The OpenAIRE Scholarly Communication Infrastructure
On Interlinking Datasets, Literature, Fundings, and Research Initiatives
OpenAIREplus project

The objectives (in a nutshell)

European data infrastructure for scholarly communication

- Facilitating discovery of research outcome across disciplines and Europe (and beyond) publications and datasets repositories
  - Promoting Open Access: publishing and citation best practices, business models for publications and datasets
  - Interlinking and contextualizing publications and datasets
- Measuring impact of research initiatives
  - Open Access vs non-Open Access
  - Funding schemes: return of investment
  - Research initiatives: research impact
- Providing both human and technical infrastructure to make this possible!
Technical infrastructure

Functional architecture

Usage policies

Publications

Datasets

Context (e.g. projects)

Information
Inference by Data Mining

Cleansing & Transforming

De-duplication

Data Source Validation

Import policies: OpenAIRE guidelines

CRIS systems (e.g. National projects)

Entity Registries (OpenDOAR, EC-CORDA)

Publications Institutional & Thematic repositories, OA Journals

Zenodo

Datasets Data repositories

Search & Browse

Get support (NOADs)

Statistics

Curate & collaborate (feedbacks from coordinators)

Deposit/Ingest (claim) Publications

Access for Third-party Service Providers

Get support (NOADs)

De-duplication

Data Source Validation

Dublin Core metadata + FP7 project + license info + registered in OpenDOAR

DataCite metadata + PID + link to pubs in OpenAIRE

Import policies: OpenAIRE guidelines

CRIS systems (e.g. National projects)

Entity Registries (OpenDOAR, EC-CORDA)

Publications Institutional & Thematic repositories, OA Journals

Zenodo

Datasets Data repositories

Get support (NOADs)

De-duplication

Data Source Validation

Dublin Core metadata + FP7 project + license info + registered in OpenDOAR

DataCite metadata + PID + link to pubs in OpenAIRE

Import policies: OpenAIRE guidelines

CRIS systems (e.g. National projects)

Entity Registries (OpenDOAR, EC-CORDA)

Publications Institutional & Thematic repositories, OA Journals

Zenodo

Datasets Data repositories

Get support (NOADs)

De-duplication

Data Source Validation

Dublin Core metadata + FP7 project + license info + registered in OpenDOAR

DataCite metadata + PID + link to pubs in OpenAIRE

Import policies: OpenAIRE guidelines

CRIS systems (e.g. National projects)

Entity Registries (OpenDOAR, EC-CORDA)

Publications Institutional & Thematic repositories, OA Journals

Zenodo

Datasets Data repositories

Get support (NOADs)

De-duplication

Data Source Validation

Dublin Core metadata + FP7 project + license info + registered in OpenDOAR

DataCite metadata + PID + link to pubs in OpenAIRE

Import policies: OpenAIRE guidelines

CRIS systems (e.g. National projects)

Entity Registries (OpenDOAR, EC-CORDA)

Publications Institutional & Thematic repositories, OA Journals

Zenodo

Datasets Data repositories

Get support (NOADs)

De-duplication

Data Source Validation

Dublin Core metadata + FP7 project + license info + registered in OpenDOAR

DataCite metadata + PID + link to pubs in OpenAIRE

Import policies: OpenAIRE guidelines

CRIS systems (e.g. National projects)

Entity Registries (OpenDOAR, EC-CORDA)

Publications Institutional & Thematic repositories, OA Journals

Zenodo

Datasets Data repositories

Get support (NOADs)

De-duplication

Data Source Validation

Dublin Core metadata + FP7 project + license info + registered in OpenDOAR

DataCite metadata + PID + link to pubs in OpenAIRE

Import policies: OpenAIRE guidelines

CRIS systems (e.g. National projects)

Entity Registries (OpenDOAR, EC-CORDA)

Publications Institutional & Thematic repositories, OA Journals

Zenodo

Datasets Data repositories

Get support (NOADs)

