Diversity of Digital Repositories in DRIVER

DELOS: Third Workshop on Foundations of Digital Libraries

ECDL 2008 in Aarhus, DK

Wolfram Horstmann | Bielefeld University

This talk …

…introduces DRIVER

… DRIVER in terms of DELOS DLRM

&

… Diversity of repositories encountered

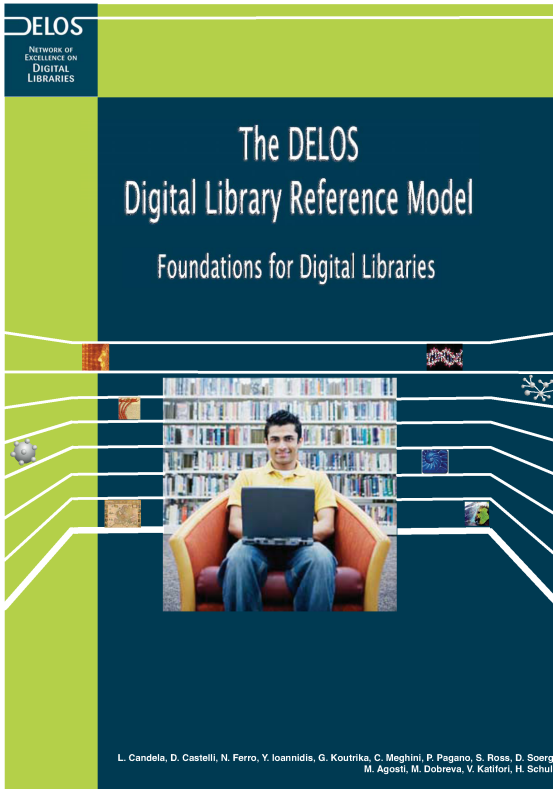
… *DRIVER*

in terms of

DELOS DLRM

DELOS DLRM Terms

- Digital Library
- Digital Library System
- Digital Library Management System
- Users
  - DL End users
  - DL Designer
  - DL System Administrator
  - DL Application Developer
- Content
- Functionality
- Policy
- Architecture
- Quality



driver

DLRM : Digital Library

- DRIVER-Community Website
  - Support for Repository Managers
  - Registration of Repositories
  - Validation of Repository-Content
  - Aggregation of distributed metadata
  - Search Repository Content
  - Future: Collections, Communities etc.
  - Advocacy for Open Access

driver

Digital Repository Infrastructure Vision for European Research

Search all repositories

Reset Search

DRIVER compliant repositories

Limit your search by

- Document Type
- Date of publication
- Document Language
- Repository
- Community
- Collection

Selected limits

hide details

No search limits.

Search History

Browse for documents

by Author
by Document type
by Subject
by Document Language
by Repository

DRIVER Collections

Anthropology
Biology
Computer networks
Computer Science
DAREnet Repositories
DART Europe Repositories
Database Systems andTheory
DINI Certificated Repositories
European Countries

See all DRIVER Collections...

Learn about Collections...

DRIVER Communities

Computer Science in the UK
DINI Repositories and Database Research
Molecular Biologists

See all DRIVER Communities...

Learn about Communities...

Create your own profile

Create your account to receive personalized services. Registered users can filter their searches or receive alerts based on their individual research interests.

Register here.

News & Announcements

Features more than 600,000 Open Access documents from over 110 European repositories in 25 languages.

For repository managers

Is your repository registered with DRIVER? If not, contact us via our helpdesk and we will guide you through the registration process.

For service providers

Develop the repository landscape!

Are you a technical person? Then you may want to have a look at the current DRIVER infrastructure deployment, or take a peek at two test installations of the driver infrastructure for specialized communities: DART and Recolecta.

Visit our official project site to download the latest DRIVER infrastructure software release D-NET v.1.0

If you have any questions or need further information please contact our helpdesk and we will be happy to assist you!

e-infrastructure

CAPACITIES

driver

Digital Repository Infrastructure Vision for European Research

Search all repositories

ticer

Reset Search

DRIVER compliant repositories

Limit your search by

- Document Type
- Date of publication
- Document Language
- Repository
- Community
- Collection

Selected limits

hide details

Repositories

- Scientific production, Tilburg University [remove]

Search History

- (repo:="Scientific production, Tilburg University") AND (ticer)
- (ticer)

Search Results

Refine your search | New Search

Found 3 documents, displaying page 1 of 1

Some alternatives for the Boolean Model in Information Retrieval

Creator(s) Paijmans, J.J.

Description -

Repository Scientific production, Tilburg University
Repository Info | Repository's web site

Language English

View document details...

The Implementation of the EU Database Directive 96/9/EC

Creator(s) Prins, J.E.J.

Description -

Repository Scientific production, Tilburg University
Repository Info | Repository's web site

Language English

View document details...

The International Ticer School: Getting inspired to shape your library of the future

Creator(s) Prinsen, Jola G.B.

