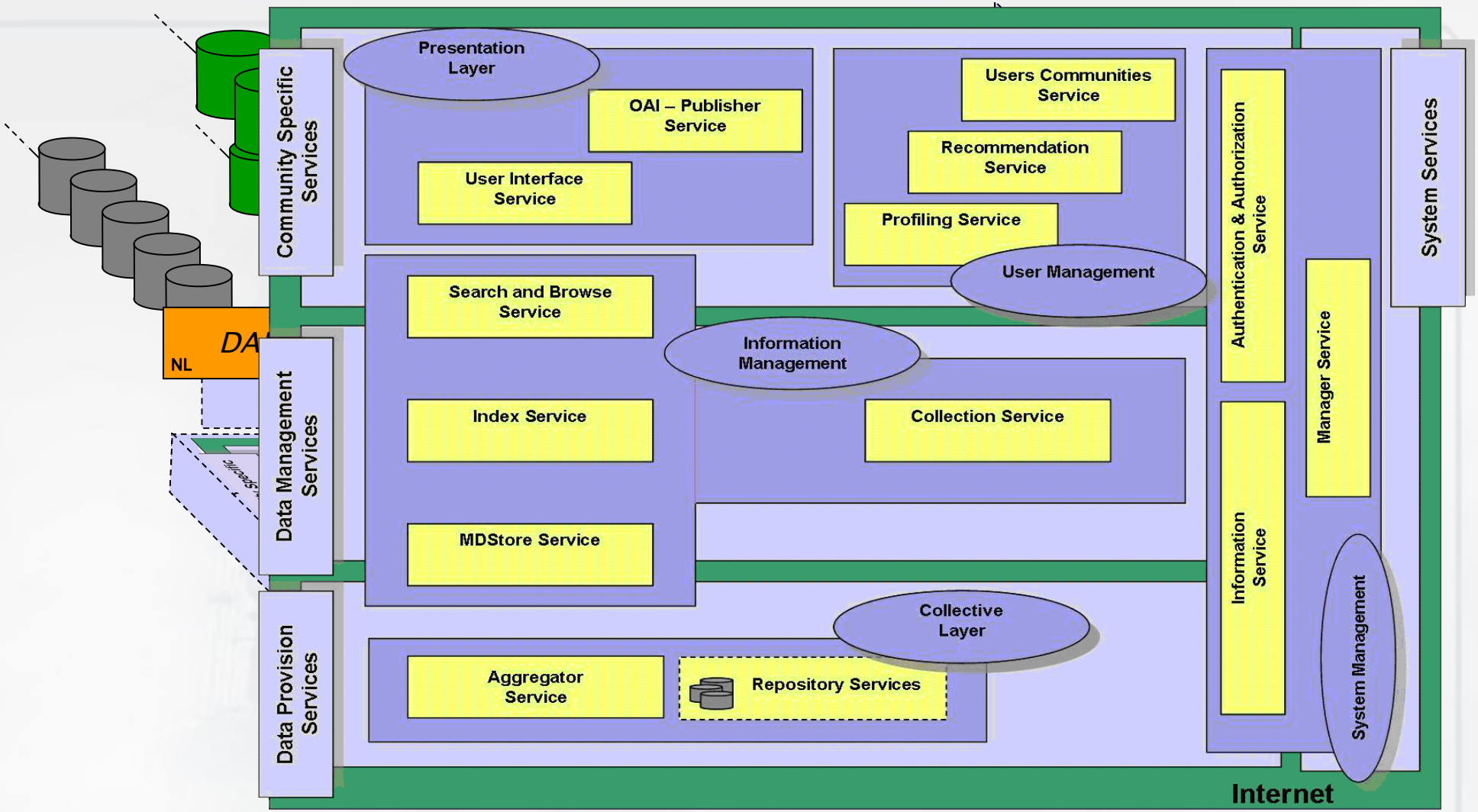
- Exclusively SOA oriented
- Objective is to „ ... support user-oriented services across digital repositories ...“ (Swan 2006)
- Service model is quite close to 'librarian' reality, even though explicitly designed for repositories

DRIVER

Harvesting + Services

europeana

DRIVER Open Service Architecture



DRIVER: Characteristics

- OAI Harvesting+
- Value added services on top of aggregated repository content
- Harvesting based model of repository federation
- Limited set of core functions
- Limited to textual objects, but currently being extended to complex and multimedia objects in DRIVER2
- => Infrastructure framework for providing platform interoperability

JCR (JSR170/283)

