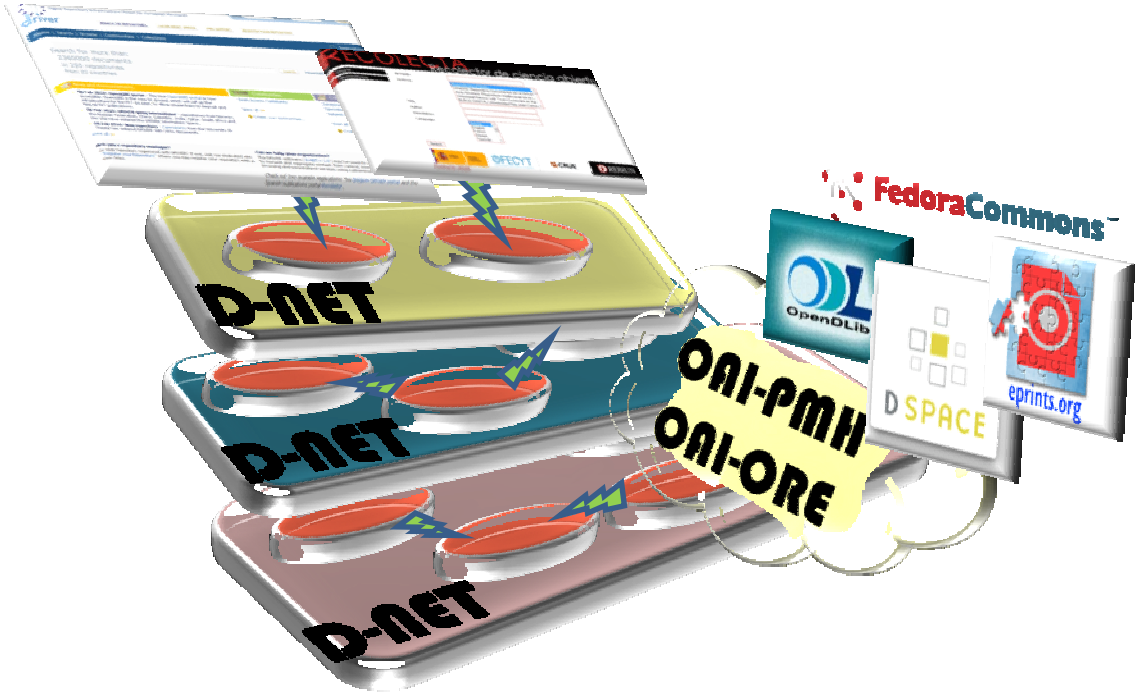
On Constructing Repository Infrastructures

The D-NET Software Toolkit

Paolo Manghi, Marko Mikulicic, Katerina Iatropoulou, Antonis Lempesis, Natalia Manola



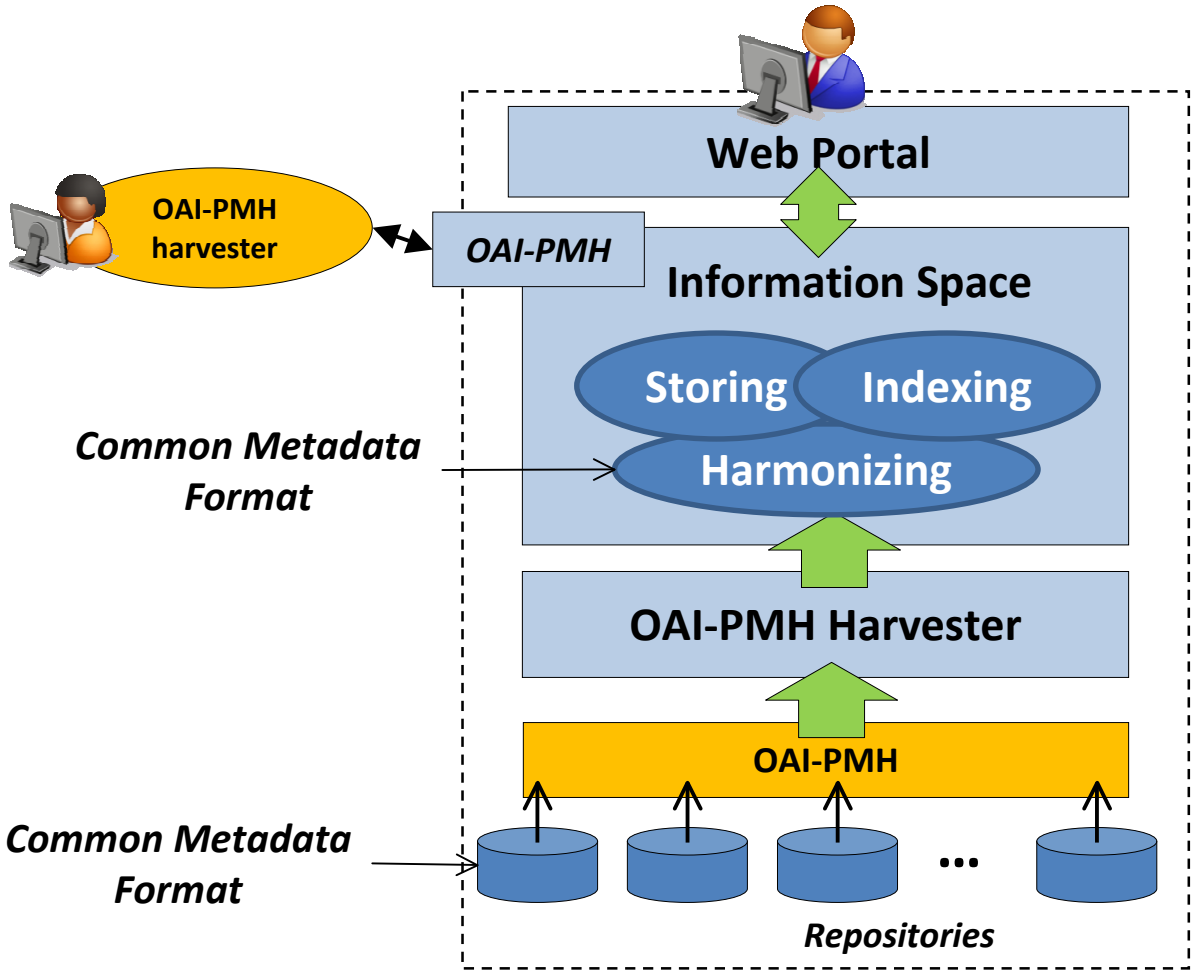
Universität Bielefeld

ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

Repository Infrastructures

- *Aggregation system*: maintaining and populating an *Information Space* by aggregating content from a collection of OAI-PMH Repositories
- *Web portal:* providing community-specific functionalities via Web User Interfaces
- Well known examples:
  - BASE (Germany)
  - DAREnet (Netherlands)
  - OAIster-OCLC (USA)
  - Others…
  - DRIVER Project
  - EFG project
  - HOPE project
  - Europeana project
  - Others…