Description -

Repository Scientific production, Tilburg University
Repository Info | Repository's web site

Language English

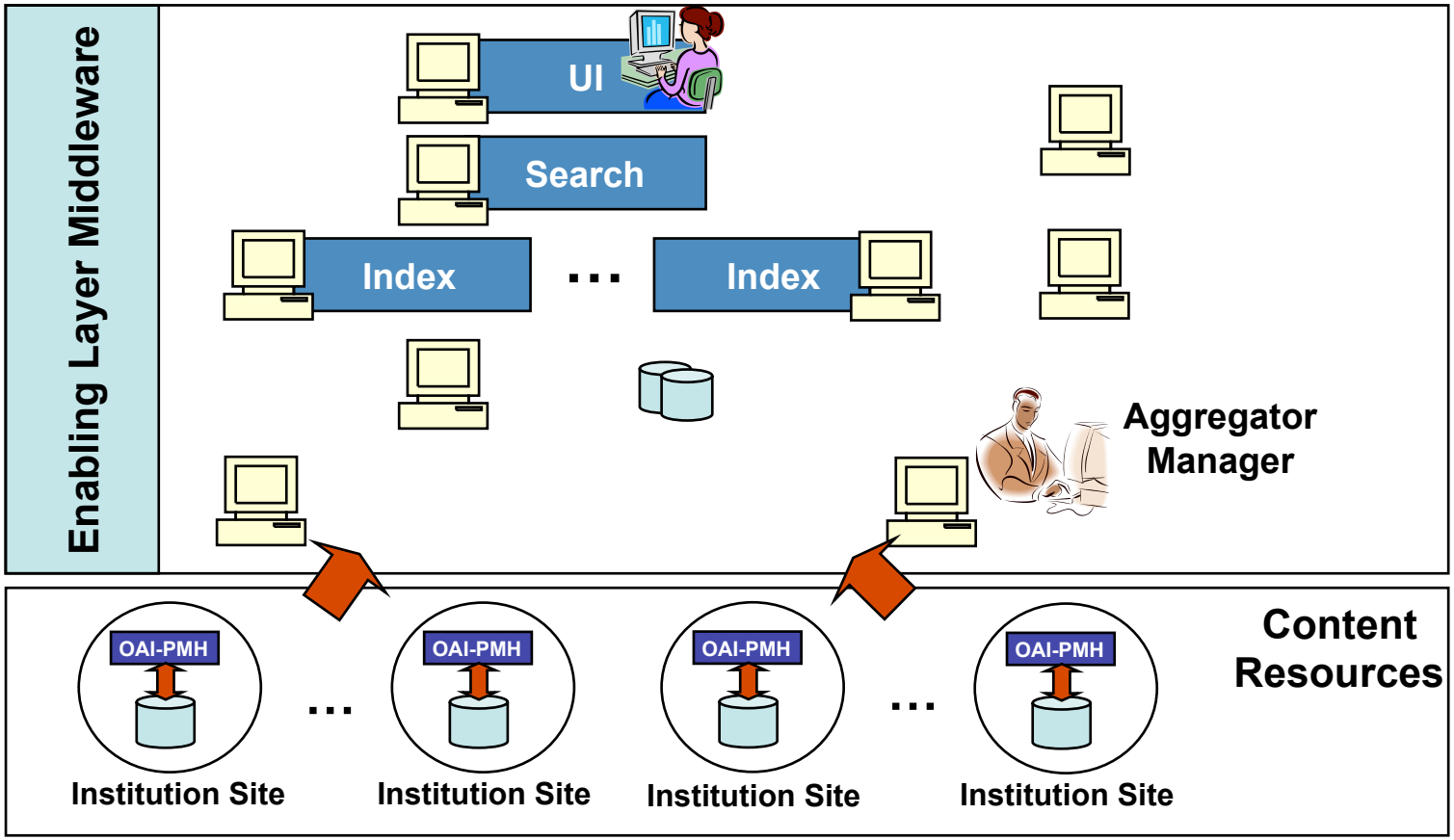
View document details...

