DL.org

Digital Library Interoperability,
Best Practices and Modelling Foundations

Content domain

Leonardo Candela, Paolo Manghi, Carlo Meghini

5th October 2010

DL.org Autumn School – Athens, 3-8 October 2010

DL.org
Digital Library Interoperability, Best Practices and Modelling Foundations

Lecture outline

- What is the Content
- Content domain in the Reference Model
- Content domain interoperability: Issues and Approaches
- Content domain interoperability approaches: DLSs Federations
- Content domain interoperability approaches: Europeana Data Model
- Hands-on Time

DL.org
Digital Library Interoperability, Best Practices and Modelling Foundations

What is [the "Digital Library"] Content

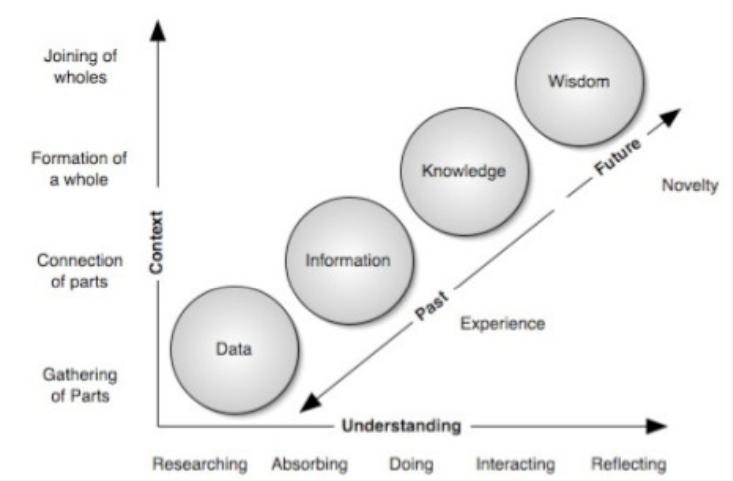
What is [the "Digital Library"] Content?

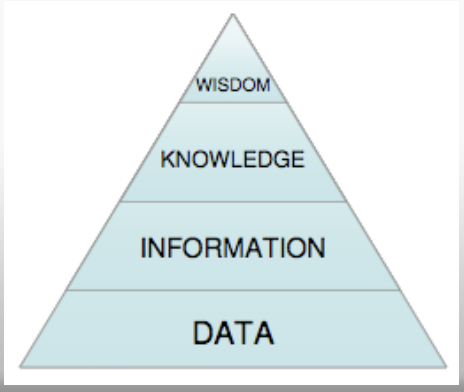
- Digital Library are tools requested to support intellectual activity having no logical, conceptual, physical, temporal or personal borders or barriers on information
  - From a content-centric system to a person-centric system
  - From static storage and retrieval of information to facilitation of communication, collaboration and other forms of interaction among scientists, researchers or the general public
  - From handling mostly centrally located text to synthesising distributed multimedia document collections, sensor data, mobile information and pervasive computing services

From data to wisdom a.k.a. the DIKW hierarchy

- **data** = raw facts
- **information** = processed data, connected data
- **knowledge** = application of information, appropriate collection of information
- **wisdom** = processed knowledge

www.systems-thinking.org





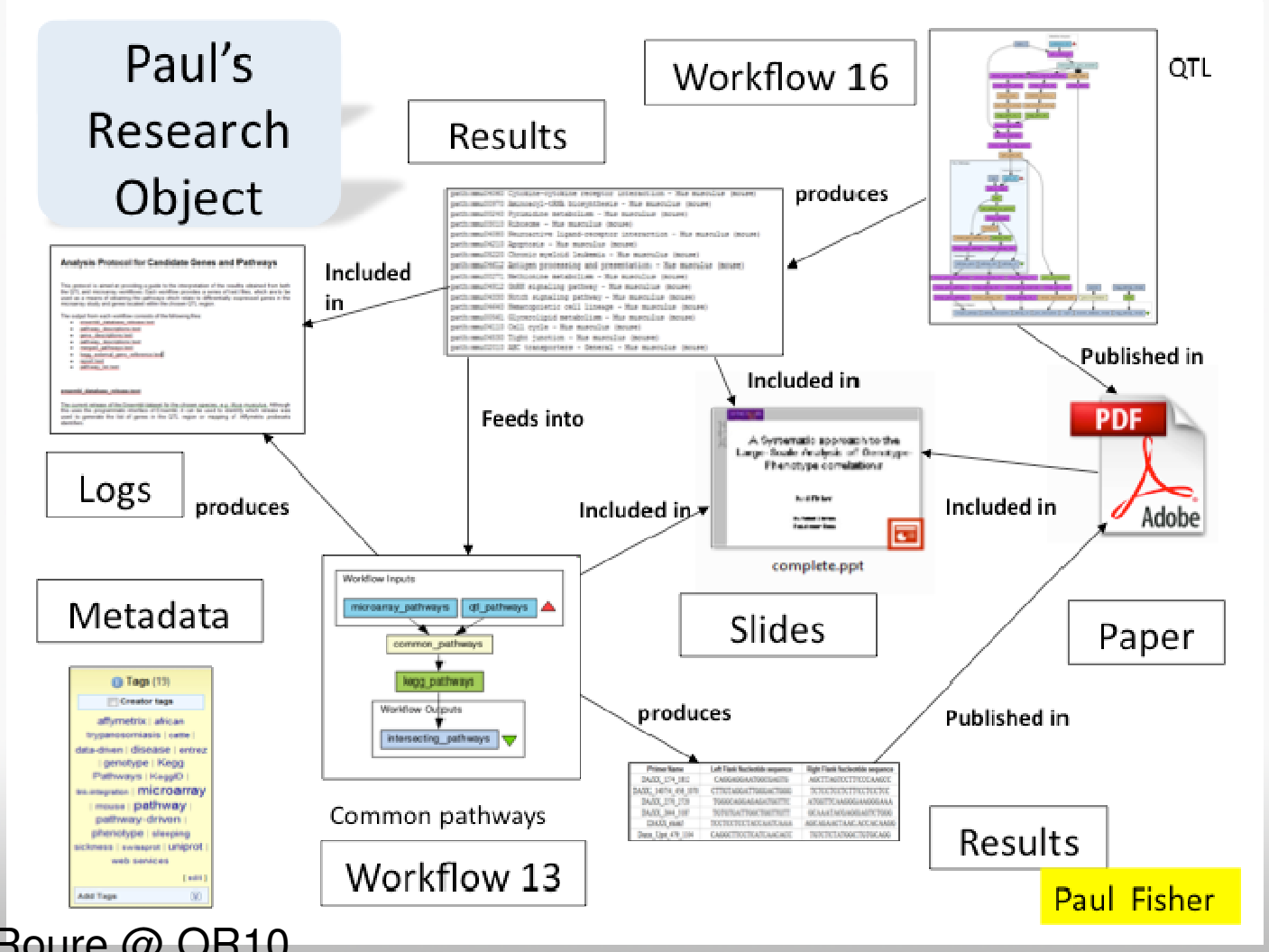
“Papers” today

Aggregations

1. The URI of the human start page for the arXiv document.
2. The formats in which the document is available: constituents of the aggregation.
3. The title of the document.
4. The authors of the document.
5. The creation and last modification date of the document.
6. Identifiers of entities that are in some manner equivalent to this document. For example, the DOI of a peer-reviewed article.
7. The versions of this document.
8. Links to other arXiv documents in the same collection.
9. Citations made by this document, and citations it received from other documents.

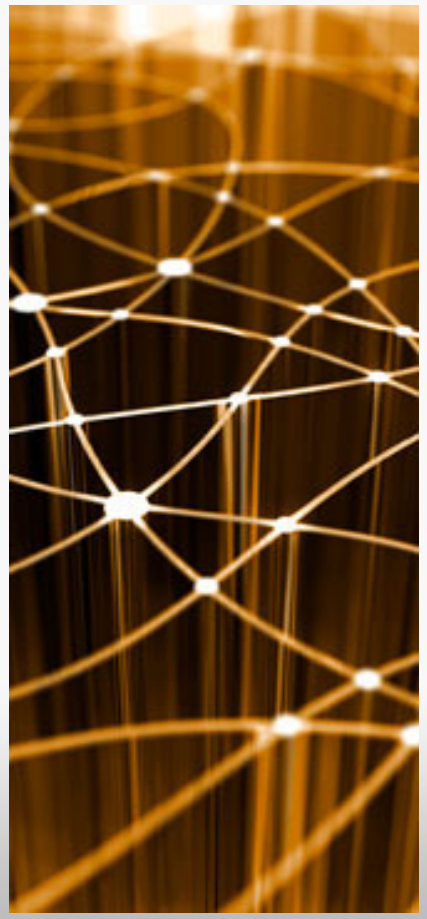
OAI Object Reuse & Exchange
Herbert Van de Sompel & Michael L. Nelson
Woodruff Library, Emory University, October 1 2009

eScience publications



David De Roure @ OR10

DL.org
Digital Library Interoperability, Best Practices and Modelling Foundations