Repository Infrastructures: Aggregation System



Extensions, updates, and refinements

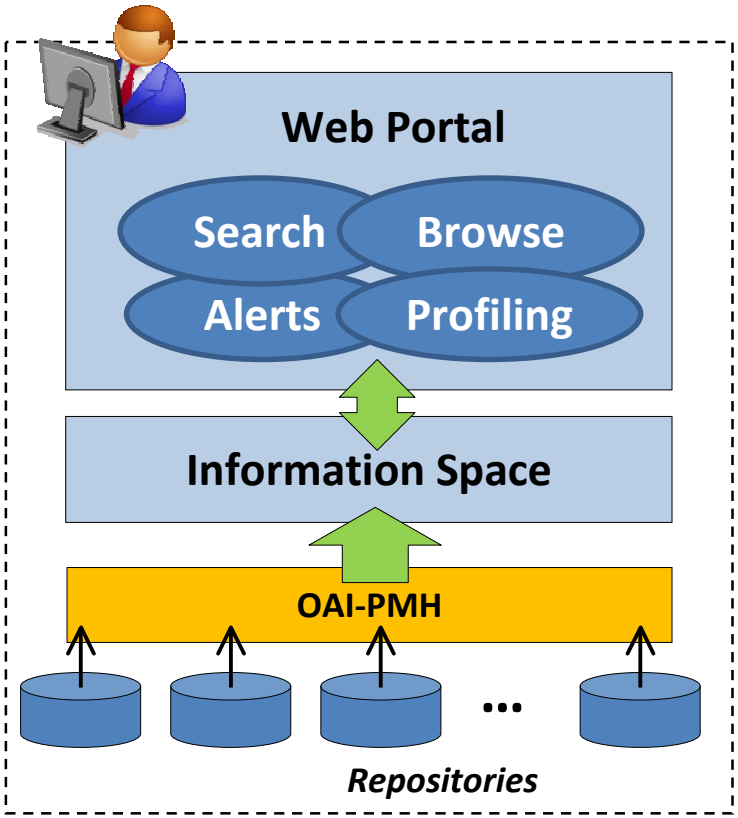
Data workflows definition

Cleaning, enriching, transforming

Repository Administration

Repository Infrastructures

- Web Portal -



Quality of Service

Extensions, updates, and refinements

Repository Infrastructure issues

- Aggregation systems
  - Arbitrary numbers of repositories (repository administration tools)
  - Harvesting and transformation actions (automated scheduling)
  - Definition of relative transformation mappings (mapping definition tools)
  - Data workflow definition
- Web portals
  - Search, browse (and more) over metadata
- Quality of service
  - Scalability, i.e., coping with ever growing incoming records
  - Robustness, i.e., data loss and availability of service

Limits of existing repository infrastructure solutions

- Limited customizability
  - E.g. pre-defined input and target metadata formats, predefined data workflows
- High-cost software extensibility
  - E.g., new functionality, new Information Spaces may require "expensive" changes
- "Manual" repository management
  - Registration, harvesting, curation (XSLT), etc…
- "Manual" administration for robustness and scalability
  - E.g., store and index replicas, system monitoring
- Constructed from scratch
  - E.g., from open source tools, writing code, often specialized

D-NET Software Toolkit

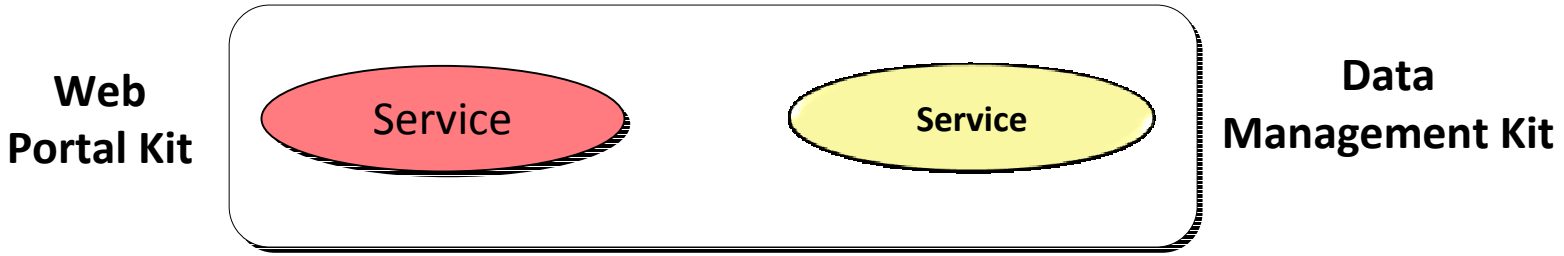
The aim..