Found 3 documents, displaying page 1 of 1

DLRM : Digital Library System

DRIVER Information Space Application

Repository Aggregation System as dynamic, distributed run-time infrastructure

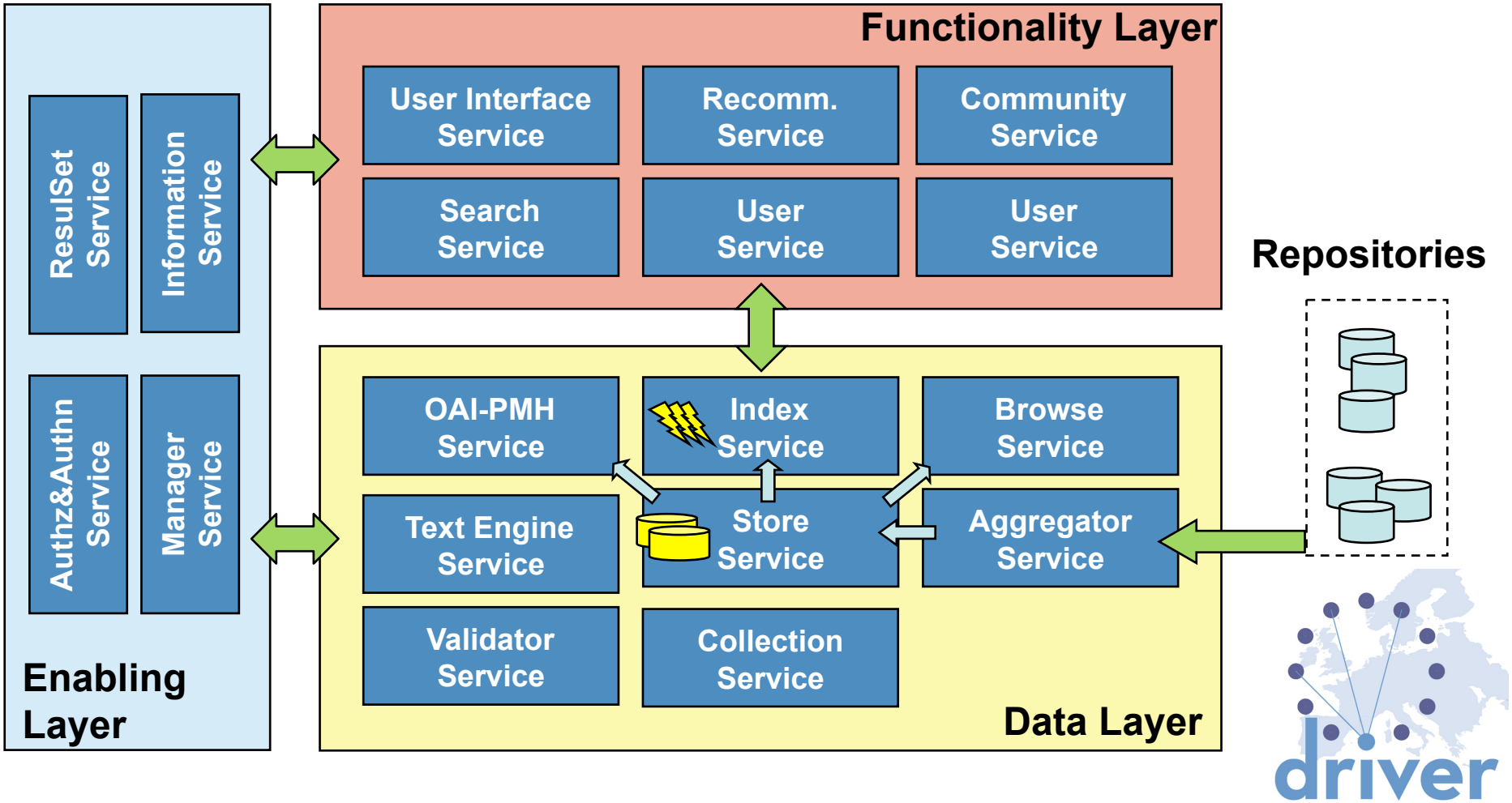


DLRM : Digital Library Management System

- D-Net Production Infrastructure
  - A set of services (deployed, hosted)
  - Ready to be administered, curated …
  - Capable of supporting multiple digital libraries
    - Q: are infrastructure admins the same as DLMS?
    - Q: A DLS may depend on many different DLMS!?
    - Q: DL System Warehouse / … Generator?
- Different from D-Net Software
  - A set of code packages
    - not deployed/not hosted