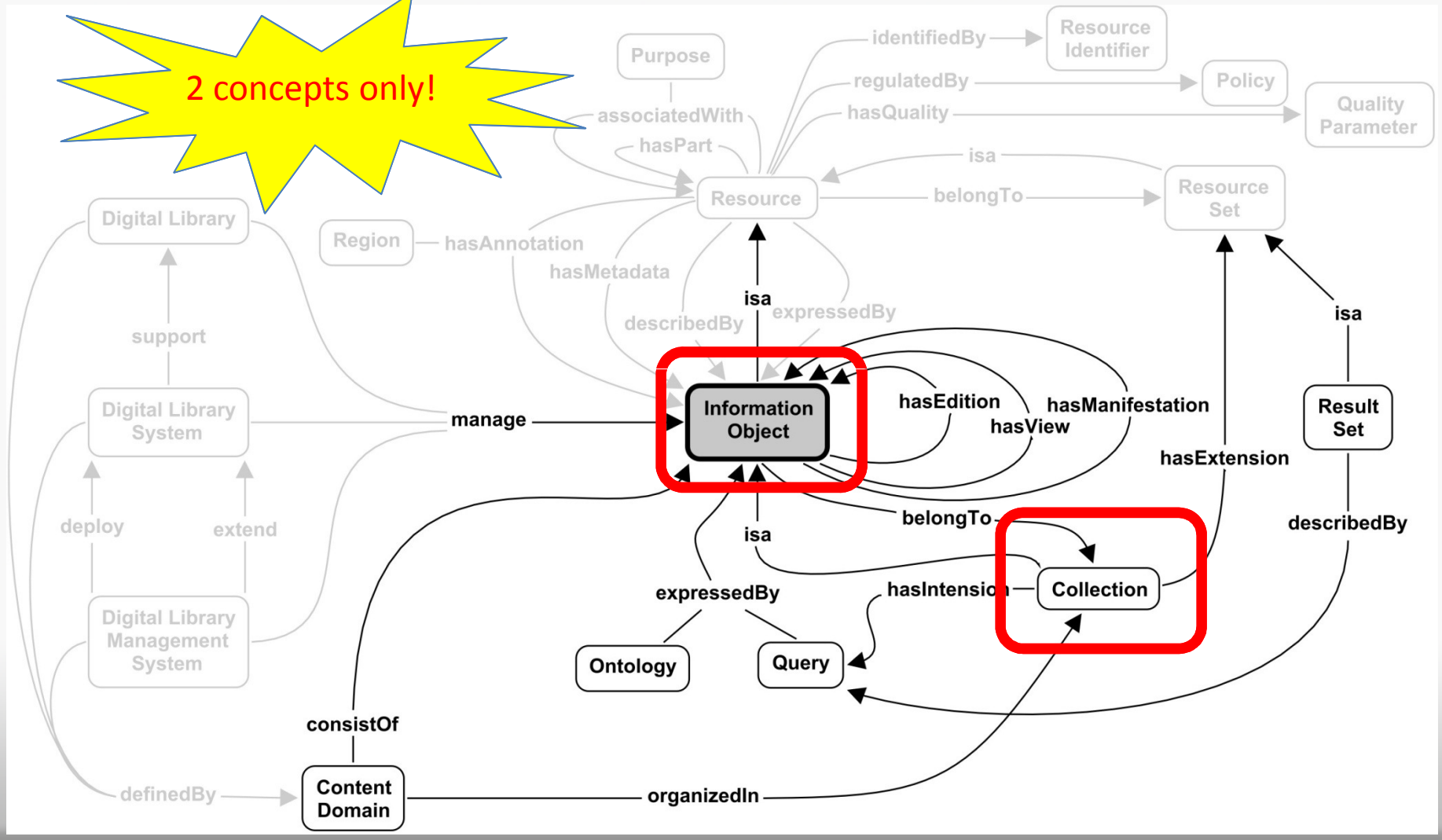
Content Domain: the Reference Model

Content Domain

One of the six main concepts characterising the Digital Library universe. It represents the various aspects related to the modelling of information managed in the Digital Library universe to serve the information needs of the *Actors*.

- Encompasses the data and information that the Digital Library handles and makes available to its users
- Encompasses the diverse range of information objects, including such resources as objects, annotations and metadata
- It is composed of a set of information objects organised in collections

Content Domain: the map



DL.org
Digital Library Interoperability, Best Practices and Modelling Foundations

Information Object

The main *Resource* of the *Content Domain*. An *Information Object* is a *Resource* identified by a *Resource Identifier*. It must belong to at least one *Collection*. It may have *Metadata*, *Annotations* and multiple *Editions*, *Views*, *Manifestations,* which are also represented as *Information Objects*. In addition, it may have *Quality Parameters* and *Policies*.

Information Object (cont.)

As an *Information Object* is a *Resource*, it inherits all its features

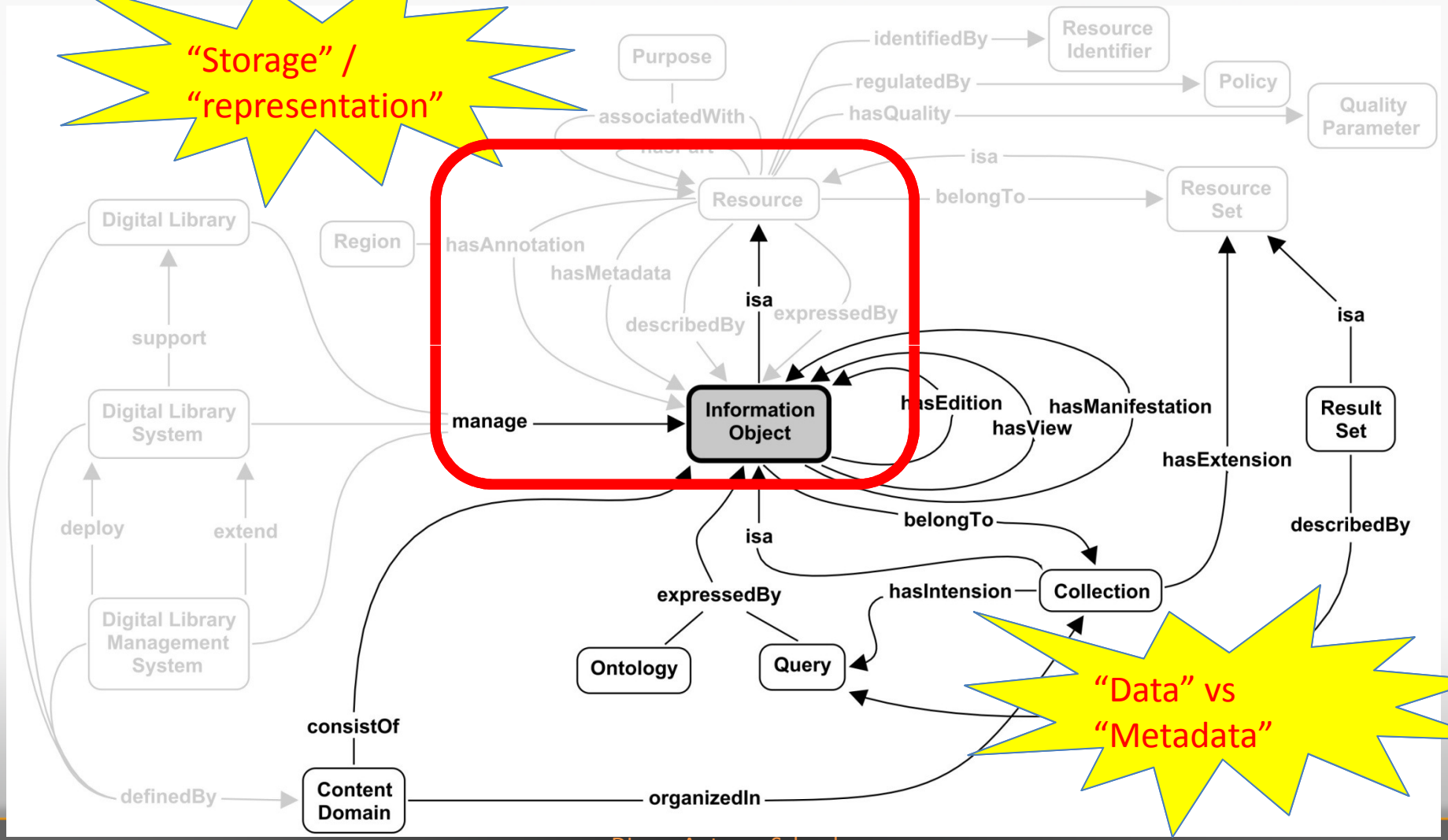
- has a unique identifier (Resource Identifier) also known as the information object identifier;
- is arranged according to a format (*Resource Format*) also known as the document model;
- can arbitrarily be composed (<*hasPart*> and <*associatedWith*>) to capture compound artefacts;
- is characterised by various *Quality Parameters* each capturing different object quality facets (<*hasQuality*>);
- is regulated by *Policies* (<*regulatedBy*>) governing every aspect of its lifetime; and
- can be described or augmented by ***Metadata*** (<*hasMetadata*>) and ***Annotations*** (<*hasAnnotation*>)

Collection

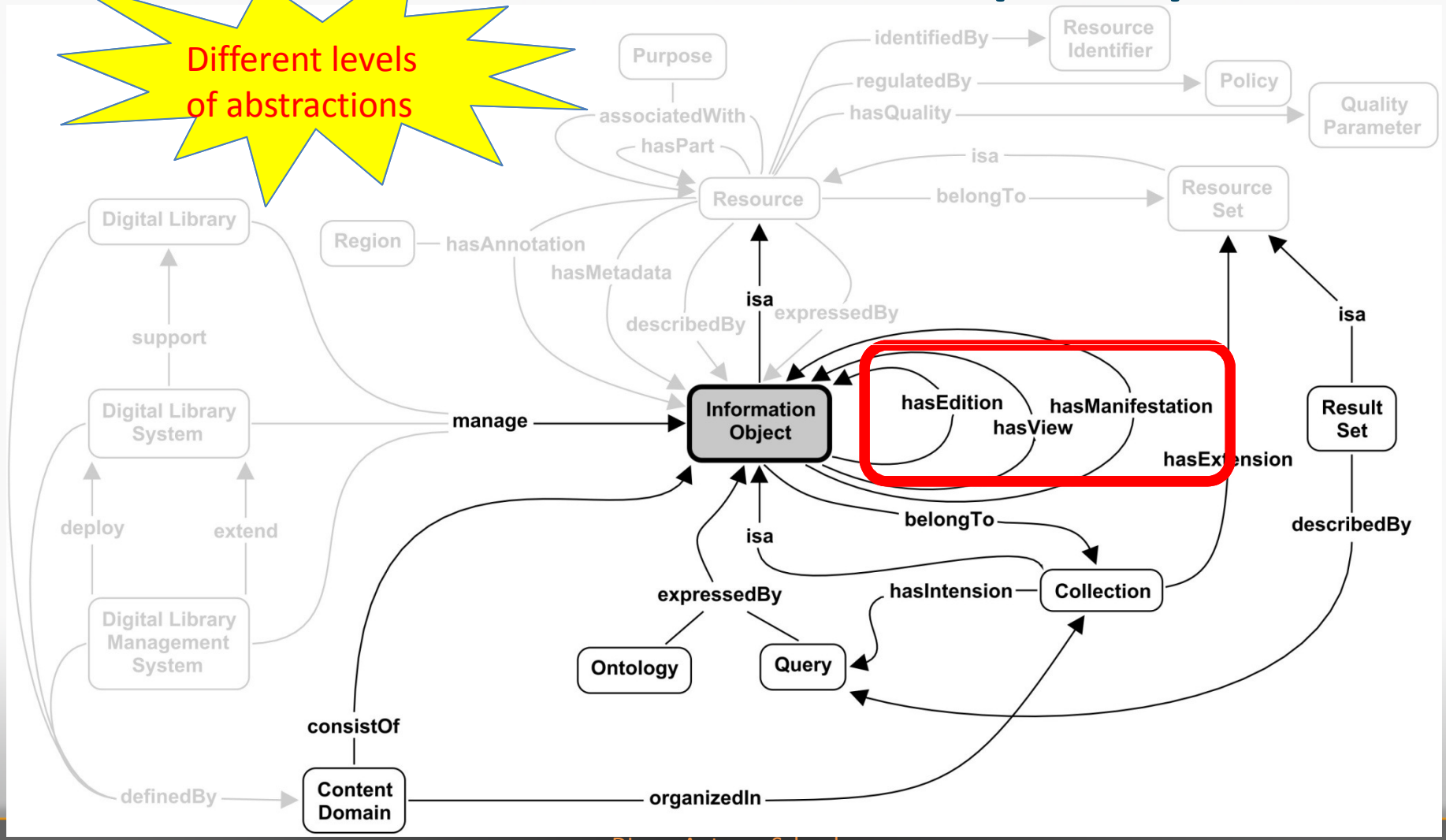
A content *Resource Set*. The 'extension' of a collection consists of the *Information Objects* it contains. A *Collection* may be defined by a membership criterion, which is the 'intension' of the collection.

- A Collection is an Information Object thus a Resource!