De-duplication

Data Source Validation

Dublin Core metadata + FP7 project + license info + registered in OpenDOAR

DataCite metadata + PID + link to pubs in OpenAIRE
OpenAIRE Data flow

OpenAIRE Portal: Discovery & Impact measure

Enriched Information Space

System Inference

End-users Feedbacks

Public Information Space

De-duplication

Native Information Space

Data source import

End-user claims
Importing from data sources

Repositories and CRIS systems

- Publication repositories (Dublin Core, OpenAIRE profile)
  - Publications with relationships to authors, projects, licenses, access rights, and repository
- Repositories aggregators (Dublin Core, OpenAIRE profile)
  - Publications with relationships to authors, projects, licenses, access rights, repository, and aggregator
- Dataset repositories (DataCite, OpenAIRE profile)
  - Datasets with relationships to authors, projects, and publications
- CRIS systems (CERIFXML, OpenAIRE profile)
  - Publications with relationships to authors, organizations, projects, etc.
  - Datasets with relationships to authors, organizations, projects, etc.
  - Projects with relationships to persons, organizations, fundings, etc.
Importing from data sources

**Entity registries**

- OpenDOAR (OpenDOAR HTTP)
  - Repositories with relationships to organizations
- EC CORDA (CORDA XML)
  - EC FP7 projects with relationships to persons and organizations
- Wellcome Trust (Wellcome Trust HTTP)
  - WT projects with relationships to organizations
- Re3data.org for data repositories?

- Exchanging data with data sources! Typically relationships to objects out of their domain
  - Publications and datasets
  - Context
Importing from end-users

**Depositing, claiming, feedback**

- **OpenAIRE Zenodo Deposition**: authors who are orphans of a repository of reference for publication and datasets can deposit file and metadata into the Zenodo repository.

- **CrossRef Claiming**: authors who have a repository of reference can “claim” their depositions into OpenAIRE by specifying the relative DOIs.

- **Feedback**: registered end-users (and OpenAIRE data curators) can give advice on how to enrich or fix the information space.
Information Space

- **Publication repositories:**
  - ~400 repositories
    - With links to EC projects: ~110 repositories + 20 Open Access Journals + 1 aggregator (NARCIS, NL)
    - With Open Access publications (DRIVER): ~390 repositories

- **Dataset repositories**
  - DataCite, DRYAD, Pangaea, others
Information Space

- **OpenAIRE Production system** (www.openaire.eu)
  - 40,000 publications linked to EC fundings

- **OpenAIREplus Beta system** (public release end of July 2013)
  - 8 Millions publications, including DRIVER OA publications, OpenAIRE publications, and datasets

- Datasets
  - DRYAD
  - DataCite
  - Geoscience journal?
Inferring data

• Relationship inference by mining publication PDFs
  • Project-publication: EC and National projects (e.g. WT)
  • Dataset-publication: DOI connections
  • Publication-publication: citations by bibliography
  • Publication-publication: text similarity

• Property inference by mining
  • Title of papers
  • Authors of papers
  • Organizations of authors at time of publication
Inference framework

OpenAIRE implements a framework and infrastructure for data inference services

- Inference service (pluggable and pipeline-able): applies an algorithm to an *input collection* and returns an *output collection* with a given degree of TRUST
  - Scope: publication/dataset files (e.g. PDF, XML), graph of objects, metadata
  - Actions: e.g. discipline-specific inference methods, NLP techniques

- Inference workflows: sequences of *inference steps* each instantiated by an inference service
  - Workflow’ intermediate and final collections are cached and can be used for testing new inference configuration or re-executing only part of the workflows
Provenance and TRUST

Each object in the information space is enriched with information relative to:

- **Data source** from which the object was collected
- **Workflow** performing the collection: data source import, end-user claim, end-user feedback, inference, etc.
- **Agent** responsible of the collection of the object: registered users, system driven mechanism
- **Trust**: value from 0 to 1 stating the level of TRUST of the object
De-duplicating data

“Equivalence relationships”

Equivalent objects are “merged” into one “representative object”

• **Publications**, the representative object stores:
  • All properties of the merged records (prioritized by degree of TRUST)
  • All files of the merged records: URLs, data source and license
  • All relationships to other objects of the merged records: projects, authors, publications, datasets, etc.

• **Organizations**, the representative object stores:
  • All properties of the merged records (prioritized by degree of TRUST)
  • All relationships to other objects of the merged records: projects, publications, datasets, etc.