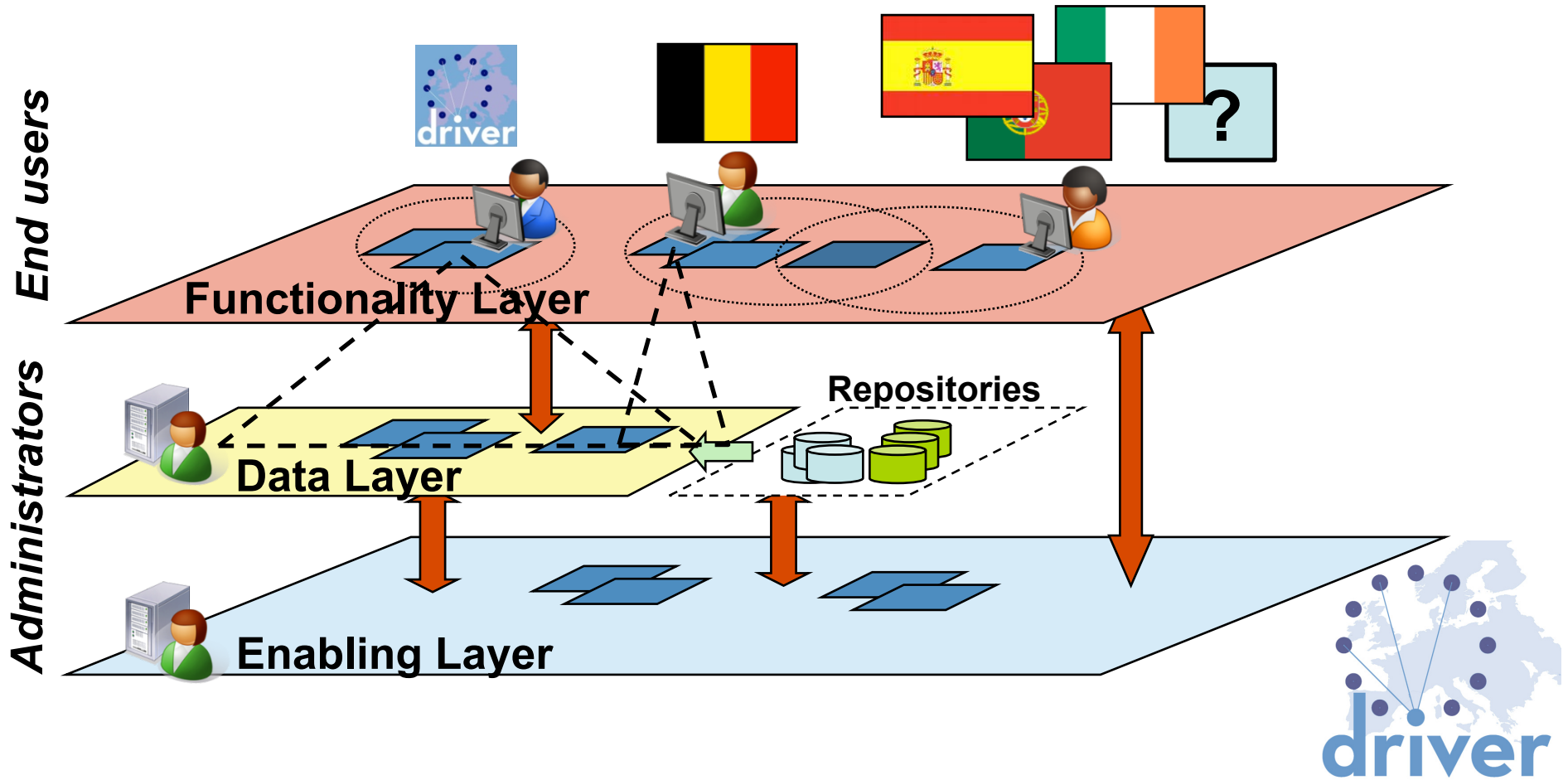
DLRM : Digital Library Management System

D-NET Software – service tool-kit



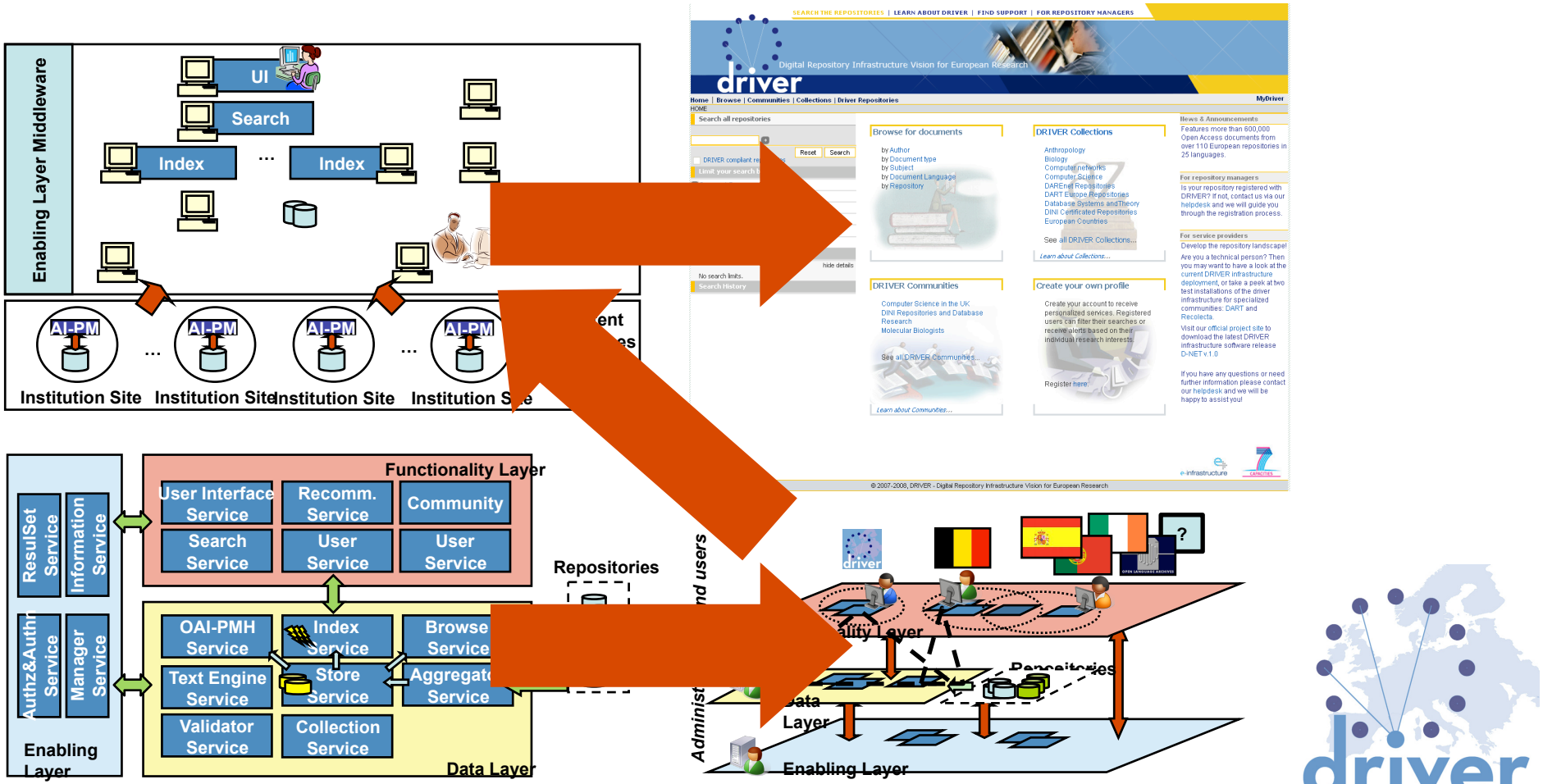
DLRM : Digital Library Management System