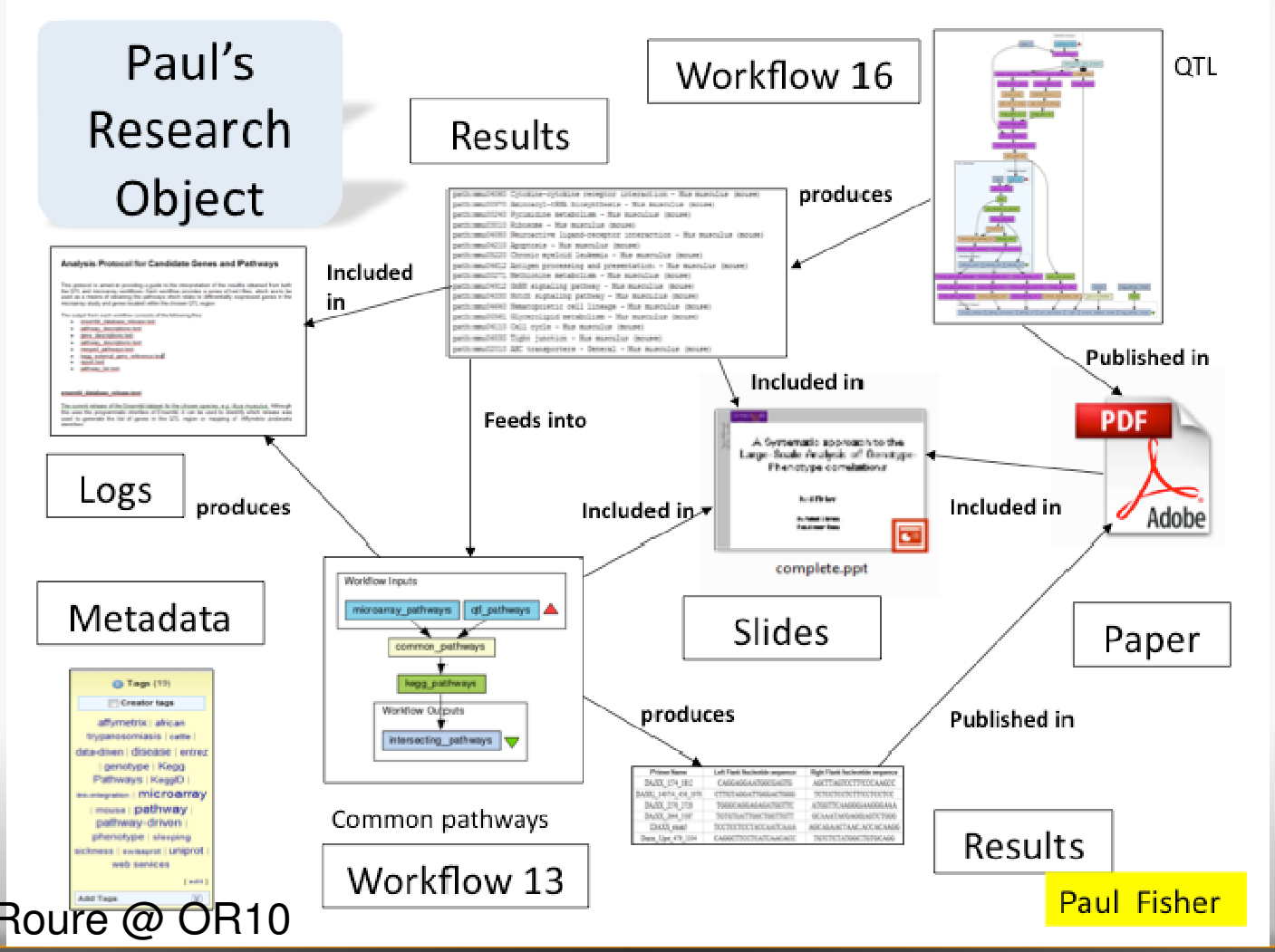
Content Domain: where is the rest?



Content Domain: where is the rest? (cont.)

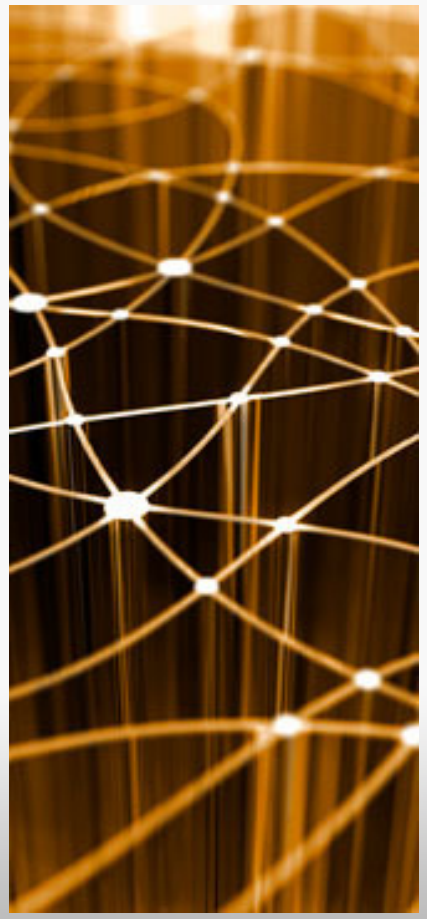


Do we have enough constructs?



David De Roure @ OR10

DL.org
Digital Library Interoperability, Best Practices and Modelling Foundations



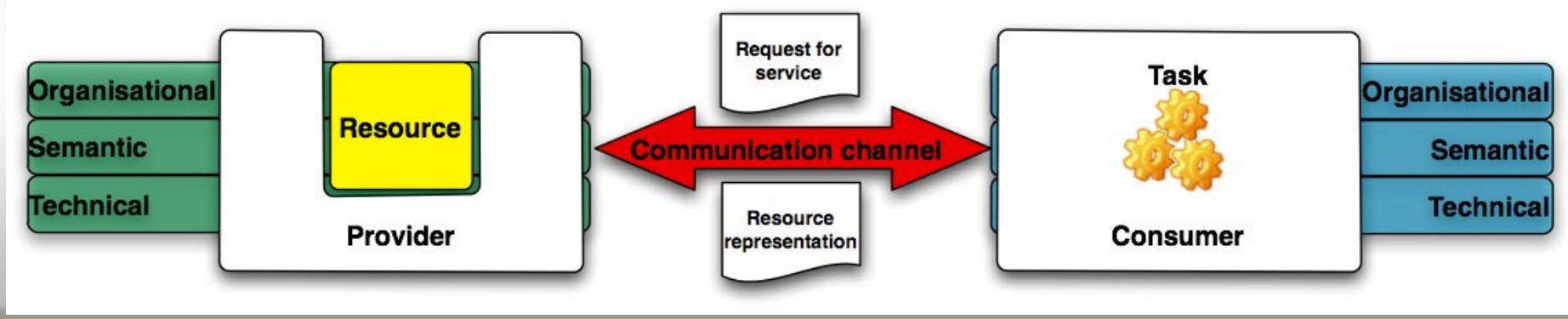
Content Domain Interoperability: Main Issues & Approaches

The Content Interoperability "monster"

- in G. Anthes, "Happy Birthday, RDBMS!", CACM, Vol. 53(5), May 2010
  - "… integration of heterogeneous data. "A special case that is still really hard is schema mapping — converting data from one format to another," ... "It sounds straightforward, but it's very subtle."
  - "… the "unsolved problem" of querying geographically distributed databases"

[Content] Interoperability Framework

- *"the ability of two or more systems or components to exchange information and to use the information that has been exchanged"* (IEEE, 1990)
- Rephrasing and elaborating
  - at least two entities *Provider* and *Consumer*
  - willing to "share" a *Resource* to perform a *Task* (has preconditions)
  - through a *Communication Channel*, involving a protocol and information representation
  - across *Organizational*, *Semantic* and *Technical* boundaries of entities



Interoperability Framework: focusing on the three key aspects

- *Organizational* deals with business goals and processes of an entity (Provider or Consumer)
- *Semantic* deals with the meaning of the exchanged resource and the rest of information
- *Technical* deals with technological solution supporting the operation of the Provider / Consumer as well as the communication among the two

- Dependencies / constraints among the three boundaries, e.g. organisational aspects has to be implemented by the technical aspects
- Implicit / hidden information, e.g. the technical aspects might implement part of the organisational aspects only
- Different approaches for diverse boundaries, e.g. human-centric vs. machine-centric
- Complete solutions involve all of them, e.g. the decision to rely on a certain technology might be useless if it is not complemented by proper organisational aspects

Interoperability Approaches

- Agreement-based Approaches
  - Include Standard-based approaches
  - Infringes autonomy, strong in effectiveness
- Mediator-based Approaches
  - An intermediary service linking Provider and Consumer
  - Strong in autonomy, development and maintenance cost
- Blending & Compound Approaches
  - Mix & compose
  - Compatibility issue among solutions
  - Alternative solutions
  - No solution exist vs for each problem there exists at least a solution

DL.org
Digital Library Interoperability, Best Practices and Modelling Foundations

How is an interoperability solution described?

1. *Overview*

   Context of the proposed approach including pointers to detailed description of it
2. *Requirements*

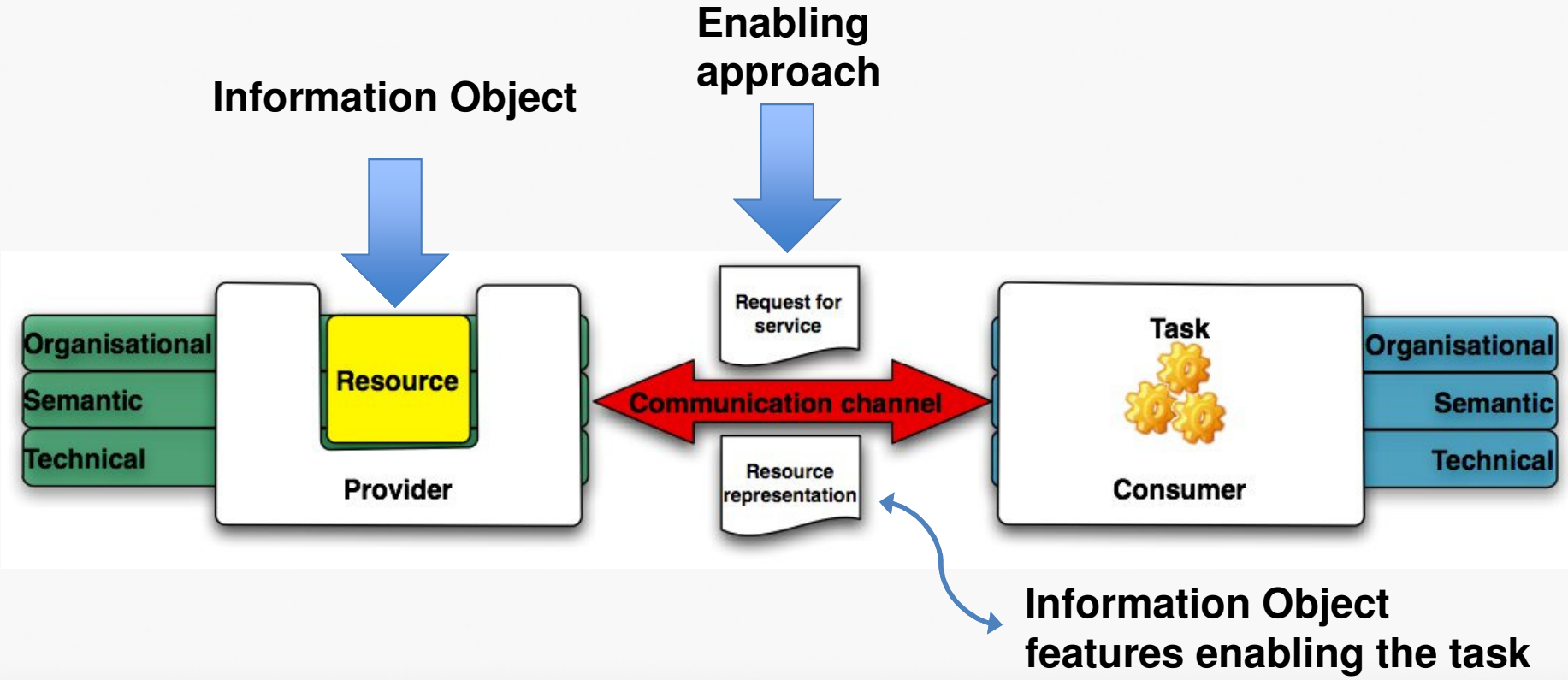
   Conditions under which the solution might be used
3. *Results*