Content Infrastructure Interoperability with High Functional Granularity

JCR: Functionality Overview

- Granular Read/Write Access - This is the bi-directional interaction of content elements. Issues with access on a property level and not just on a "document" level
- Versioning - Transparent version handling across the entire content repository, providing the ability to create versions of any content and select versions for any content access or modification.
- Hard- and Soft-structured Content - An Object Model that defines how hard and soft-structured content could be addressed.
- Event Monitoring (Observation) - Possible use of JMS based notification framework allowing for subscription on content modification.
- Full-text Search and filtering - Entire (non-binary) repository content indexed by a full-text search engine that enables exact and sub-string searching of content.
- Access Control - Unified, extensible, access control mechanisms.
- Namespaces & Standard Properties - Defining default standard properties that will maintain namespace uniqueness and hierarchy.
- Locking and Concurrency - Standardized access to locking and concurrency features
- Linking - A standard mechanism to soft/hard link items and properties in a repository along with providing a mechanism to create relationships in the repository.
- ... more: Specification Document

JCR: Characteristics

- “ ... lay the foundations for a true industry-wide content infrastructure”: focus is on interaction with DMSs
- Industry standard: a very different type of community!
- Limited take-up in industry
- Immediate impact may be limited because of limitation to Java implementation ...
- ... but offers a very granular set of functional primitives!

iRODS

Microservices and Rules for Interoperability

iRODS: Microservices

- Micro services are small, well-defined **procedures/functions that perform a certain task**
- Users and administrators can chain these micro-services to implement a larger **macro-level functionality** that they want to use or provide for others.
- The task that is performed by a micro-service can be quite small or very involved. We leave it to the micro-service developer to choose the proper **level of granularity** for their task differentiation.
- Sample micro-services are “createCollection”, "assignAccess", "createPhysicalFile", “computeChecksum” and “replicateObject”.

iRODS: More Microservices

- Workflow Services:
  - nop, null - no action
  - cut - not to retry any other applicable rules for this action
  - succeed - succeed immediately
  - fail - fail immediately - recovery and retries are possible
- System Micro Services - Can only be called by the server process.
  - msiSetDefaultResc - set the default resource
  - msiSetRescSortScheme - set the scheme for selecting the best resource to use
  - msiSetDataObjPreferredResc - specify the preferred copy to use in case of multiple copies
  - msiSetDataObjAvoidResc - specify the copy to avoid
- User Micro Services - can be called by client through irule.
  - msiDataObjCreate - create a data object
  - msiDataObjOpen - open a data object
  - msiDataObjRead - read an opened data object
  - msiDataObjWrite - write
  - msiDataObjUnlink - delete
  - msiDataObjCopy - copy
  - msiDataObjRename - rename a data object