D-NET Software – multiple digital libraries



DLRM : Architecture

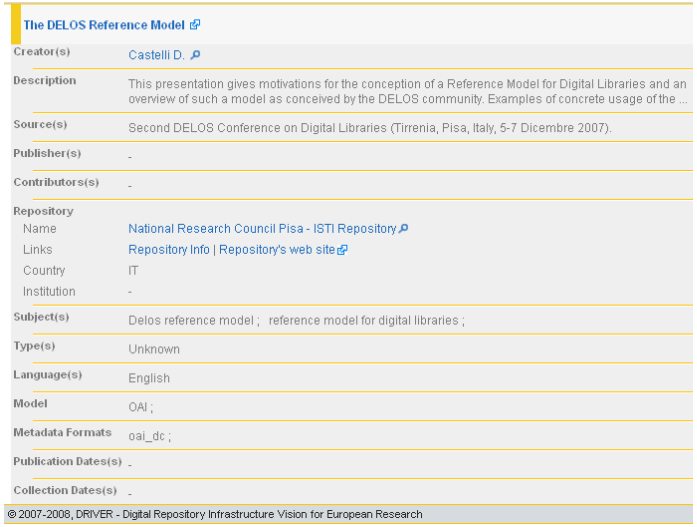
How DLs are relaized through DLS &DLMS



DLRM : User

- DL End users
  - Researchers (e.g. informations seekers)
  - Repository Managers (e.g. validation tools)
- DL Designer
  - DRIVER consortium
- DL System Administrator
  - Service deployer / aggregation manager / RM?
- DL Application Developer
  - DRIVER technical partners – CNR, NKUA, ICM, UNIBI

DLRM : Content



The DELOS Reference Model

Creator(s)	Castelli D.
Description	This presentation gives motivations for the conception of a Reference Model for Digital Libraries and an overview of such a model as conceived by the DELOS community. Examples of concrete usage of the ...
Source(s)	Second DELOS Conference on Digital Libraries (Tirrenia, Pisa, Italy, 5-7 Dicembre 2007).
Publisher(s)	-
Contributors(s)	-
Repository	
Name	National Research Council Pisa - ISTI Repository
Links	Repository Info | Repository's web site
Country	IT
Institution	-
Subject(s)	Delos reference model ; reference model for digital libraries ;
Type(s)	Unknown
Language(s)	English
Model	OAI ;
Metadata Formats	oai_dc ;
Publication Dates(s)	-
Collection Dates(s)	-

- Today
  - Scholarly articles
  - Metadata only – Simple DC >> DMF
  - System resources – users, repositories …
- Future
  - Full-text – Text-Mining, Classification …
  - All kinds of metadata-formats – MARC, MODS etc.
  - All kinds of materials – data, image, models etc.
  - Complex Objects – enhanced publications, relations

driver

Search all repositories

delos

Reset Search

DRIVER compliant repositories

Limit your search by

Document Type

Date of publication

Document Language

Repository

Community

Collection

Selected limits

No search limits. hide details

Search History

(delos)