   Changes resulting from the usage of the solution
4. *Implementation guidelines*

   How the changes are produced
5. *Assessment*

   Qualitative evaluation of the proposed solution

Content Interoperability Scenario



Content Interoperability Issues and Solutions

- IO Features: ID, structure, any attribute (e.g. provenance)
- Approaches
  - Agreement-based
    - OAI-PMH, OAI-ORE, DublinCore
  - Mediator-based
    - metadata mapping, Application Profiles
  - Blending

Metadata Schemas and Application Profiles

- Metadata Schemas
  - 1995 DublinCore
    - Identifier, Title, Creator, Contributor, Publisher, Subject, Description, Coverage, Format, Type, Date, Relation, Source, Rights, Language
    - cross-domain
    - simplicity vs precision/completeness -> cost
    - “not one size fits all”
- Application Profiles
  - metadata schema defined by combining elements from existing schemes

OAI-PMH

- A 2001 simple protocol for metadata harvesting
  - metadata centric (?!?!)
  - data provider vs service provider
  - six verbs (Identify, ListMetadataFormats, ListSets, ListRecords, ListIdentifiers, GetRecord)
  - http & XML

http://www.openarchives.org/pmh

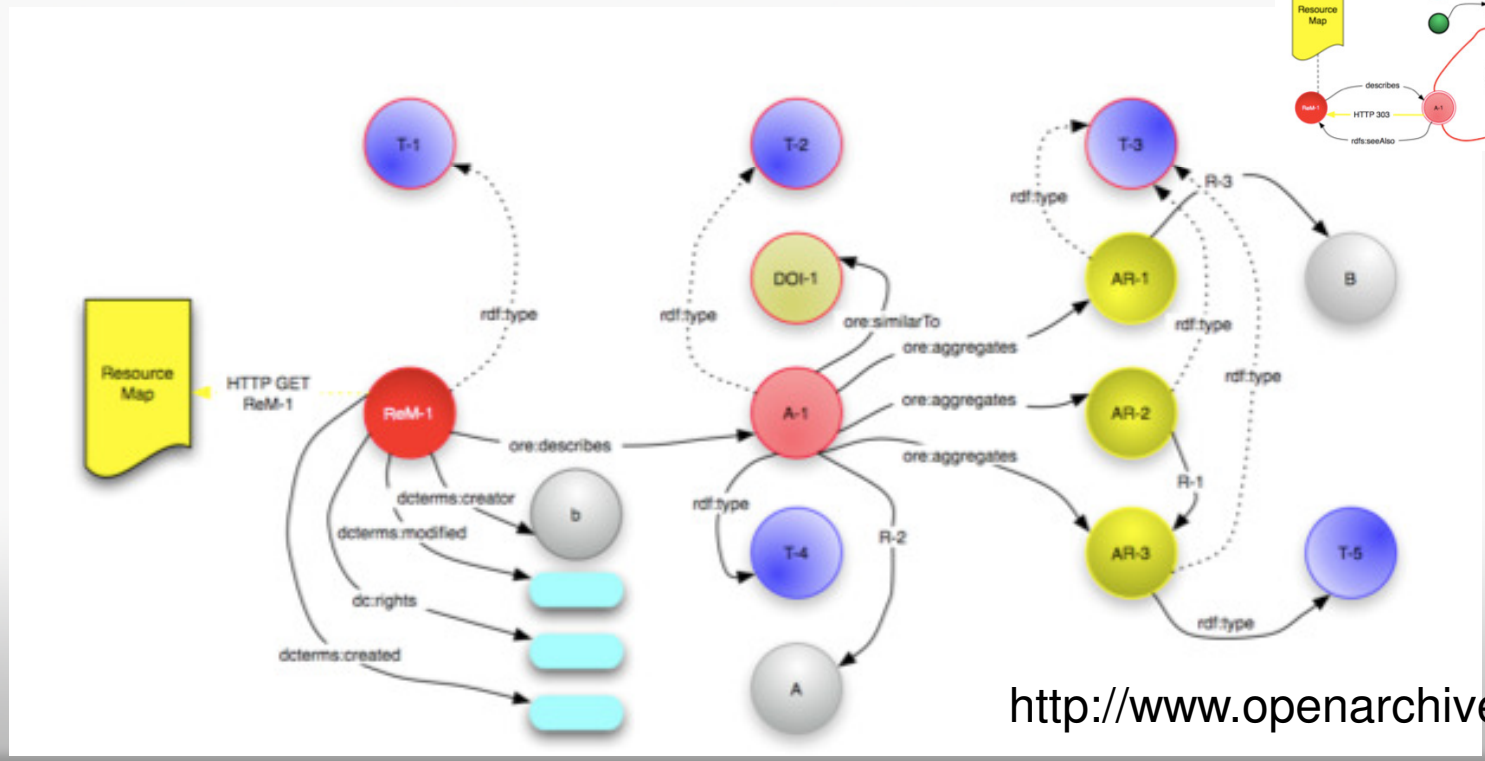
Linked Data

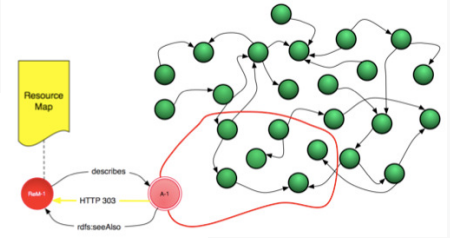
- A 2006 (?) set of principles for publishing structured data on the Web
  - resource - an item of interest
  - URI - global identifier for a resource
  - representation - data corresponding to the state of a resource
  - information resource - a “document” containing information
  - non-information resource - anything else
  - associated description - representation describing a Semantic Web resource

http://www.linkeddata.org

OAI-ORE

- A 2008 approach for Aggregated Resources
  - web principles adherence



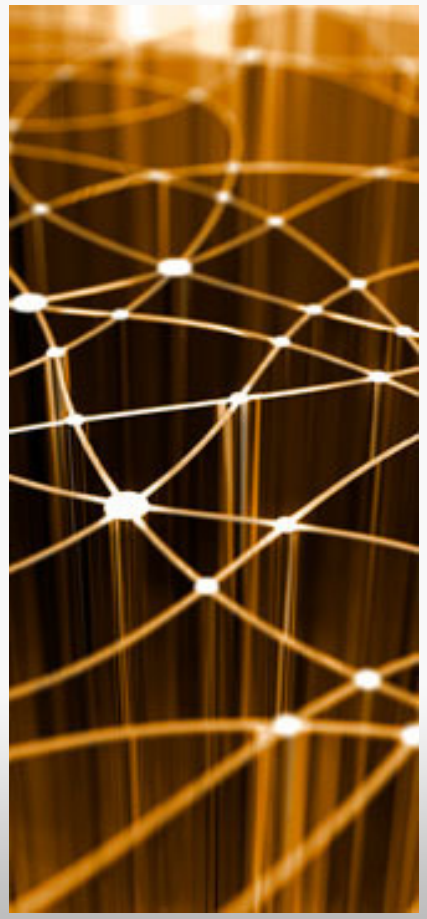


http://www.openarchives.org/ore

Lesson learned

- "Open" / "Publish" Comprehensive Information Objects (features) => "recall"-orientation
  - not implies "free" (policies are always there)
  - machine-orientation
- Web-orientation and standards
  - no new / custom protocol
  - no complex protocol
- Information Objects semantic & data quality are yet the hard problem (but mitigated by shared agreements)

DL.org
Digital Library Interoperability, Best Practices and Modelling Foundations



Content Domain: Interoperability Solutions "Interoperability Patters in Digital Library Systems Federations"

Paolo Manghi

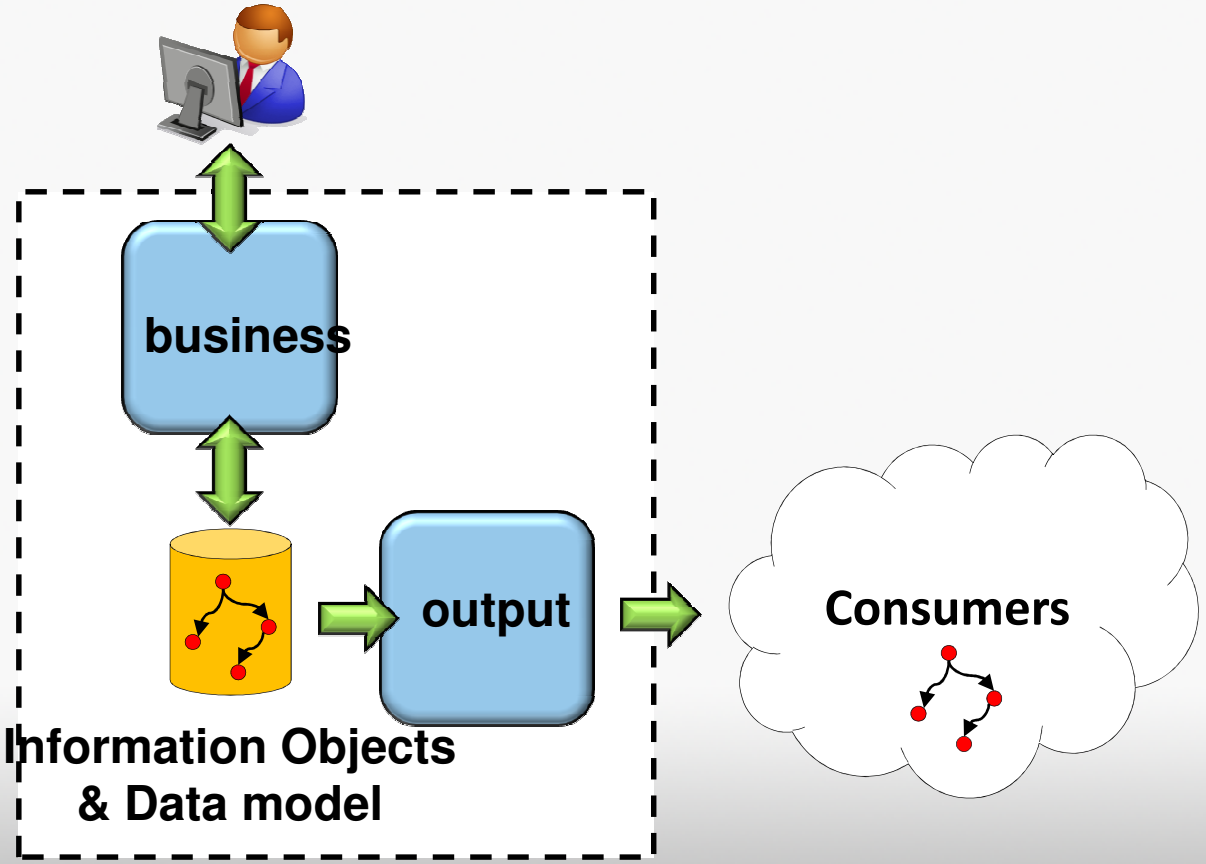
http://mc.gunet.gr/live/dlorg.php

DL.org
Digital Library Interoperability, Best Practices and Modelling Foundations