"General-purpose framework for the easy development of domain-specific repository infrastructures"

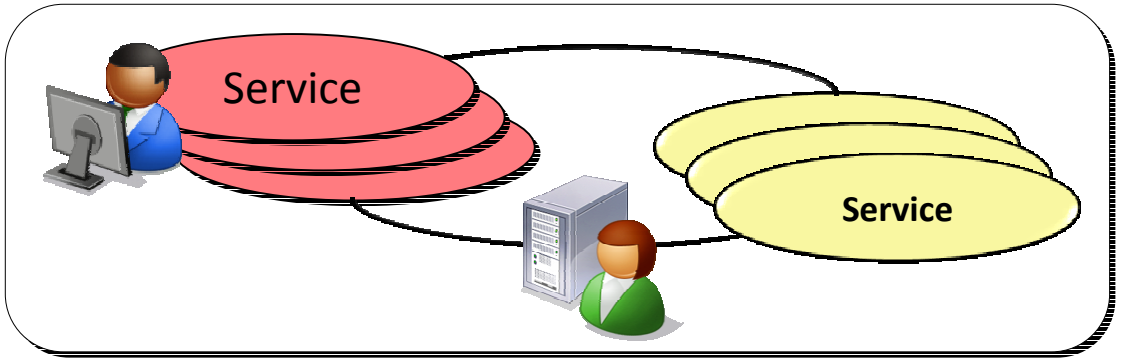
- Aggregation systems
  - Arbitrary metadata formats
  - Repository administration tools
  - Personalized and automated data workflows (data "manipulation")
- Web portals
  - Arbitrary metadata formats
  - Personalized end-user functionality

D-NET Software Toolkit
The solution…

- *Service Kits* supporting "personalizable" repository infrastructure functionality



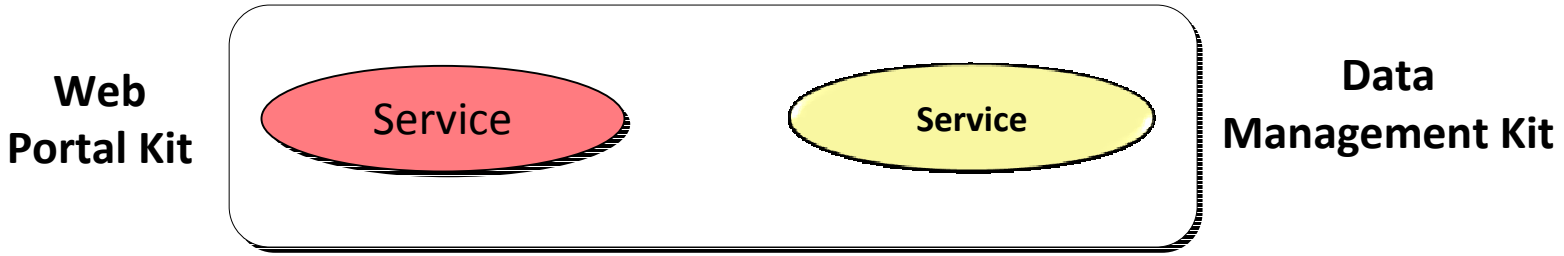
- *Service-oriented infrastructure features* to support sustainable production systems



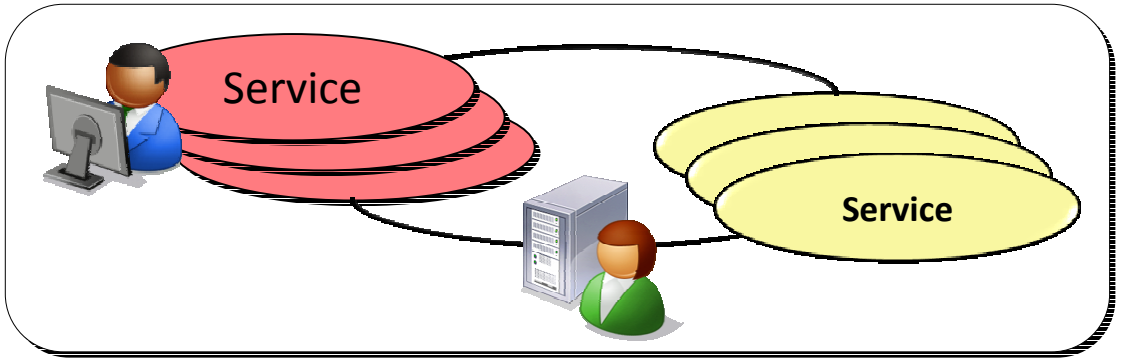
D-NET Software Toolkit

The solution…

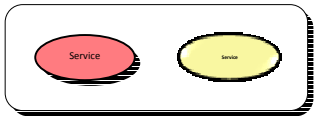
- ***Service Kits* supporting "personalizable" repository infrastructure functionality**