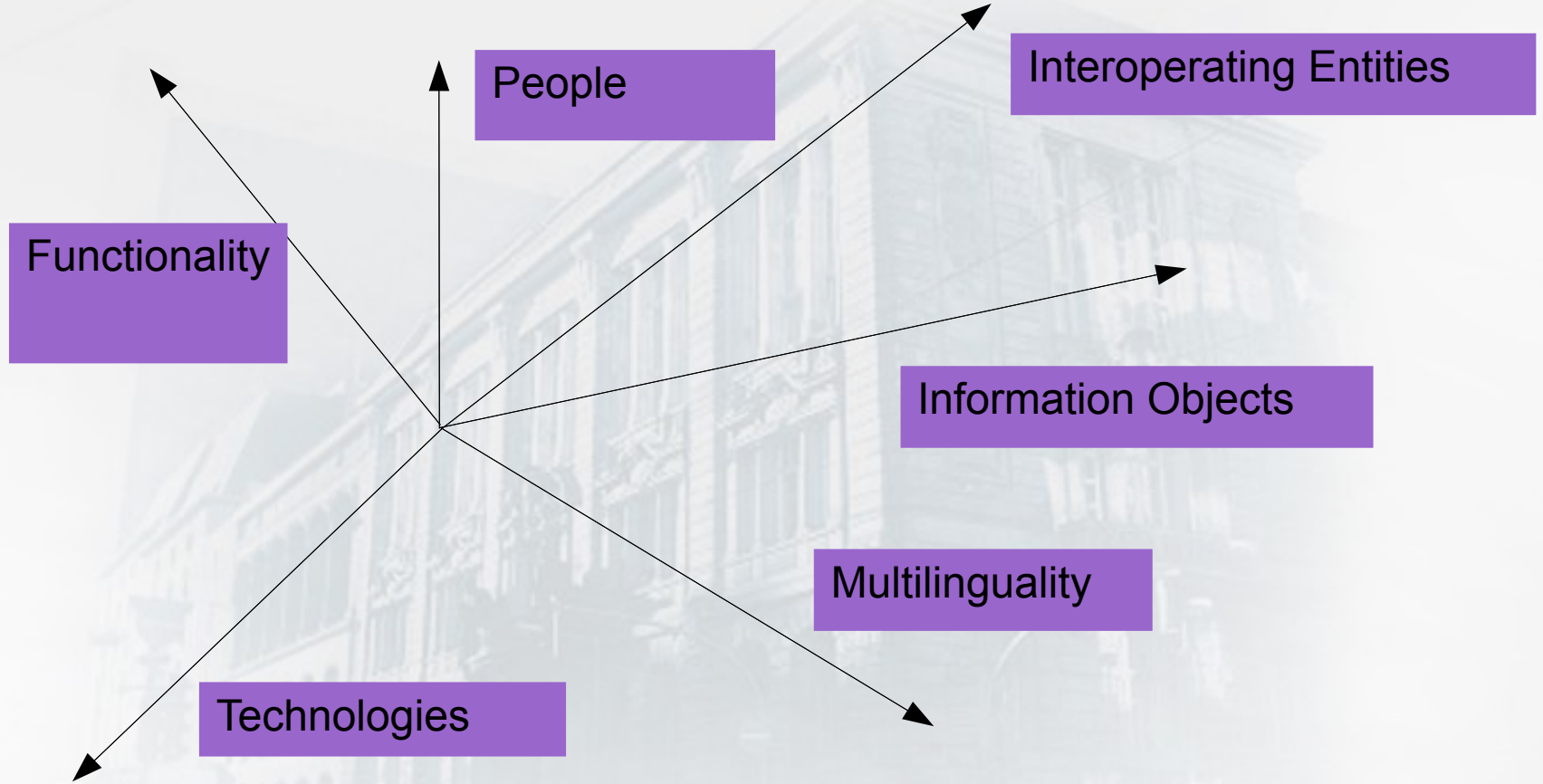
iRODS: Rules

- Rules are **definitions of actions** (or macro-level tasks) that need to be performed by the server.
- These definitions are made in terms of micro-services and other actions.
- Basically a rule is specified with a line of text which contains 4 parts separated by the '|' separator:
  actionDef | condition | workflow-chain |recovery-chain
  - **'actionDef'** is the name of the rule. It is an identifier which can be used by other rules or external functions to invoke the rule.
  - **'condition'** is the condition under which this rule applies. i.e., this rule will apply only if the condition is satisfied.
  - **'workflow-chain'** is a sequence of micro-services/rules to be executed by this rule.
  - **'recovery-chain'** are the rules to be called when execution of any one of the rules in the workflow-chain failed.

iRODS: Characteristics

- Server based platform for application development
- Hiding distribution
- Centralised, unified API (unlike web services!)
- Relatively low abstraction level
- Not well suited for legacy systems

Interoperability: 6 vectors on 4 abstraction levels



Entities & Objects

- **Interoperating Entities**
  - Cultural Heritage Institutions (libraries, museums, archives)
  - Digital Libraries,
  - Repositories (institutional and other),
  - eScience/eLearning platforms
- **Objects of Inter-Operation**
  - full content of digital information objects (analogue vs. born digital),
  - representations (librarian or other metadata sets),
  - surrogates,
  - functions,
  - services