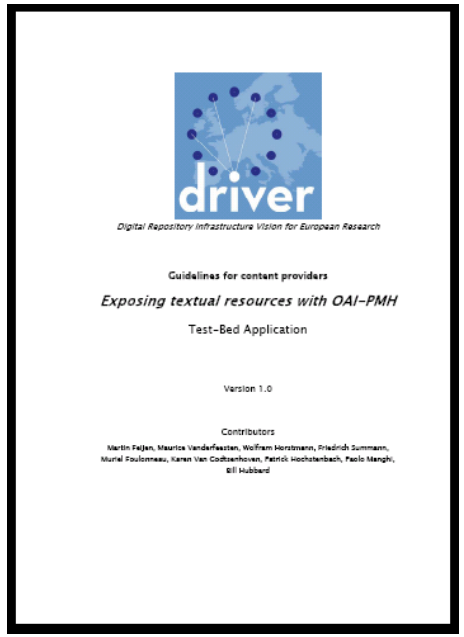
The DELOS Reference Model

Creator(s)	Castelli D.
Description	This presentation gives motivations for the conception of a Reference Model for Digital Libraries and an overview of such a model as conceived by the DELOS community. Examples of concrete usage of the ...
Source(s)	Second DELOS Conference on Digital Libraries (Tirrenia, Pisa, Italy, 5-7 Dicembre 2007).
Publisher(s)	-
Contributors(s)	-
Repository	
Name	National Research Council Pisa - ISTI Repository
Links	Repository Info | Repository's web site
Country	IT
Institution	-
Subject(s)	Delos reference model ; reference model for digital libraries ;
Type(s)	Unknown
Language(s)	English
Model	OAI ;
Metadata Formats	oai_dc ;
Publication Dates(s)	-
Collection Dates(s)	-

DLRM : Functionality

- Contents
  - Registration of content
  - Aggregation
  - Indexing
  - Search / Browse
  - Collect / Customize
- Repository Managers
  - Registration
  - Validation

DLRM : Policy


driver

Digital Repository Infrastructure Vision for European Research

Guidelines for content providers

Exposing textual resources with OAI-PMH

Test-Bed Application

Version 1.0

Contributors

Martin Feijen, Maurice Vanderfeesten, Wolfram Horstmann, Friedrich Summann, Muriel Foulonneau, Karen Van Godtsenhoven, Patrick Hochstenbach, Paolo Manghi, Bill Hubbard

- Content Policy
  - Extrinsic policy: local repository
  - Intrinsic policy: DRIVER guidelines …
- User Policy
  - Free to use, authentication optional
  - Administrator policy >> consortium/confederation
- System Policy
  - Open standards + technical agreements

driver

DLRM : Quality

- Content Quality Parameters
  - Amount of referenced objects
  - Coverage of subject domains
  - Amount of aggregated sources
  - Normalization of metadata
  - Richness of metadata
  - …
- System Quality Parameters
  - Reliability, robustness, scalability

Interim conclusion

DRIVER is easily expressed in terms of the DELOS DLRM

(Q: where is the notion of infrastructures)

… *Diversity of repositories in DRIVER*

Diversity phenomenon *1/4*

- 1000+ institutional repositories
- Problems are the standard, not the exception
  - Repositories do not respond or deliver error-msg
  - Harvesting process is slow or dies
  - Incremental Harvesting not supported
  - Data contain only references without any full-text
  - Links to the document do not work
  - XML file is not well-formed
  - Field content varies

Diversity phenomenon *1/4*

- 1000+ institutional repositories
- Problems are the standard, not the exception
  - Repositories do not respond or deliver error-msg
  - Harvesting process is slow or dies
  - Incremental Harvesting not supported
  - Data contain only references without any full-text
  - Links to the document do not work
  - XML file is not well-formed
  - Field content varies

Diversity phenomenon *2/4*

- E.g. Field Contents Diversity: <dc:creator>

<dc:creator>Barry Wellman,Jeffrey Boase,Kakuko Miyata</dc:creator>
<dc:subject>Barry Wellman,Jeffrey Boase,Kakuko Miyata The Mobile-izing ....</dc:subject>

<dc:title>Talk P. Bruzzone</dc:title><dc:creator>Bruzzone </dc:creator>
<dc:creator>Pierluigi</dc:creator>

Diversity phenomenon *3/4*

- E.g. Field Contents Diversity: <dc:language>

EN: 9910
ENG: 771
En: 566
Eng: 1
English: 24084
English (United States): 63
English and Greek: 1
English and Russian: 1
English/Japanese: 1
English; Russian: 1
English=en: 1
Translation into English: 2

en: **1279115**
en-CA: 865
en-US: 3
en-es: 5
en-us: 8
en;: 2
en_UK: 618
en_US: 18456
eng: **186787**
eng : 92
eng + dut: 2
eng;: 17
eng; fre; ger;: 141
....

Diversity phenomenon *4/4*

- E.g. Field Contents Diversity: <dc:type>

Text - 2597764
text - 2540080
TEXT - 1128327
Dataset - 580448
Image - 522643
image - 514585
Electronic Thesis or Dissertation
Tese ou Dissertacao Eletronica
article - 345093
Article - 337314
Artikel - 310858
Other - 219331
still image - 212178
Rezension - 198773
preprint - 191784
…