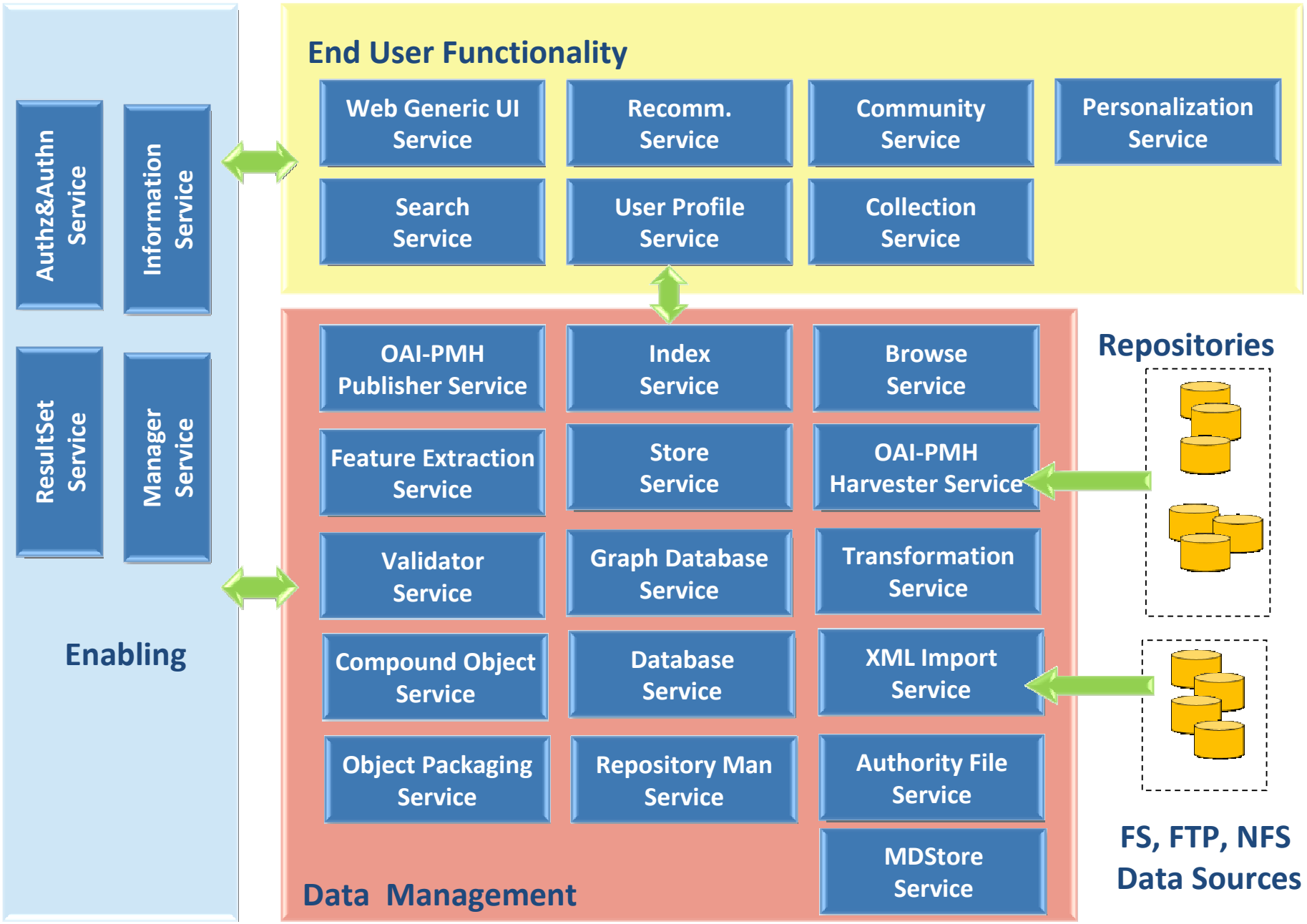


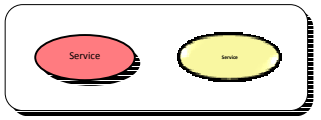
- *Service-oriented infrastructure features* to support sustainable production systems



D-NET: Service Kits





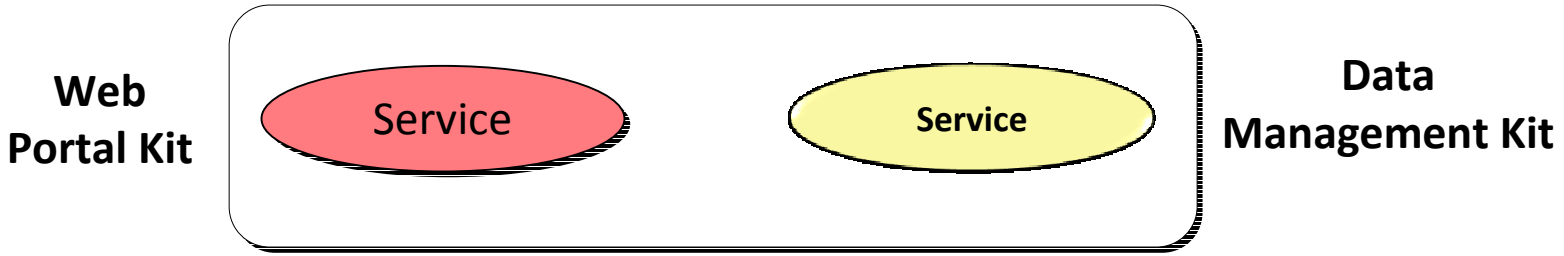


D-NET: service kits properties

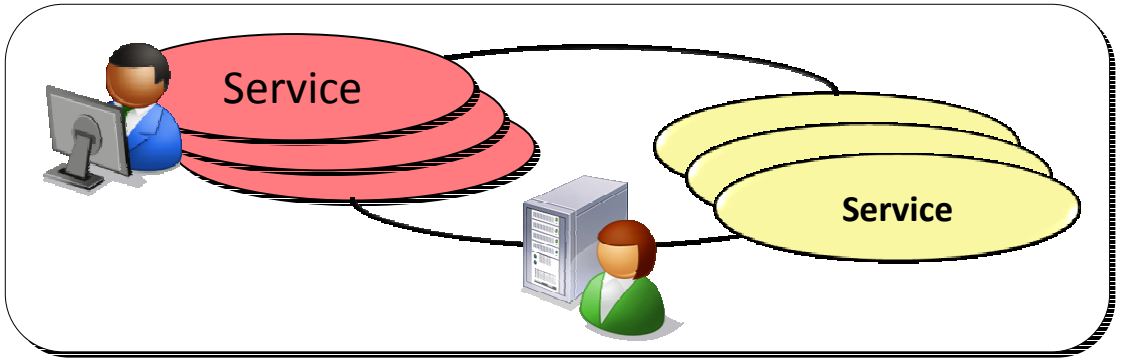
- Service modularity (“LEGO approach”)
  - Functionality in “isolation” (e.g. index, storage, transformation) to enable tailored data workflows
  - Service Customizability
  - Parametric services, e.g., any metadata format (XML schema)
  - Service Extendibility
  - New functionality can be easily integrated with existing ones

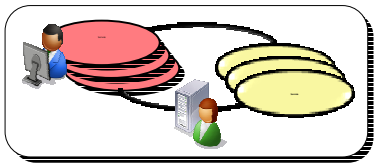
D-NET Software Toolkit
The solution…

- *Service Kits* supporting "personalizable" repository infrastructure functionality



- ***Service-oriented infrastructure features* to support sustainable production systems**