Outline

- Digital Library System Federations
- Interoperability issues
- Data impedance mismatch
  - Structural, semantic and granularity mismatch
- Solution: D-NET Software Toolkit

Digital Library Systems



Digital Library Systems Federations (DLSFs)

- Motivations
  - On-line availability of “fragmented” research outcomes
  - Multidisciplinary character of modern research
  - Increased speed of research life-cycle, i.e., immediate availability and access to research outcome
  - Others…

DLSFs

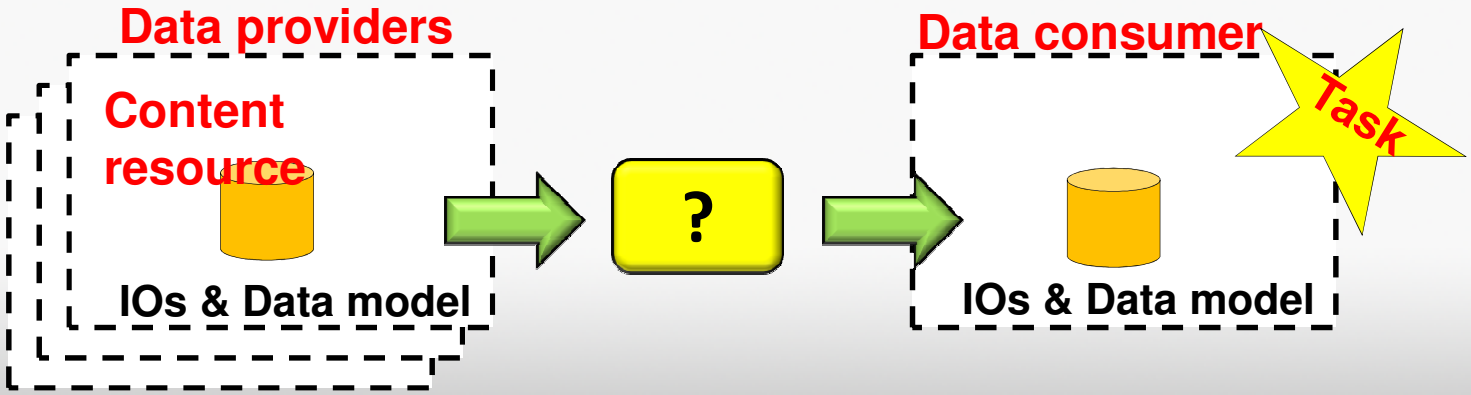
- OAI-PMH archive/libraries/repository federations
  - e.g., Europeana, OCLC-OAIster, BASE, NARCIS
- Community-oriented data infrastructures
  - e.g., DRIVER, SAPIR, CLARIN, EFG, HOPE, D4Science

DLSFs and the DL.org interoperability framework

- Providers = Digital Library Systems or Data Providers
- Consumer = Service Provider, software system specially devised for
  - Collecting input content resources (information objects, e.g., metadata, payloads, compound objects) from a set of data providers
  - From input information objects, producing a uniform "information space" of output information objects, required by the consumer to perform a given task

DLSFs and the DL.org interoperability framework

- Providers = Digital Library Systems or Data Providers
- Consumer = Service Provider



DLSFs: content interoperability

- "Obstacles" encountered by a data provider (DLS) willing to offer useful information objects to a service provider to accomplish its task
- "Obstacles" encountered by a service provider willing to accomplish its task by accessing the information objects of a data provider which it considers useful

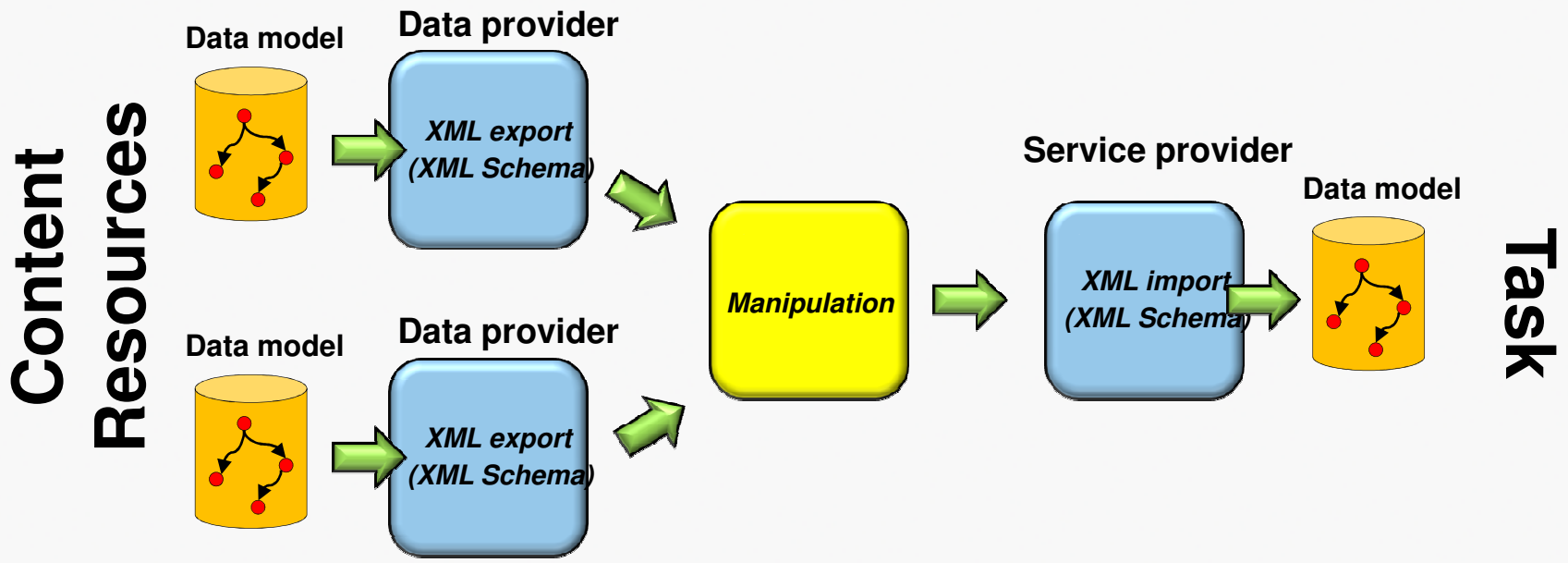
DLSFs: content interoperability issues

- Low-level issues: “How to exchange objects”
  - Identifying common on-the-wire data-exchange practices
- High-level issues: “How to harmonize information objects data models”
  - Resolve data impedance mismatch problems arising from distinct data models of data and service providers

Low-level issues: "How to exchange information objects"

- Adoption of XML as lingua-franca and standard data-exchange protocols, e.g., OAI-PMH, OAI-ORE, ODBC
  - XML schema for data model
  - Data providers implement exporting components: information objects → XML files
  - Service provider implement importing component: XML files → information objects
- Worth noticing:
  - Equal data models does not mean equal XML schemas
  - Data and service providers may manage information objects as XML files (e.g., native XML DBs)

High-level issues: data impedance mismatch




```
<Article>
        <Title> "Interoperabilty patterns…"
        </Title>
        <Authors> "Paolo Manghi, Leonardo Candela…"
        </Authors>
        <Date > "September 2010"
        </Date>
</Article>
```

Definitions:

- Schema *path*: **Article.Title**
- Schema *leaf*: **"September 2010"**

High-level issues: data impedance mismatch

- Data model impedance mismatch
  - Data and service providers XML schemas do not match, either structurally (schema paths) or semantically (schema leaves)
- Granularity impedance mismatch
  - XML encodings of information objects at the service provider and data providers adopt different levels of granularity.

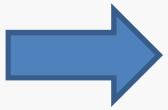
Structural heterogeneity
(Data model impedance mismatch)

Data provider

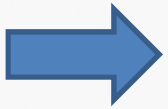
Article
- Title
- Authors
- Date

Article
- Title
- Authors
- Date

Loss



Casting



Service provider

Article
- Title
- Authors

Article
- Title
- Creators
- DateOfCreation

Semantic heterogeneity
(Data model impedance mismatch)

Data provider

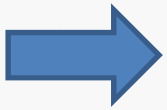
Article

Title "Interoperability…"

Authors "Paolo Manghi, …"

Date "September 2010"

Formats



Service provider

Article

Title "Interoperability…"

Authors "Manghi, P., …"

Date "01-09-2010"

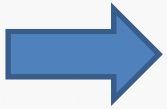
Article

Title "Interoperability…"

Authors "Paolo Manghi, …"

Date "September 2010"

Dervation/ Inference



Article

Title "Interoperability…"

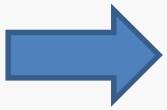
Authors "Paolo Manghi, …"

Date "September 2010"

TitleLanguage "EN"

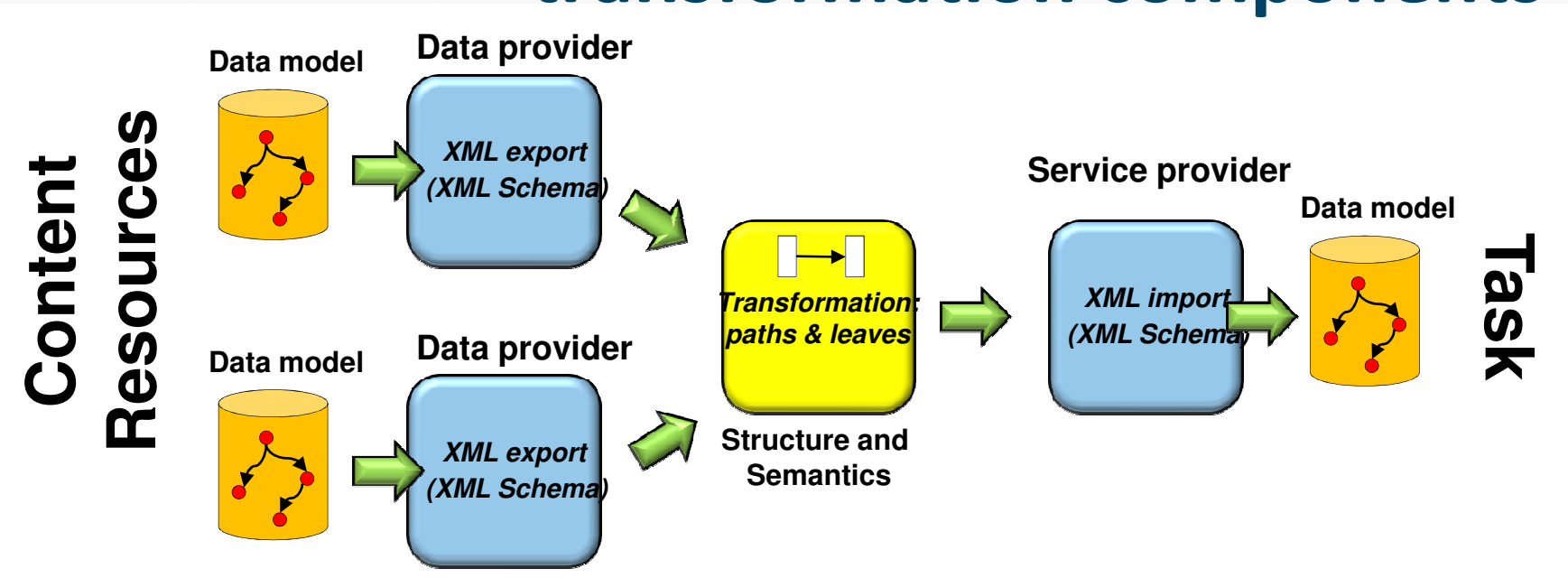
Semantic&Structural heterogeneity (Data model impedance mismatch)