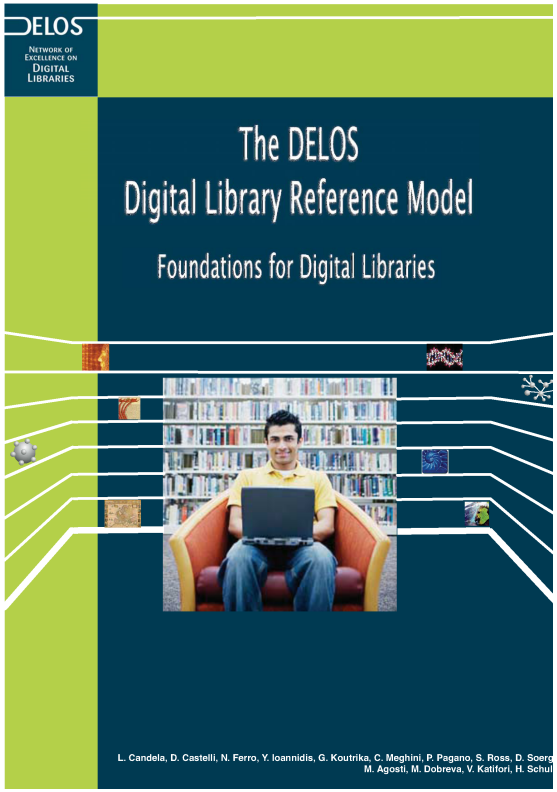
NULL - 8082
Clippings
Text
 - 5004
This record contains content - 1473
Dark Item - 1145
Photographs; Mosaics; - 1
1 photographic print mounted on cardboard: b&w; 22 x 16 in. - 1
80534 - 1
Section V Flow Control - 1
Acrylic on board: 63 x 114 cm. - 1
79508 - 1
Colored graphite on paper: 47 x 63 cm. - 1
texts, vocabulary ('stories, monologues, word list' - 1
15121 - 1

Diversity of repositories …

- DRIVER today: Institutional Repositories
  - Remember: I give only one example of diversity
- DRIVER future: all kinds of repositories
  - Image, audio, video material
  - Primary data, analyzed data, models, simulations
  - What about software-repositories?
  - Second-level repositories / enrichments
    - Entity recognition, Auto-Classification, Mash-Ups…

Diversity in DLRM Terms

- Digital Library
- Digital Library System
- Digital Library Management System
- Users
  - DL End users
  - DL Designer
  - DL System Administrator
  - DL Application Developer
- Content
- Functionality
- Policy
- Architecture
- Quality



driver

Digital Library

- Diversity origin
  - 1000 IRs are (parts of) 1000 local digital libraries

- DRIVER answers
  - Website offering normalized content
  - Website offering support for Repository Managers

Digital Library System

- Diversity origin
  - Different repository platforms and policies
    - DSPACE, EPRINTS, OPUS, FEDORA, OWN
  - …
- DRIVER answers
  - DRIVER Information Space Application
    - Normalized through OAI-PMH
    - Still platform differences

Digital Library Management System

- Diversity origin
  - not directly visible at the level of DLMS
  - Different aggregations of repositories
    - DAREnet, DINI, DART, RECOLECTA etc.

- DRIVER answers
  - D-Net Production Infrastructure
    - Capable of representing aggregations,
    - Capable of forming instances of DLSs

User

- Diversity origin
  - DL End users
    - 1000+ Repository Managers: individual contacts
    - 10000+ metadata and content providers
    - 100000+ information seekers
- DRIVER answers
  - Special user roles (Rep.Man. and Agg.Man)

Content

- Diversity Origin
  - Degrees of freedom for simpleDC
  - (Complexity of underlying world)
- DRIVER answers
  - Normalization / Transformation
  - Individual repository profiles

Functionality

- Diversity origin
  - Lose web forms ???
- DRIVER answers
  - Indirect: Registration and Validation functions

Policy

- Diversity Origin
  - Content Policy
    - 1000+ local / regional policies
  - User Policy
    - Varying authentication >> metadata providers

- DRIVER answers
  - DRIVER Guidelines
  - Validation as self-audit

Quality

- Diversity Origin
  - Content Quality Parameters
    - No benchmarking / monitoring
  - System Quality Parameters
    - No benchmarking / monitoring

- DRIVER answers
  - Benchmarking through validation score
  - System monitoring through regular reports

IR Diversity Conclusion

- Diversity challenges normal !?
  - Distributed system, many users, policies etc.
  - Still high homogeneity through OAI-PMH/ simpleDC >> far too good to drop
- Compliance enforcement main problem
  - Awareness in repository managers major factor
  - Guidelines validation scoring & reports may help