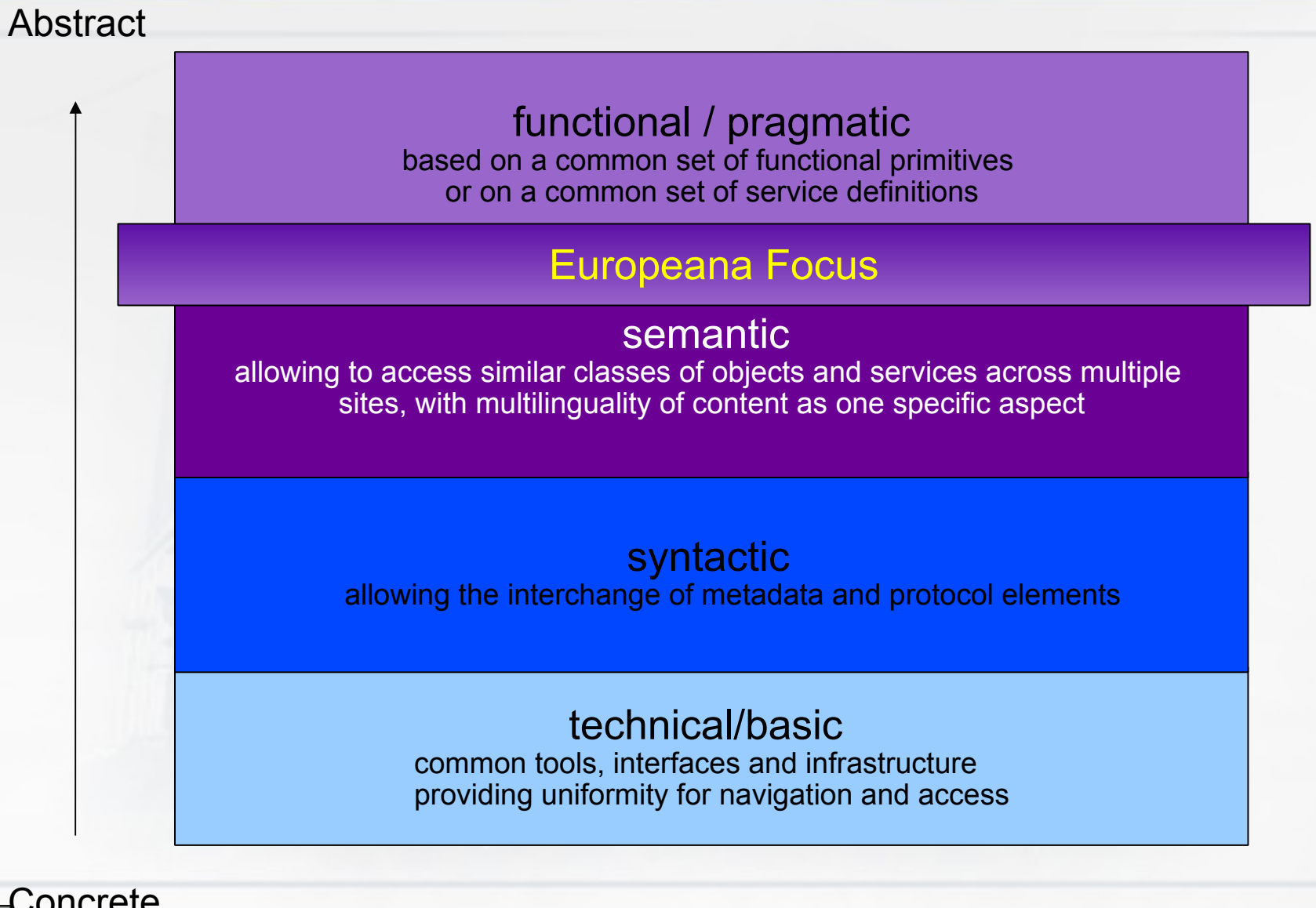
Functionality and Technology

- **Functional Perspective of Interoperation**
  - Exchange and/or propagation of digital content (OA/Non OA)
  - Aggregation of objects into a common content layer (push vs. harvesting / pull)
  - interaction with multiple Digital Libraries via unified interfaces
  - operations across federated autonomous Digital Libraries (such as searching or meta-analysis for e. g. impact evaluation)
  - common service architecture and/or common service definitions or aim at building common portal services.
- **Interoperability Enabling Technology**
  - Z39.50 / SRU+SRW
  - harvesting methods based on OAI-PMH
  - web service based approaches (SOAP/UDDI)
  - Java based API defined in JCR (JSR 170/283)
  - Web crawlers & search engines

Language & People

- **Multilingualism**
  - Multilingual / localised interfaces,
  - Multilingual Object Space
    - dynamic query translation,
    - dynamic translation of metadata or
    - dynamic localisation of digital content.
- **Design and Use Perspective**
  - manager,
  - administrator,
  - end user as consumer or
  - end user as provider of content,
  - content aggregator,
  - a meta user or a
  - policy maker.

Abstraction Layers



The Nasty Bit: Data Quality

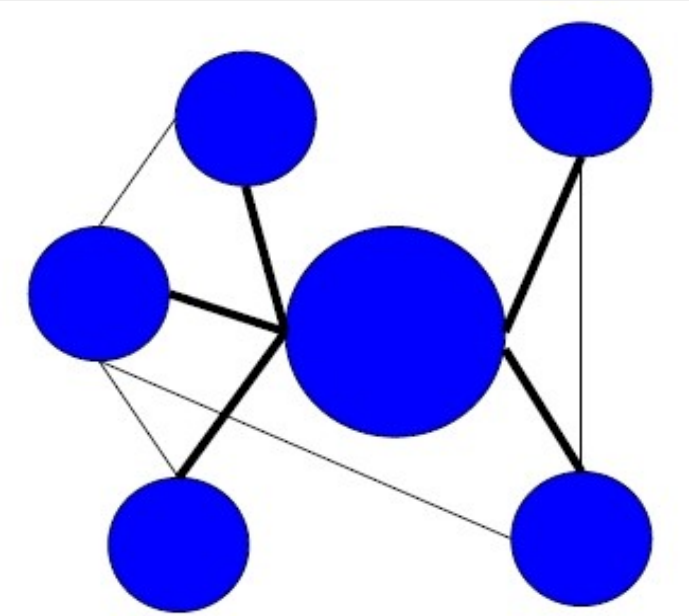
- A perfect framework combining
  - solid object modeling
  - well understood functional primitives
  - including authorisation methods
  - as well as using aligned semantic elements
  - and fully multilingual