• **Persons** (*ongoing*), the representative object stores:
  • All properties of the merged records (prioritized by degree of TRUST)
  • All relationships to other objects of the merged records: organizations, publications, datasets, etc.
Exploiting relationships
Measuring impact of “research initiatives”

- **Research initiative**: activity funding, favoring or developing science and willing to measure its impact in terms of research outcome
  - Example: European Commission, e-IRG EGI infrastructure, National funding agencies

- OpenAIRE supports research initiatives by providing:
  - **End-user claim services**: manually identifying research output linked to the relative initiative
  - **Relationship inference services**: tools to automatically identify publications linked to the research initiative

- **Current research initiatives**:
  - OA vs non-OA publications w.r.t. EC projects (facets: organizations, countries, EC funding schemes, EC subjects)
  - Publications w.r.t. EGI infrastructure (facets: EGI virtual organizations, EGI disciplines)
Enabling Technology

• D-NET Software toolkit
  • Service-oriented data infrastructure enabling technology
  • Adoption: Projects (DRIVER/II, OpenAIRE/plus, EFG/1914, HOPE, EAGLE) and nations (Spain-Recolecta, Poland, Belgium, Argentina (in progress))
  • By: CNR-ISTI (IT), Uni Athens (GR), ICM (PL), Uni Bielefeld (GE)

• INVENIO Repository
  • Customizable repository platform: workflows and data models
  • Adoption: CERN digital library and 30+ institutions world-wide
  • By: CERN (Switzerland), collaboration from DESY, EPFL, FNAL, SLAC
Data-publication Linking Issues

Data models for scholarly communication

Dataset modeling issues
- Data typology: e.g. raw data, secondary data, software, experiments
- Data granularity: e.g. DB, DB record, DB queries, DB query results
- Data scope: e.g. discipline-specific (BADC format), cross-discipline (INSPIRE), general-purpose (DataCite)

General issues
- Contextual entities: e.g. authors, organizations, patents, funding schemes, tools, devices
- Publication and data IPRs: e.g. access control, propagation of IPRs, ownership
- Provenance: e.g. location, hosting data source, generating agent (human, machine, device, facility)
- Relationships: e.g. semantics, modes (static, dynamic, inferred)

New Trends
- Enhanced publication data models, i.e. packaging existing publications and datasets into one new consumable object (research objects, executable objects)
Data-publication Linking Issues

Information systems for scholarly communication

Information System Typologies

• Data infrastructures: aggregation of existing pubs and datasets for discovery and re-use
• Data Centers, Journals, Repositories: deposition, publishing, preservation of pubs and datasets

Data infrastructure issues

• Top-down approach: e.g. integrating existing repositories
• Bottom-up approach: e.g. guidelines for repositories to facilitate data and publication linking, mandates to include references to fundings, best practices on citation
• Relationship inference: equivalence (de-duplication), reference, citation, etc.

General functionality issues

• Reading/visualization, discovery (search/browse/navigation), automated reuse, enrichment/annotation, cross-discipline investigation, research impact, community views (VREs), export APIs and exchange formats
Data-publication Linking  
**LCPD2013 Workshop**

- First Workshop on *Linking and Contextualizing Publications and Datasets*
- CFP to be made officially public in May, important dates:
  - Research paper submission: June 24th, 2013
  - Notification of acceptance: July 29th, 2013
  - Camera ready version: August 31st, 2013
  - Workshop day: 26th of September 2013
OpenAIRE project factsheet

- Coordination
  - University of Athens - GR
  - Goettingen University Library - DE
  - CNR-ISTI - IT

- Technical production & operation
  - 5 partners with expertise in technologies for Digital Libraries and Data Infrastructures

- General
  - Starting date: Dec 1, 2011
  - Duration: 30 months
  - Total budget: 5.2 Mi

- Scientific communities
  - EMBL-EBI – biology
  - DANS – social sciences
  - STFC/BADC – climate

- Networking Organization
  - 5 libraries, active in OA movement

- National Open Access Desks
  - All member states
  - Norway, Switzerland, Turkey, Iceland
Questions?

For more:
OpenAIRE infrastructure: http://www.openaire.eu
D-NET software: http://www.d-net.research-infrastructures.eu
INVENIO: http://invenio-software.org/