Future of repository diversity

… an excursion to data infrastructures …

5th e-Infrastructure Concertation
Barcelona
6 June 2008

e-Infrastructures as Standardisation Drivers

DATA TRACK

Chair : Krystyna Marek
Rapporteur: Wolfram Horstmann

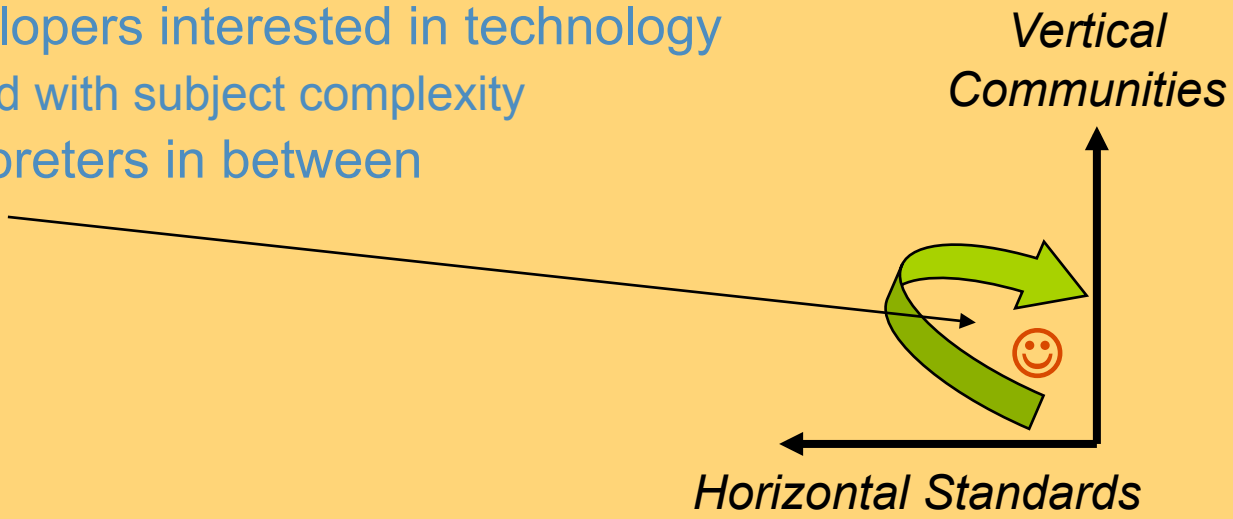
Participants in the session

- By call
  - Repository infrastructures
    - NMBD
    - IMPACT
    - DRIVER-II
    - EuroVO-DCA, EuroVO-AIDA
    - Genesi-DR
    - METAFOR
  - User communities
    - D4Science (Diligent)
  - Communications
    - BELIEF-II
  - Scientific data infrastructures (2008)
    - DIESIS, PESI, PARSE.Insight,
  - More / Observers
    - Spanish Innovation Min / BeInGrid / SimDat / EGEE / EUASIA / ETICS / ..

Underlined: participants new as compared to 4th concertation

Contextualizing the session

- Data-Projects are „user“ driven
  - horizontal standards play role of a commodity
  - Are supplemented by „vertical“ / community standards
  - Diversity is a declared objective
- Inherent paradox of subject differentiation and standards
  - Different Perspective than in Connectivity and Middleware
  - Researchers mainly interested in their subject
    - Standardization too technical
  - Standard developers interested in technology
    - overwhelmed with subject complexity
  - We need interpreters in between



Standards-Use

W3C	ISO no individual mentioning	OASIS	IEEE	IETF	ETSI
• HTML, URI/URL, XML • Web Services (WSDL, SOAP) • Ontologies/ Semantic Web (e.g. RDF, SPARQL, SKOS)	• Vocabularies (language, country, dates) • Virtual research environments • Geographic MetaData & information and services, • Archiving/OAIS	• Web Services (UDDI) • A&A (SAML/ XACML) • GeoXACML (OGC) • Business Markup / Process (ebXML, BPEL) • Stateful WebServices (WSRF) • Remote User Interfaces (WSRP)	• Architecture (HLA) • Simulation (DIS)	• --	• --

Underlined = updated information; Red = proactive contributons

Standards-Use

OGF	***OAI***	***DCMI***	***LOC***	***IVOA*** *(subject based)*	***Other*** *(subject based)*
• „Usage of other people‘s work“ • OGSA, DAIS, GridFTP, GLUE, DRMAA, JSDL	• Resource exposure/ aggregation (OAI-PMH) • Object Re-use and Exchange (OAI-ORE)	• Simple Metadata (DCMES) • Virtualizing (DC-Collection)	• Web-Service queries (SRU/W-CQL)	• Metadata • Resource Registry	• [Ontologies, Vocabularies, Terminology Services] • INSPIRE Metadata Draft (?) • OGF/OGC WebProc. Standard (WPS) WebService (OWS) and Web Catalogue (CSW) • Classification systems (e.g. ISSCAAP, ISSCFV, ISSCFG) • features representation (e.g. GML for GIS) • metadata (e.g. AgMES for Agricultural, SDMX for Statistics)

Underlined = updated information; Red = proactive contributons

Observations: *Challenges*

- Grand Challenge
  - Bridging between generic and subject specific
- Specific Challenges
  1. Protocols
  2. Identifiers
  3. Abstract Data Model
  4. Metadata

Observations*: Differing Networking*

- Point to Point Network
  - community by community definition of data standards
  - Community by community interfaces
  - *adhoc* standardization
- Layered Network
  - Minimal generic requirements for repositories
  - Thin interoperability layer …

Conclusion

- DRIVER easily expressed in terms of DLRM
  - Where is the notion of infrastructures (relation of x DLMS)
- Diversity of institutional repositories a consequence of distribution and complexity
  - Becomes clear in DLRM but may interrelation of basic concepts let it appear redundant
- Diversity of data repositories might require more
  - Extend foundations for (social) subject communities?