Article
- Title "Interoperability…"
- Authors "Paolo Manghi, …"
- Date "September 2010"



Article
- Title "Interoperability…"
- Creator
  - Name "Paolo"
  - Surname "Manghi"
- Creator
  - Name "Leonardo"
  - Surname "Candela"
- Date "September 2010"

Tackling the data model impedance mismatch: transformation components



Use-cases:

- All data providers have the same XML schema
- Data providers have different XML schemas

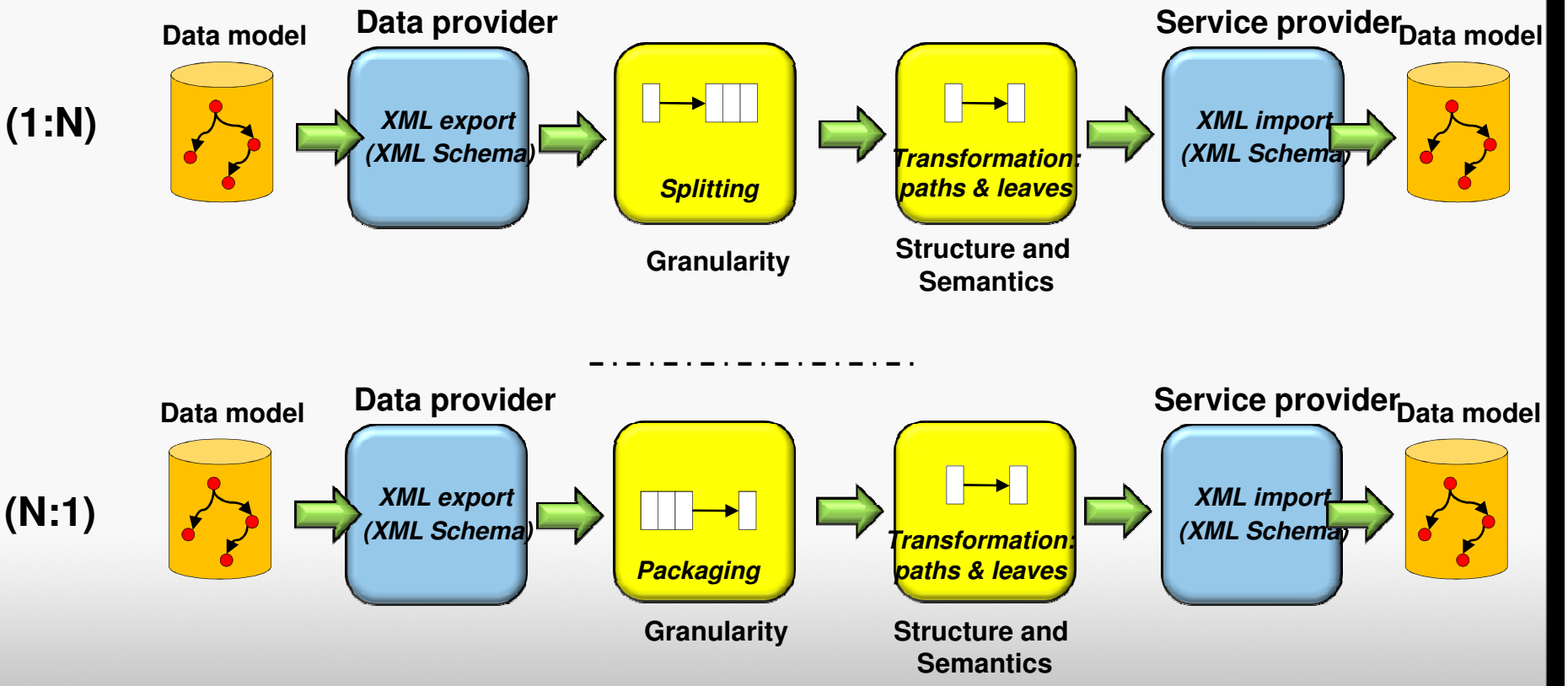
Data providers with equal XML schema (Data model impedance mismatch solutions)

- The transformation component considers one mapping from such common XML schema onto the service provider schema
  - Output schema leaves (identified by output schema paths) are generated by processing input leaves (identified by schema paths) through transformation functions F
- The complexity of the F's can be arbitrary:
  - feature extraction functions: taking a URL, downloading the file (e.g., HTML, PDF, JPG) and returning content extracted from it
  - conversion functions: translation from vocabulary to vocabulary
  - transcoding functions: leaf format to leaf format (e.g., date formats);
  - regular expression: generating one leaf from a set of leaves (e.g., generating a person name leaf by concatenating name and surname originally kept in two distinct leaves).

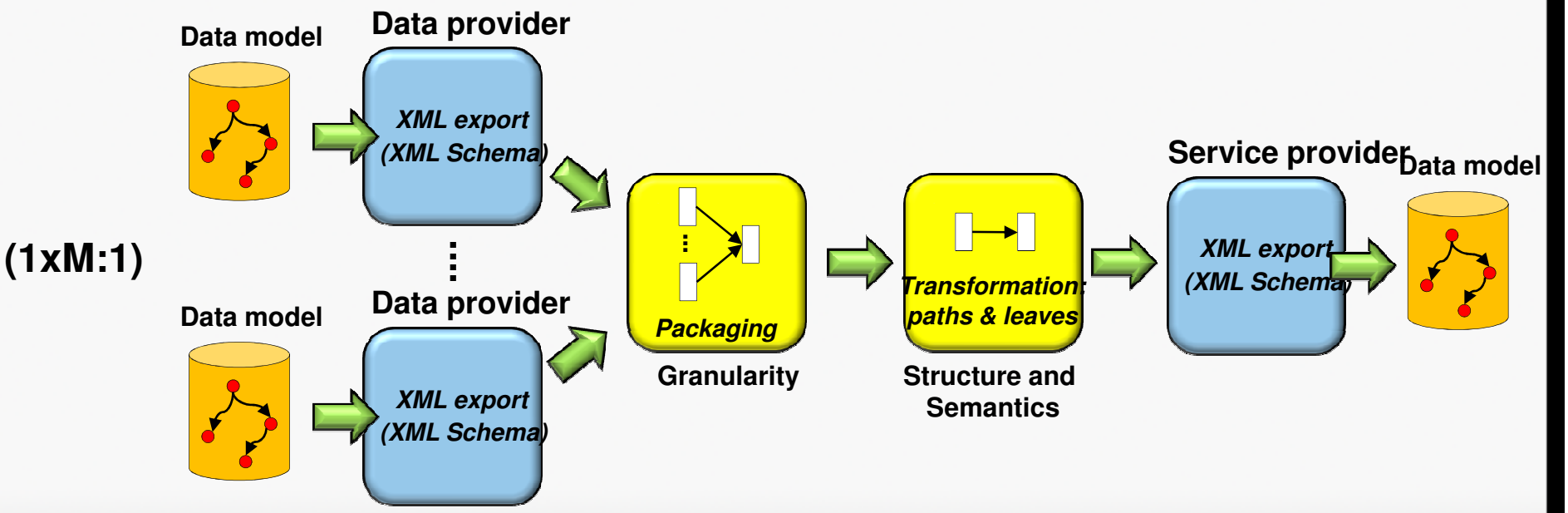
Data providers with different XML schemas (Data model impedance mismatch solutions)

- The transformation component must consider multiple mappings from the diverse input XML schemas onto the service provider XML schema of the service provider
- Simple scenario: pre-determined set of data providers
  - Providing one transformation component as the one described for the previous scenario for each set of data providers with the same schema
- Complex scenario: undetermined number of data providers is expected, possibly bearing different XML schema
  - Providing general-purpose components, capable of managing (create, remove, update) a set of “mappings”
  - Mappings are named lists of pairs (input paths, F, output path)
  - The component may allow for the addition of new F’s

Granularity impedance mismatch



Granularity impedance mismatch



Architecture of interoperability solutions

- "Bottom-up" federations, e.g., DAREnet-NARCIS,
  - Realized by organizations who have control over the set of participating data providers,
  - Agree on common data model and XML schema so that no interoperability issues occur
- "Open" federations, e.g., the DRIVER repository infrastructure, OpenAIRE system, Europeana
  - Federations "attractive" to data providers, which are willing to adhere to given "data model" specifications ("guidelines") in order to join the aggregation
  - Transformation: data providers are responsible of structural interoperability (typically light-weight transformation issues); semantics interoperability is typically responsibility of service provider
  - Packaging/splitting not required

Architecture of interoperability solutions

- "Community-oriented" federations, e.g., the European Film Gateway project
  - Data providers handling the same typology of content invest on the realization of a service provider to enable cross-provider functionality
  - Define a common data model on the service provider
  - Packaging/splitting: if needed, typically occurs at the service provider side
  - Transformation: may occur at the data provider side (before XML export takes place) or data providers are directly involved in the definition of mappings on the service provider
- "Top-down" federations, e.g., OAIster-OCLC project, BASE search engine
  - Realized by organizations willing to deliver a service provider to offer functionality over data providers whose content is openly reachable.
  - Service provider deals with any interoperability issues

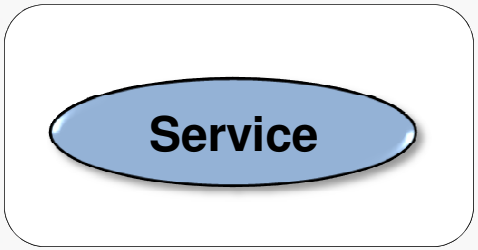
D-NET Software Toolkit

D-NET Software Toolkit: general-purpose DLCLs

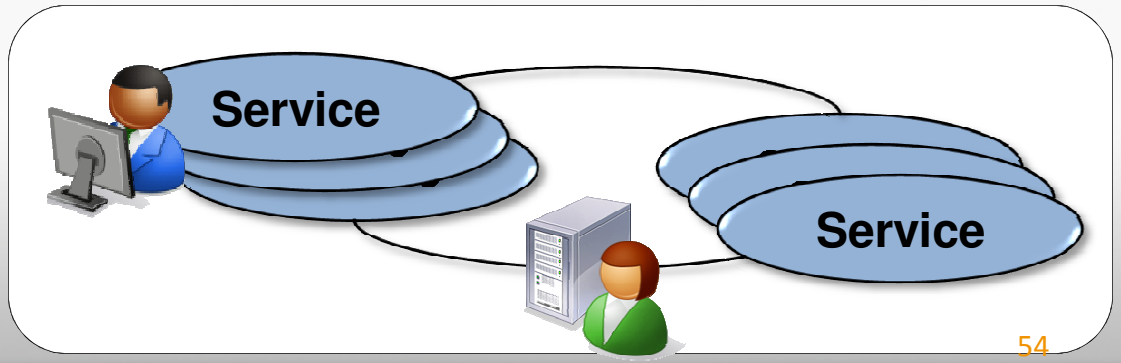
- General-purpose framework for the realization and maintenance of context-specific DLCLs
- Management of information objects of arbitrary data models
- Management of DLSs of several typologies (e.g., OAI, ODBC, FTP)
- Construction of personalized and automated data workflows
- Management of robustness and scalability parameters
- DLSs life-cycle administration tools
- Extensibility with new functionality