  may still result in a dramatic lack of interoperability:
- When operating on 'dirty', heterogeneous data!
- This is a truth both trivial and critical

Interoperability 2.0

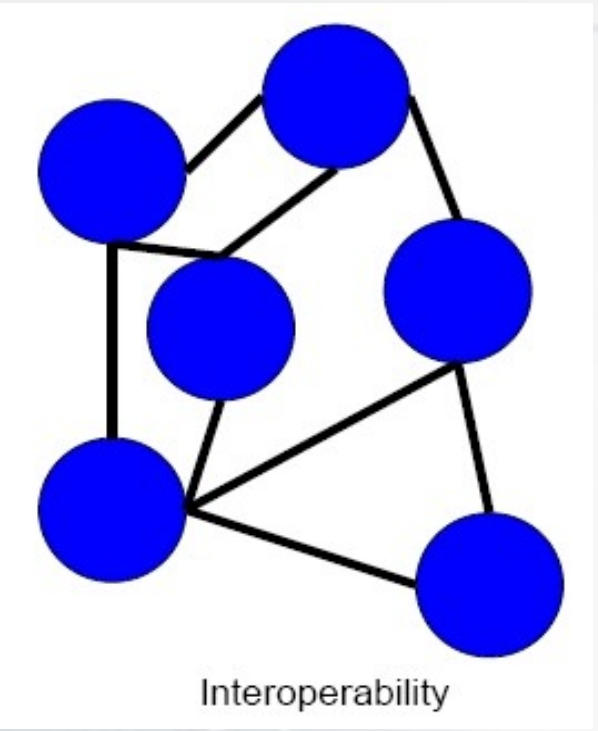
- The 'interoperability' notion has **almost been burnt** by abuse.
- You can even find **„Interoperability 2.0“** in google (although the disaster is limited to the *IMS Global Learning Consortium Learning Tools Interoperability v2.0 Working Group)*
- Term is used in **surprising contexts**: „Microsoft is committed to solving the real-world interoperability challenges of our customers ...“ => “Document Interoperability Initiative” http://www.microsoft.com/interop/principles/default.mspx
- “While the best example of a communications system based on open standards is the Internet, perhaps the best counter-example lies in the proprietary world of the desktop computing environment, which is dominated by Microsoft’s closed operating system (Windows).“ (ECIS, not altogether unbiased!)

Intraoperability



Intraoperability

vs.



Interoperability

- "I think the word **"interoperability"** is being similarly abused. When a single vendor or software provider makes it easier to connect primarily to his or her software, this is more properly called intraoperability." (Bob Sutor)

Back to Interoperability 1.0

- Re: Motivation (ECIS, slightly modified – DL instead of 'devices'):
  - In today's networked ICT environments, DLs do not function purely on their own, but must interact with other DLs.
  - A DL that cannot interoperate with the other products with which consumers expect it to interoperate is essentially worthless.
  - It is interoperability that drives competition on the merits and innovation.
  - The ability of different DLs to interoperate allows consumers to choose among them.
  - Because consumers can choose among them, DLs must compete with one another, and it is this competition that has driven innovation in the software industry.
- => Interoperability enables **Diversity** and **Competition**.
- Therefore let us keep the concept 'clean' and meaningful.
- This is what DL.org basically is about

An Interoperable Europeana

- Europeana = data + functionality + an API exposing both
- Europeana ≠ a portal: it provides a portal based on its own API
- D2.5: „Europeana can be thought of as a network of inter-operating contextualised object surrogates ... **This network in turn is an integral part of the overall information architecture of the WWW**.“ → embed Europeana in Linked Data and in the biggest interoperability space ever created: the WWW!!
  - → **ORE**, **DC** and **SKOS** are used as basic building blocks of the EDM which in turn generates specialisations of these building blocks
  - ORE is used for **defining aggregations**, DC for **assigning properties** to objects and SKOS for **contextualising** these objects.
  - For property modelling EDM must be able to incorporate properties from **more specialized contributor models**.
- Europeana in this sense will be a definite success ***once Google starts using our API***