On Constructing Repository Infrastructures
The D-NET Software Toolkit

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

Web Portal

OAI-PMH

Information Space

Storing

Indexing

Harmonizing

OAI-PMH Harvester

Repositories

Common Metadata Format

Local Metadata Format

OAI-PMH harvester

Extensions, updates, and refinements

Data workflows definition

Cleaning, enriching, transforming

Repository Administration
Repository Infrastructures
- Web Portal -

Web Portal
- Search
- Browse
- Alerts
- Profiling

Information Space

OAI-PMH

Repositories

Quality of Service
Extensions, updates, and refinements
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 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

Service Kits
- Web UI
- Search
- Index
- Store
- Transformator
- User Profiling
- Others

Dynamic, distributed Run-time Infrastructure

Enabling Layer Middleware

Repositories
Repository Administration Tools
Repository Infrastructures in D-NET

Deployment of web portals

Service Kits
- Web UI
- Search
- Index
- Store
- Transformator
- User Profiling
- Others

Dynamic, distributed Run-time Infrastructure

Web Portal
Aggregation system
Web Portal deployment
Modularity, customizability, sharing (and orchestration)

DRIVER Project
Modularity, customizability, sharing (and orchestration)

EFG Project

Authority File
Harvester
MDStore
Transformer
Index
Search
Web UI
OAI-PMH

Local Schema

Metadata Formats
Modularity, customizability, sharing (and orchestration)

OpenAIRE Project

OpenAIRE Export format: Dublin Core + project ID + License info

OpenAIRE Internal Format: OpenAIRE Export format + repo info

Metadata Formats

XML download

Project and Participants Format (from EC)

Search format: Project Papers Participant

Web UI
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

• 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, Uniwersytet Warszawski, Poland