D-NET Software Toolkit
The solution…

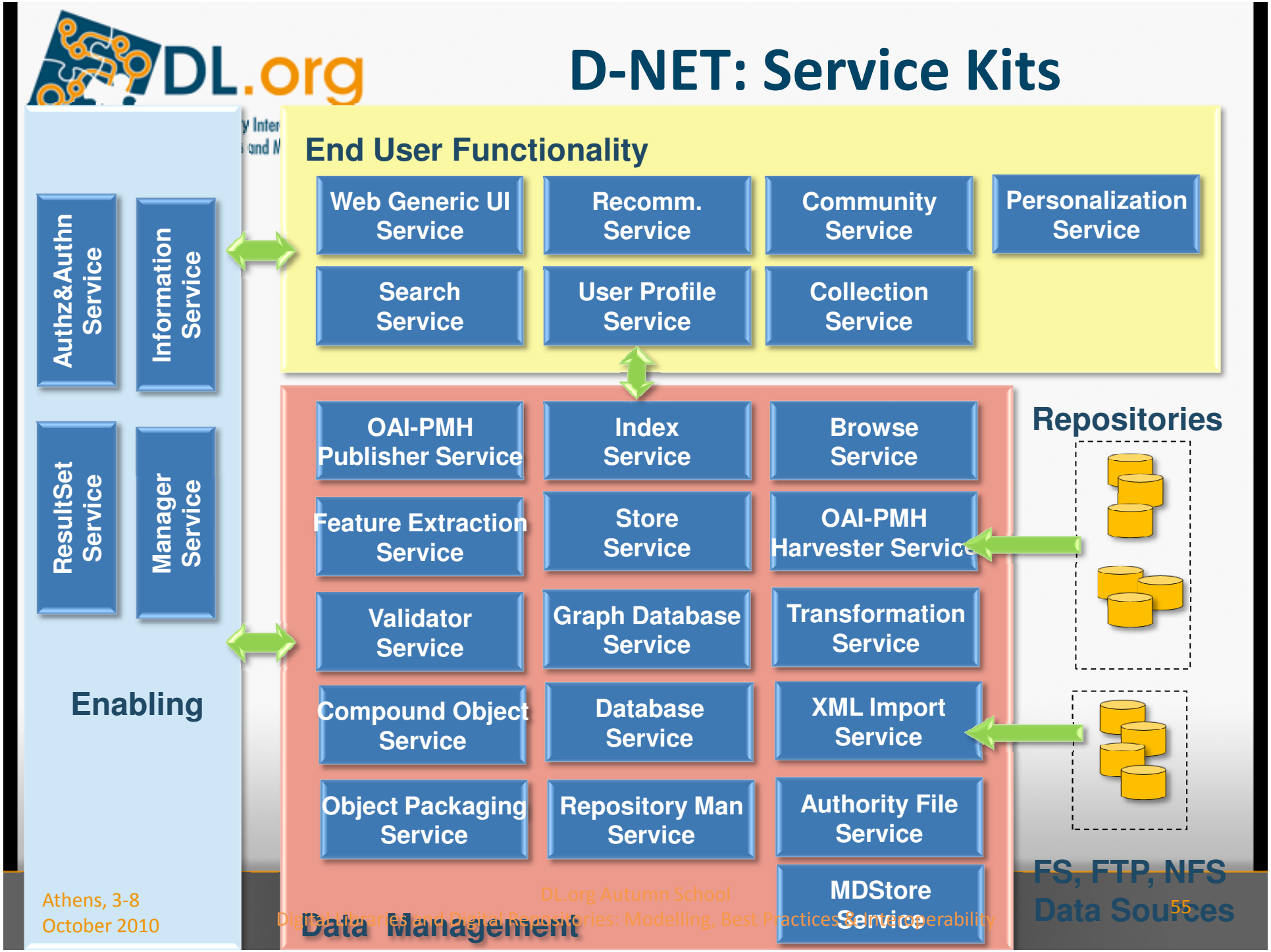
- Service Kits supporting realization of “personalized” DLSFs by exploiting customizability, extensibility and modularity features



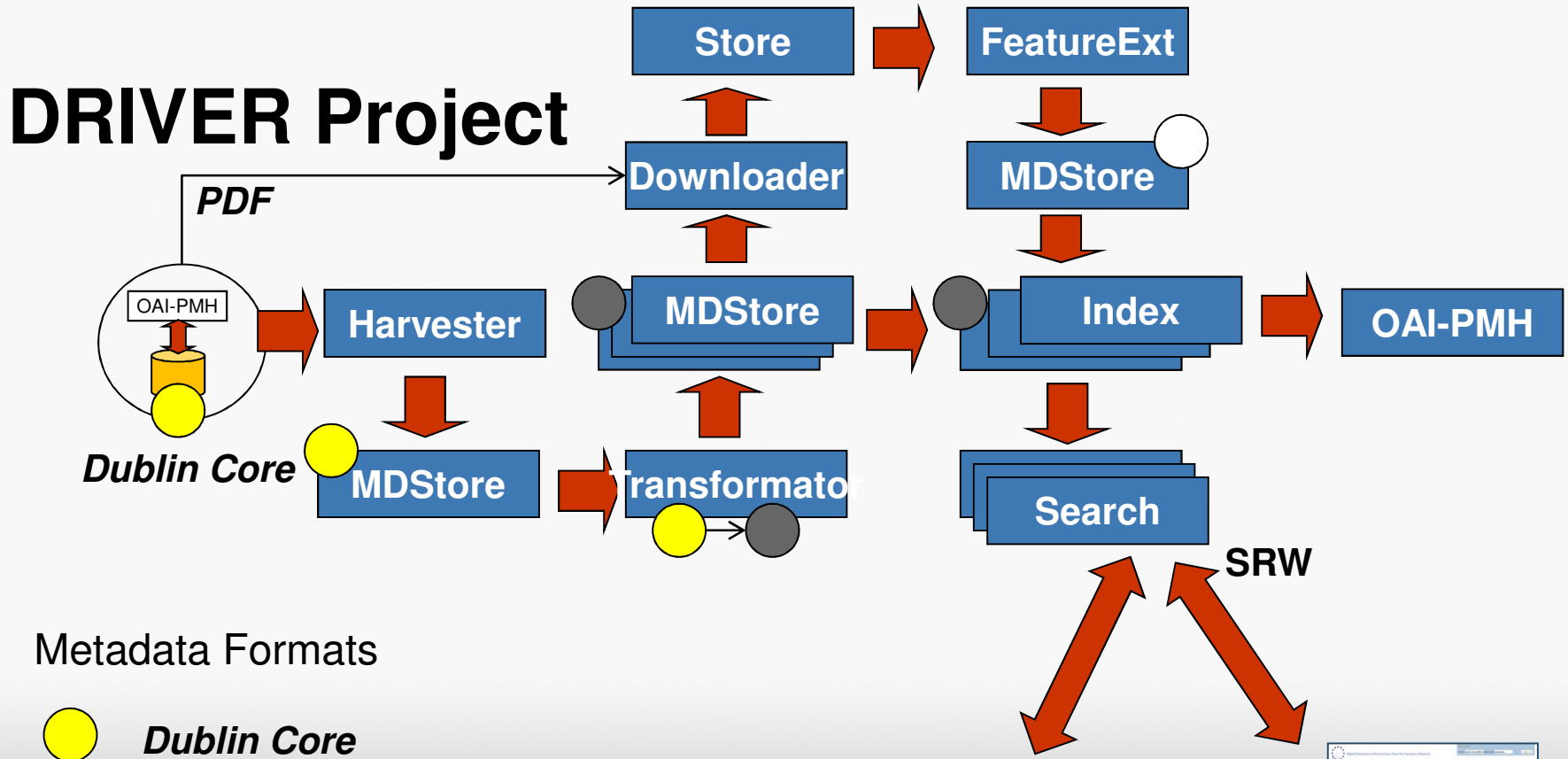
- Service-oriented infrastructure features (autonomicity, distribution and sharing) to support scalable and robust production systems



D-NET: Service Kits

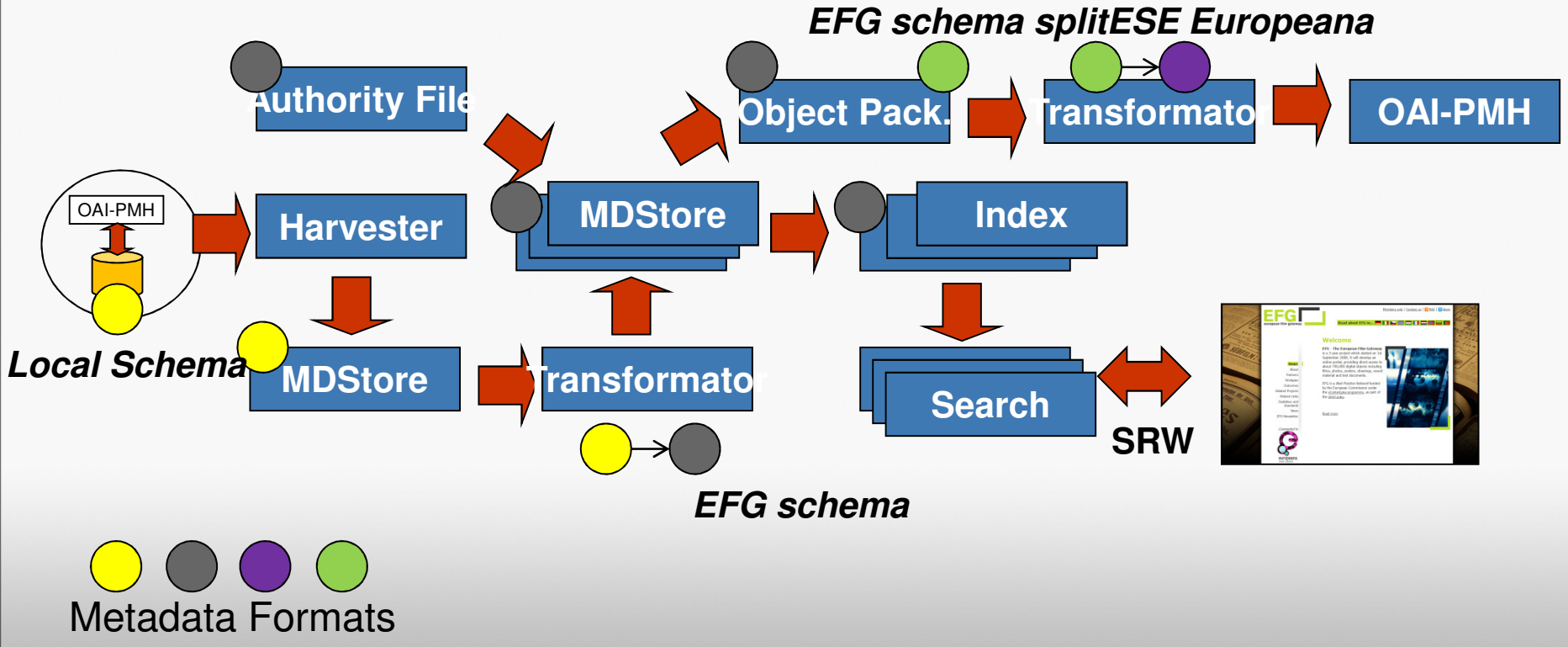


Modularity, customizability, sharing (and orchestration)



