Introducing CRIS Information in Repository Federations
The OpenAIRE Project Experience

Workshop on CERIF-CRIS and Repositories
Introducing CRIS Information in Repository Federations

- (Open Access) Repository federations
  - DRIVER EC project

- Repository federations and integration with CRIS-like information
  - OpenAIRE EC project
OA Repository federations

The DRIVER project experience
DRIVER project

Objectives

- Building an *Information Space* aggregating Open Access research publications from institutional repositories in Europe and making it available for access and elaboration to any interested third-party community

Brussels, 12th of September, 2011 - CERIF-CRIS and Repositories
**Issues**

- Scalability
  - Arbitrary number of repositories
  - Arbitrary number of users
  - Arbitrary number of consuming applications

- Data «harmonization»
  - Heterogeneous metadata formats (structure and semantics)
DRIVER technical objectives

Aggregation System

Data harvesting and harmonization

Community Application

Community Application

Community Application

OA Research Publications

Brussels, 12th of September, 2011 - CERIF-CRIS and Repositories
Adopted solution

**DRIVER and DRIVER-II**

- **Technology**
  - D-NET Software toolkit to implement aggregation system and custom applications

- **Networking - advocacy and dissemination**
  - DRIVER Guidelines for Repository Managers
  - Open Access mandates
  - Confederation of Open Access Repositories (COAR)
    - EC and beyond
Technology

D-NET Software Toolkit

End User Functionality Area
- Generic UI Service
- Recomm. Service
- Community Service
- Search Service
- User Service

Information Service
- OAI-PMH Pub Service
- Index Service
- Browse Service
- Feature Extr Service
- Store Service
- OAI-PMH Harv Service
- Validator Service
- Collection Service
- Transformation Service
- Authz&Authn Service
- Authority File Service
- OAI-ORE Service
- Enabling Area
- Comp Objs Service
- Object Packing Service
- Repo Manager Service

Repositories

http://www.d-net.research-infrastructures.eu
Technology

**D-NET Software Toolkit**

- **Customizability**
  - Services offer general-purpose functionalities, e.g., index and store can be configured to handle any metadata formats.

- **Compositionality**
  - Services can be combine into personalized workflows.

- **Scalability and robustness**
  - Content, functionality and users.

- **Autonomicity**
  - Services can self-administrate their actions and interactions.

- **Openness:**
  - Facilitating extension, update, integration of new functionality services.

**Used in many EC projects, i.e., DRIVER, EFG, EFG1914, HOPE, OpenAIRE, OpenAIREplus and by the national initiative Spain-Recolecta (Concorcio Madrono)**

Brussels, 12th of September, 2011 - CERIF-CRIS and Repositories
Networking

- DRIVER Guidelines for Repository Managers
  - DRIVER Metadata Format (DRIVER data model)
    - DC + vocabularies (dc:type) + value format
    - Repository info
    - Location
  - Protocol
    - Constraints on well-behaved OAI-PMH

- Wide international network
  - Seeds of COAR (Confederation of Open Access Repositories)
DRIVER production infrastructure

DRIVER search portal

National portals

End users

Functionality Layer

EU Open Access Repositories

Administrators

Data Layer

Enabling Layer

Brussels, 12th of September, 2011 - CERIF-CRIS and Repositories
DRIVER production infrastructure: content

- 309 harvested repositories
- 39 countries from Europe and beyond
- 3,680,000+ OA publication metadata records

http://search.driver.research-infrastructures.eu
CRIS and repository federations
The OpenAIRE project experience
OpenAIRE project

Objectives

- **Aim**: producing reliable statistics: to evaluate the impact of the EC Open Access mandate, i.e., **Special Clause 39**

- **How**: Building a European data infrastructure where EC project information can be interlinked with the publications they funded
DRIVER infrastructure

High-level data flows

End-users

Administrators

Search Stats

Articles

End-users

Third-party Applications (e.g., portals)

DRIVER Data Model

Articles

DRIVER Information Space

DRIVER infrastructure

DRIVER Compliant Repositories

Metadata: Biblio

Brussels, 12th of September, 2011 - CERIF-CRIS and Repositories
OpenAIRE Infrastructure

High-level data flows

End-users

Administrators

Search
Projects
Articles
Stats

OpenAIRE Data Model

Articles → Projects

Deposition/Ingestion

End-Users ("authors")

Metadata: Biblio, Project, License (Files)

OpenAIRE Compliant Repositories

Metadata: Biblio, Project, License

EC Databases

CORDA

Third-party applications

End-users

Orphan repository

OpenAIRE Information Space

OpenAIRE Infrastructure

Formats: KE-CRIS/OAR DC
proprietary (on request)