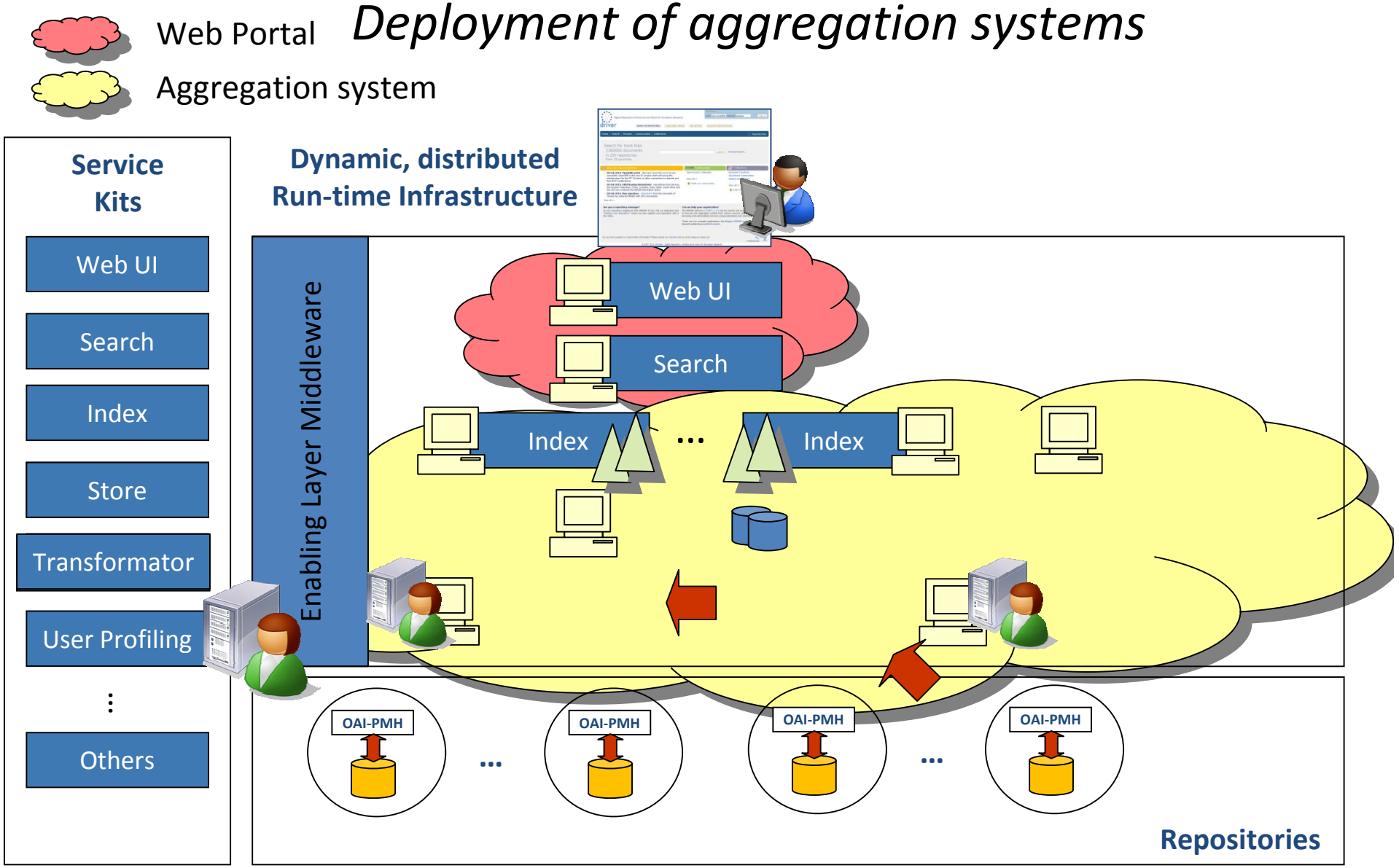
D-NET: service oriented features

"Enabling the operation of scalable, robust and autonomic applications"

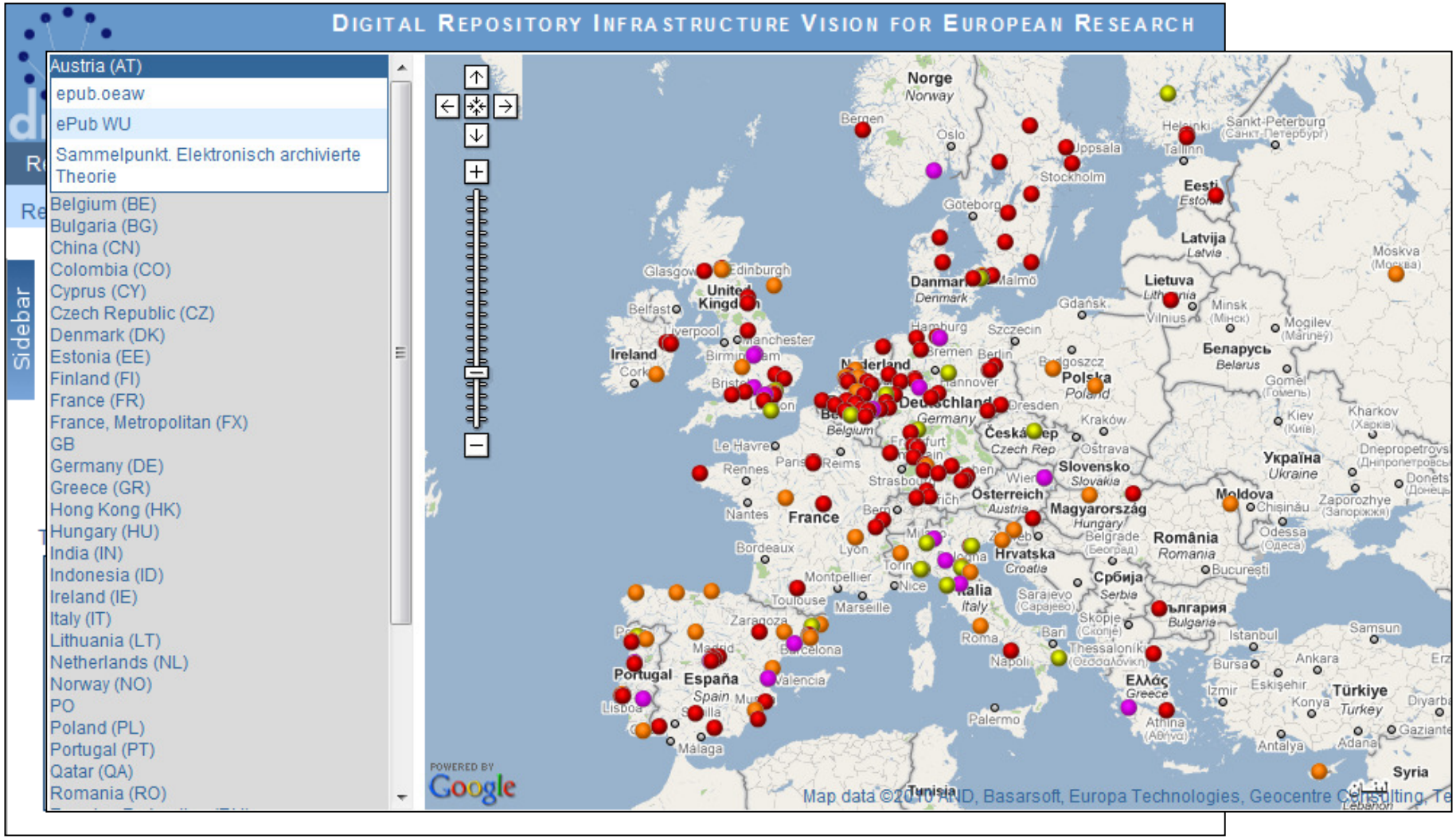
- Distribution
  - Services *can* be distributed, workload distribution, robustness and replicas
- Sharing
  - Services (and hardware) can be shared across several applications (reducing overall cost)
- Autonomic behavior by orchestration
  - Services can be orchestrated automatically to accomplish certain tasks ("workflow automation")
  - Reduced maintenance and administration cost