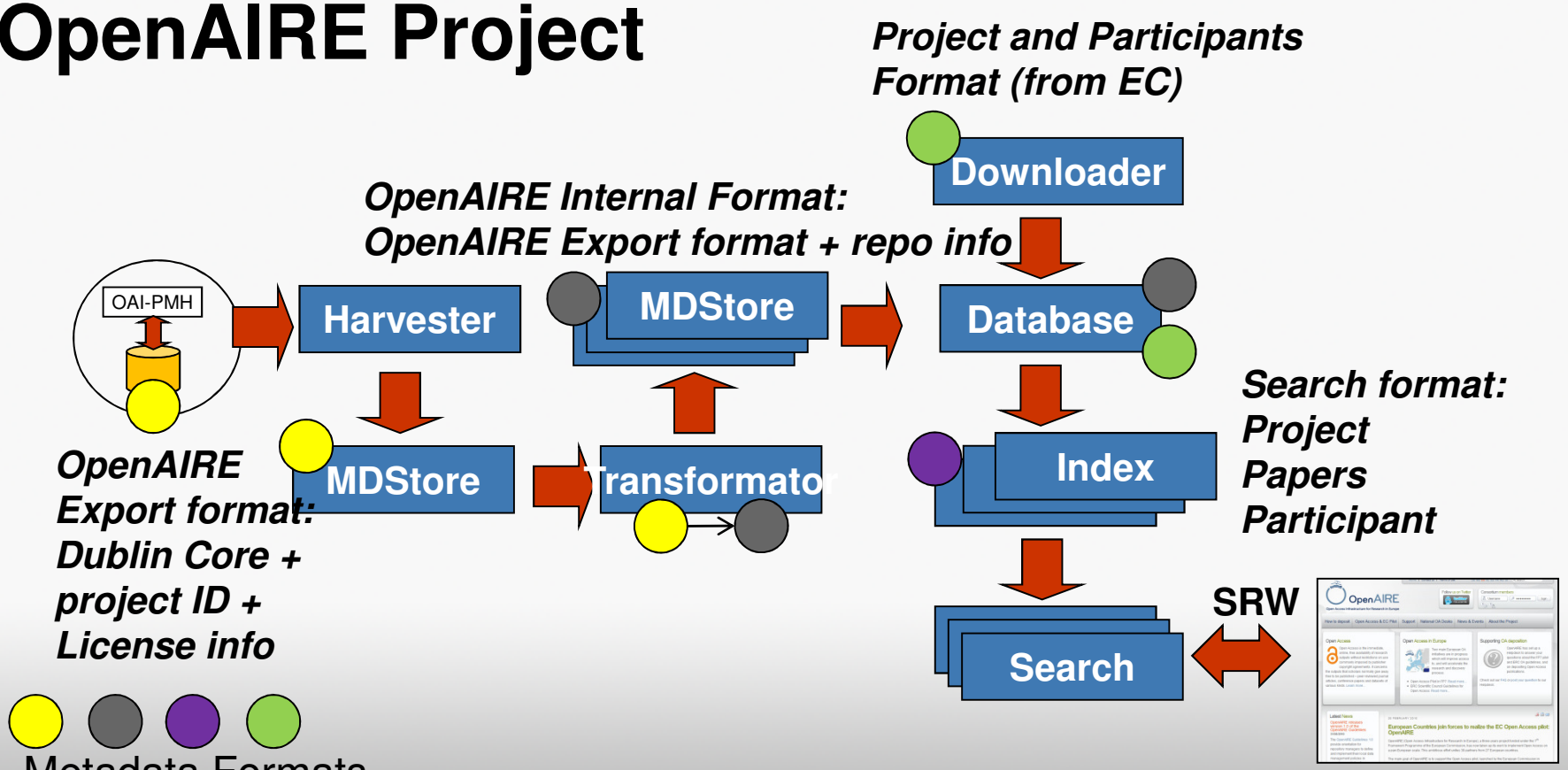
Modularity, customizability, sharing (and orchestration)

EFG Project



Modularity, customizability, sharing (and orchestration)

OpenAIRE Project



D-NET's uptake

- DRIVER project
  - 250 repositories (34 countries), 2,300,000+ items
  - search.driver.research-infrastructures.eu
- European Film Gateway EC project
  - 14 archives, 300,000 items, compound object data model
  - www.europeanfilmgateway.eu
- OpenAIRE EC pilot
  - Harvesting, depositing and statistics of publications and EC project data
  - www.openaire.eu
- HOPE project
  - +20 archives, millions of items, compound object data model
  - www.iisg.nl/news/hope.php
- ScholarLynk
  - R2D2 Project: Microsoft Research Cambridge and D-NET

Experimentation

- Experimentation of deployment of new D-NET repository infrastructures
  - China, India, Portugal, Belgium, Spain, Slovenia
  - Upcoming: Greece and Bulgaria

D-Net Software Toolkit

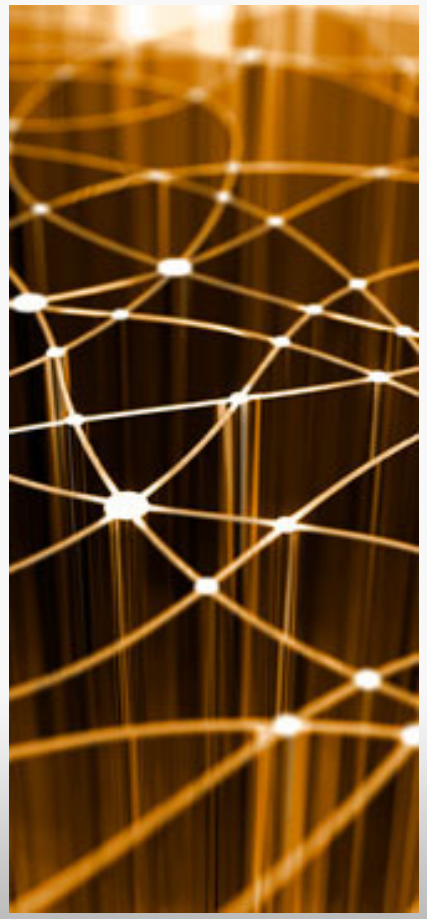
- Software packages
  - Open Source Apache License
  - Release v1.0 (production) and v1.2 (beta)
  - Release v2.0 (beta): Enhanced Publication
- Under continuous refinement

  - www.d-net.research-infrastructures.eu

Technical Team

- CNR-ISTI: Istituto di Scienze e Tecnologie Informatiche, Centro Nazionale delle Ricerche, Pisa, Italy
- NKUA: Department of Informatics and Telecommunications, National and Kapodistrian University of Athens, Greece
- UNIBI: Universität Bielefeld, Germany
- ICM: Interdisciplinary Centre for Mathematical and Computational Modeling, Uniwesytet Warszawski, Poland

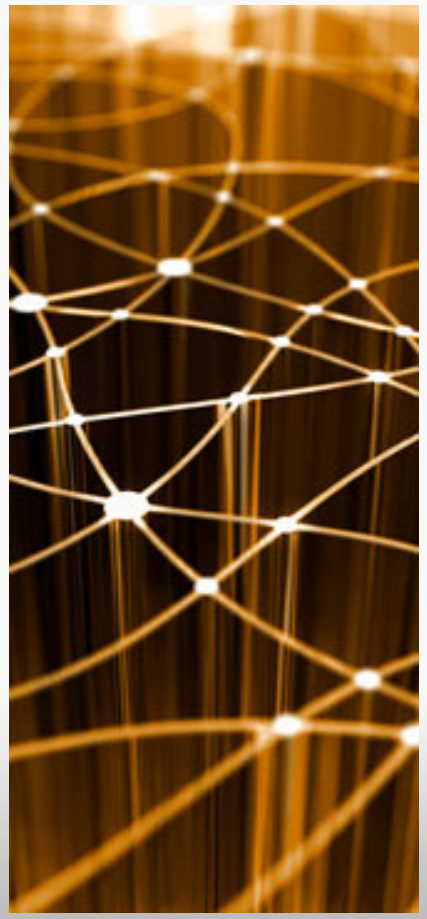
DL.org
Digital Library Interoperability, Best Practices and Modelling Foundations



Content Domain: Interoperability Solutions "Europeana Data Model"

Carlo Meghini

DL.org
Digital Library Interoperability, Best Practices and Modelling Foundations

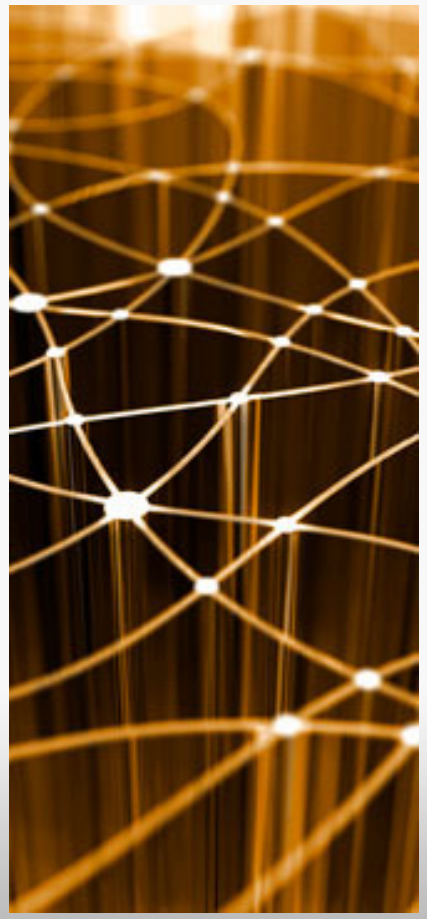


Content Domain: Hands-on Time

Exercises

- Indentify and produce RM Content domain enhancements
  - Each enhancements should be equipped with a motivation.
  - Enhancements might be on the introduction of new concepts and/or relationships, on the revision of existing definitions as well as on exemplars;
- Select one (or more) “DL” system and describe its content domain by relying on the Reference Model;
- Identify and describe a content-oriented interoperability solution (Overview-Requirements-Results-Implementation Guidelines-Assessment);
- Revise one (or more) Cookbook content-oriented interoperability solutions;
- Work on the Content domain part of the interoperability scenario;

DL.org
Digital Library Interoperability, Best Practices and Modelling Foundations



Thank you