Protocols: SRW, OAI-ORE, OAI-PMH, HTTP
Functional requirements

- **System administrators**
  - Data “harvesting”
    - Repositories: articles
    - CORDA Databases: projects (http://www.cordis.europa.eu)

- **End-Users**
  - “Authors”
    - Article *ingestion* (orphan of a repository: metadata and files)
    - *Article deposition* (with a repository of reference: metadata)
  - “Anonymous”
    - Access to full-text search and faceted browsing
    - Access to data stats
Issues

- Scalability and metadata «harmonization»
  - Similar to DRIVER project
- Unavailability of CRIS-like information
  - Most repositories do not retain project information within publication metadata
- Stats precision
  - Duplicates must be identified and solved
Adopted solution

Technology
- **D-NET Toolkit services** for aggregation system
- Extension of D-NET with **new services** where needed: format conversion and authority file management
- **Invenio Repository** (CERN) for orphan repository

Networking
- EC mandatory policies: **SC39**
- **Confederation of Open Access Repositories (COAR)**
- **OpenAIRE guidelines** for content providers
  - DRIVER guidelines + license info + project info
OpenAIRE production infrastructure: content

- OpenAIRE compliant repositories: 40+
  - Around 15+ compliant but empty
- European Repositories (OpenDOAR): 850+
  - From OpenDOAR
- Publications: ~ 1,100
  - From both repositories and end-users
- Projects (FP7): 16,603
  - CORDA is the authoritative source

http://www.openaire.eu
Data acquisition

Current solutions

- **From EC databases (CORDA):**
  - CORDA is the authoritative source for projects
  - OpenAIRE information space synchronized with CORDA

- **From Repositories:**
  - “OpenAIRE Guidelines for repository managers”
    - Project information made available through OpenAIRE (e.g., add-ons for DSpace and ePrints UIs, OAI-PMH interfaces, HTTP search access)
  - Validation prior ingestion: D-NET Validator Services
Data acquisition

Current solutions

- From End-Users (authors or others on behalf of them):
  - Publication “ingestion”: Orphan Repository
  - Publication “deposition”, i.e., “claiming”
    - Browsing and selecting publications from DRIVER Information Space
    - By specifying DOIs

- Data curation by system administrators
  - Value “harmonization” (e.g., DRIVER-like)
  - Identification and resolution of duplicates
By EC mandatory policies

- **Author-centric approach**: “suggest/force” researchers to deposit their article metadata also in OpenAIRE, not only in their repositories of reference.

- **Article-centric approach**: “suggest/force” researchers to deposit their articles “out there” with special license/project “tags”.

- **EC-centric approach**: “suggest/force” project coordinators to deposit publication in EC databases (e.g., SESAM) adopting the strict ingestion-deposition approach of OpenAIRE.
Improving data acquisition

Studies and experimentations

OpenAIRE research activities (WP7)

- **CRIS systems**: define OpenAIRE compatibility for CRIS systems (i.e., inclusive of publication metadata) and enable registration and harvesting of CRIS systems
  - Knowledge Exchange CRIS/OAR initiative
  - Cooperation with EuroCRIS (ref. Keith Jeffery)
  - To be continued in OpenAIREplus...

- **Usage stats**
  - Collecting stats from repositories

- **Research data**
  - Linking publications and data
OpenAIREplus project

- Upgrading to CERIF’s data model
- Including research data
  - Data Access and Interoperability Task Force initiative (DAITF), in cooperation with EUDAT project
- Integrate the DRIVER infrastructure into the OpenAIRE infrastructure
- Extra functionalities
  - Include «user feedbacks»
  - Include advanced citation tools
  - Others...
Questions?
DRIVER Actors

System Managers:
DRIVER Consortium

User communities

Service providers:
DRIVER Consortium and user communities

Aggregator managers:
DRIVER Consortium and others

Data providers:
EU Repositories

Brussels, 12th of September, 2011 - CERIF-CRIS and Repositories
OpenAIRE Architecture

**Provision**
- Alerts
- Format
- Stats
- Usage stats

**Ingestion support (DRIVER, CrossRef)**

**Portal**
- access to full-text search and browse

**DB Service**
- Search-SRW
- Index
- Stats
- Repos
- Articles
- Projects

**Ingestion**
- deposition
- ingestion
- access to stats

**Information Space**
- Orphan

**Federation**
- OpenAIRE compliant Repositories (guidelines compliance)

**Harvester**

**Mediation**
- ePrints-DSpace UI publisher
- DC, project, CRIS-OAR
- OAI-PMH publisher
- OAI-ORE publisher

**Formats:**
- DC, project, CRIS-OAR

**Harvester**
- OpenDOAR Import Module
- EC DBs Import Module

**Transform**
- curation

**MDStore**
- A
- OpenAIRE format: DC + project + license
- B
- OpenAIRE format + repo info