Repository Infrastructures in D-NET

Deployment of aggregation systems

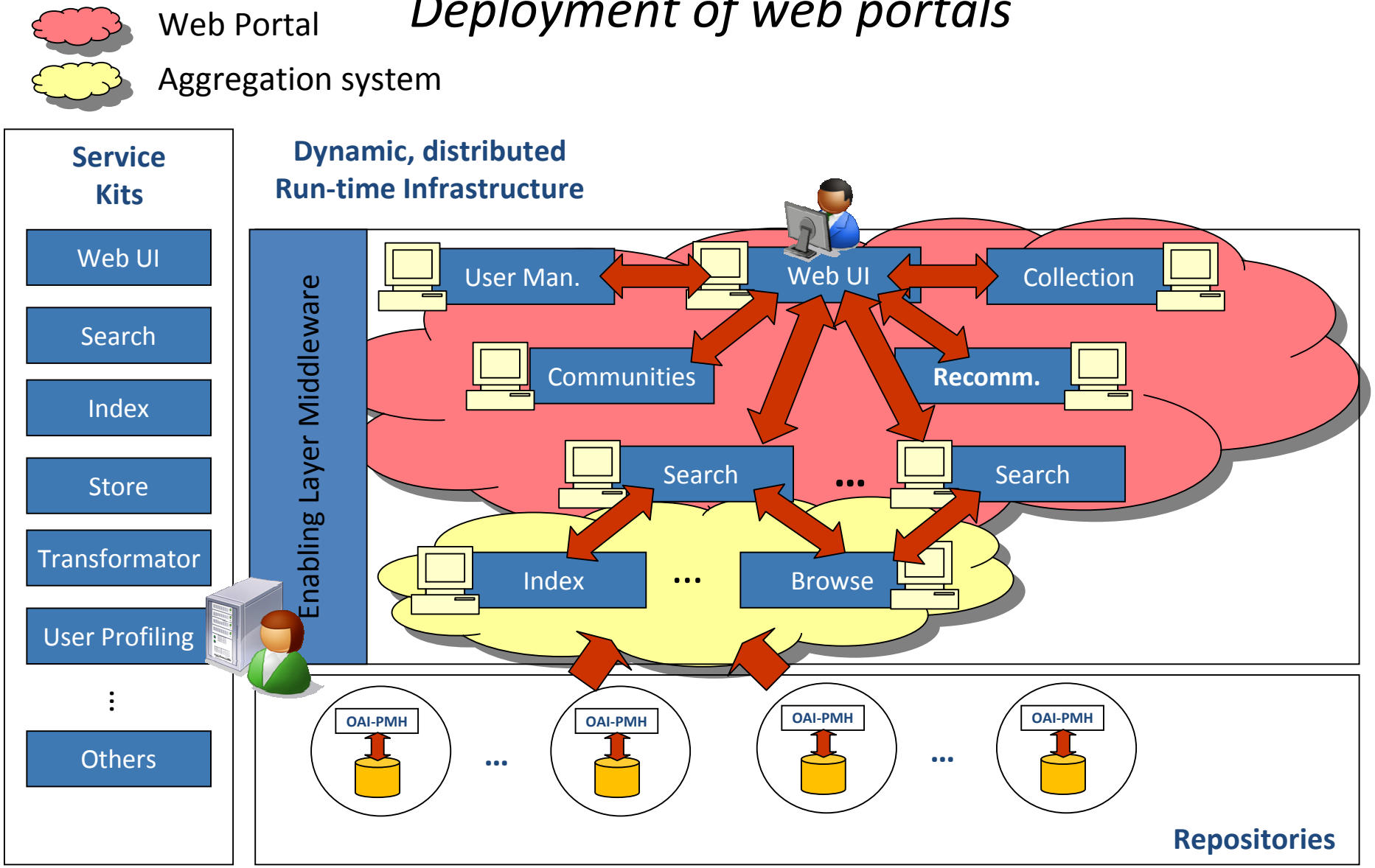


Repository Administration Tools



Repository Infrastructures in D-NET

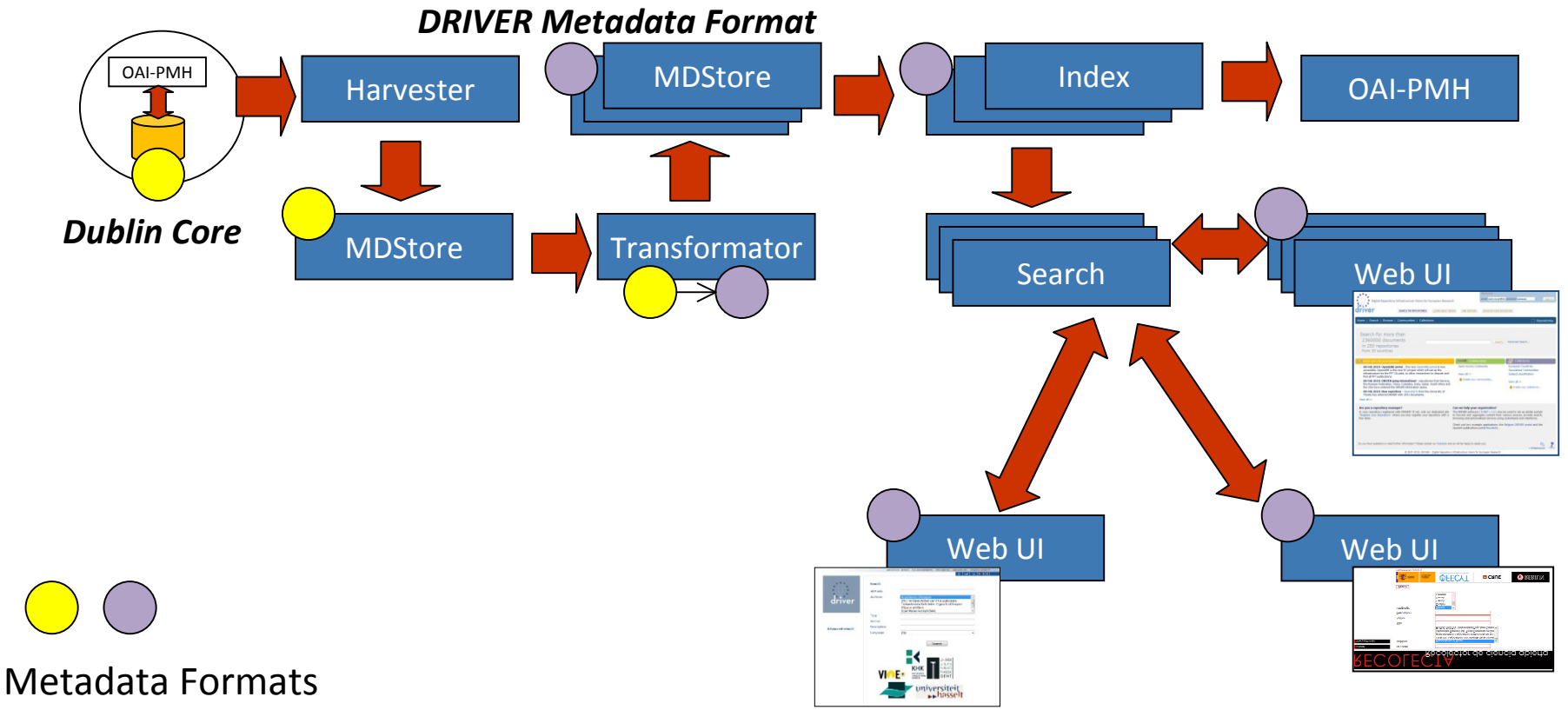
Deployment of web portals



Web Portal deployment

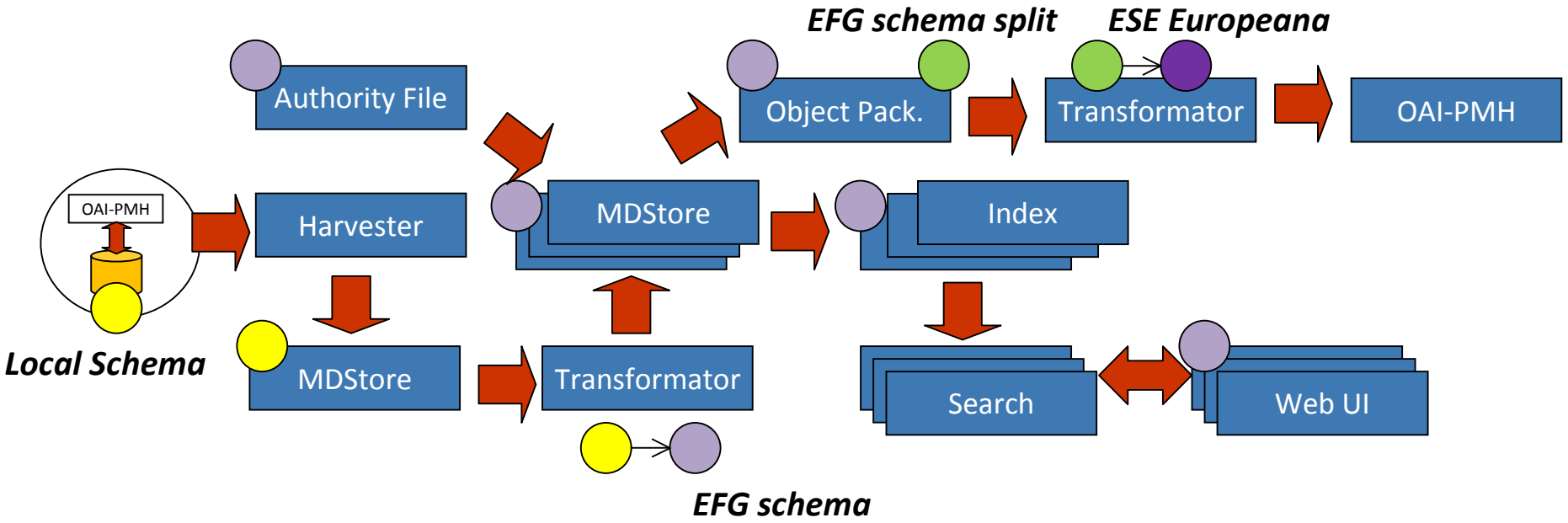
Modularity, customizability, sharing (and orchestration)

DRIVER Project

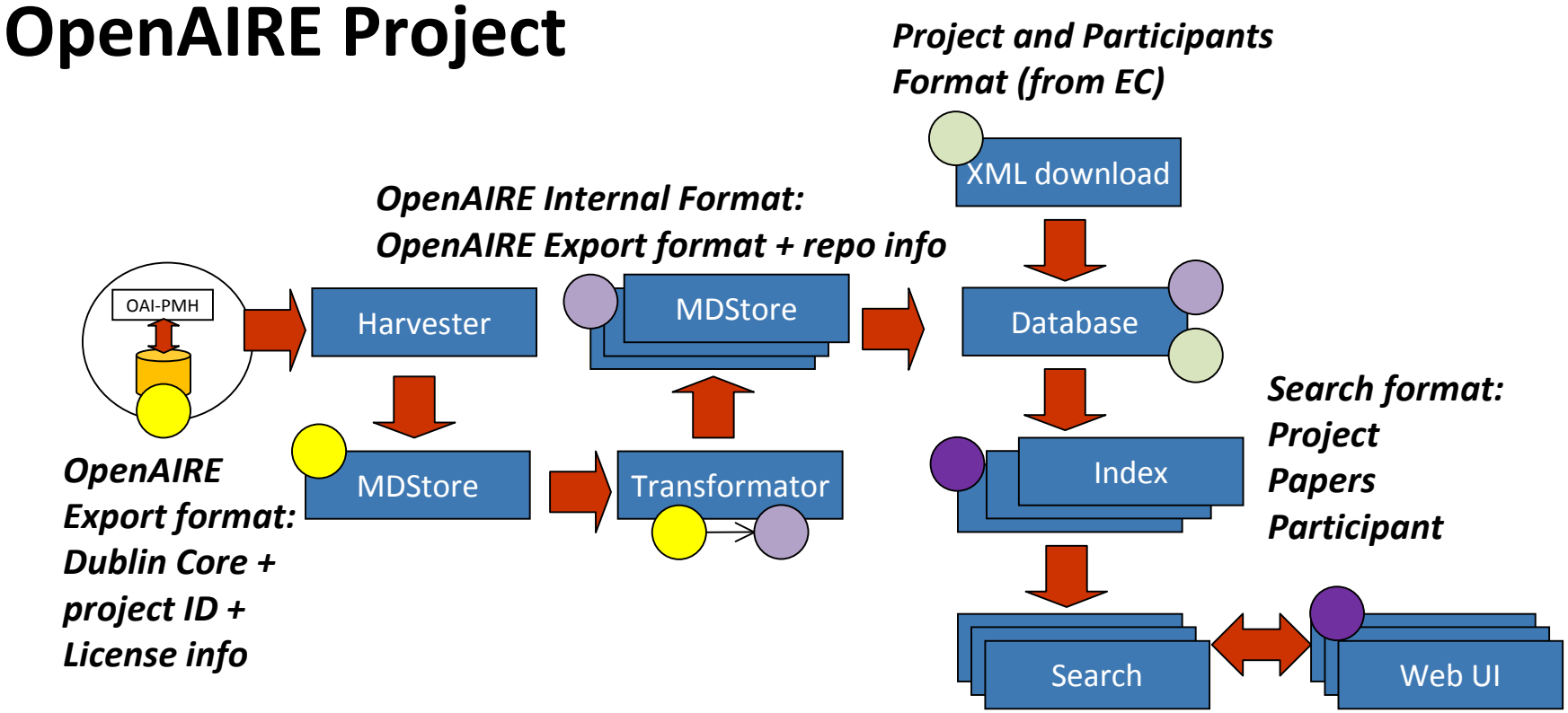


Modularity, customizability, sharing (and orchestration)

EFG Project



Modularity, customizability, sharing (and orchestration)



D-NET benefits: summary

- Functionality Services
  - Customizability: parametric services
  - Modularity: data workflow definition
  - Openness: designed to integrate new functionality
  - Data Repository management tools
    - Services for metadata collection, transformation, integration and web access
    - GUIs for harvesting and aggregation
- Sustainable Production Systems
  - Scalability and Robustness
    - By distribution, replicas and sharing
  - Sharing
    - Cost optimization (services and hardware)
  - Autonomic behaviour
    - Reduced maintenance and administration cost

D-NET's uptake

- DRIVER project
  - 250 repositories (34 countries), 2,100,000+ 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

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

D-NET and standards

- Service Resources are implemented as Web Services and accessed through the corresponding Web Service Interface
  - Parameters calls are enveloped into SOAP messages
  - The Enabling Services are also compatible with REST
- XML is the lingua-franca for the whole system
  - Resource internal status, i.e. Resource profiles, are represented as XML files conforming to a given schema
  - Profiles are kept into the Information Service, whose underlying engine is an Exist XML engine

D-NET and standards

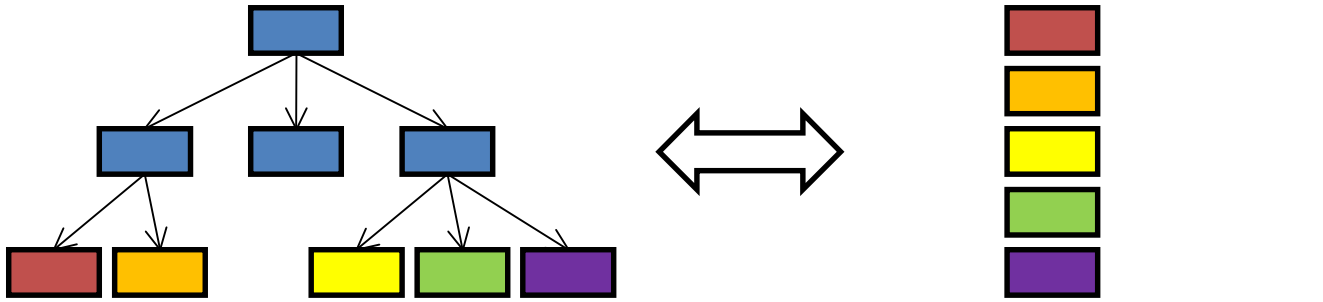
- Subscription and Notification Service
  - Any Service can subscribe to events regarding any DRIVER Resource: creation, deletion, and specific action accomplished by a resource
  - The Subscription and Notification mechanism is compliant with the OASIS Standards WS Base Notification 1.3 and WS Topics 1.3
- Authorization and Authentication Service offers security contexts to all Resources according to the Access Control Markup Language standard (XACML)

D-NET standardization

- *Information Service* system mediation
  - All relevant *resources* register their profile into the IS; e.g. Services, collections, indexes, users, etc.
  - Services can access system relevant information through the IS, in a common standard way, with no need to statically know the locations of other Services
- *ResultSet* mechanism
  - Standard interfaces and tools for data exchange
  - By reference or value
  - Paging modes, transformation, caching

D-NET Framework Assumptions

- Service registration & discovery: infrastructure
- ResultSets: data exchange
- “Flattenizable” Metadata: generic services



Metadata format **Flat Metadata format**