Paolo Manghi
National Research Council (IT)

Moving content across the OpenAIRE infrastructure boundaries

RDA Plenary
Data and computing infrastructures for open scholarship Workshop
Session 2: Linking up science: Interoperability aspects and services

22nd of September, 2015 - Paris
**Open Access Infrastructure for Research in Europe**

The point of reference for Open Access in Europe

- NOADS: National Open Access Desks
- Monitor and foster the adoption of Open Access policies at the local level
- Support researchers at the implementation of the Data Pilot
- Gold OA: FP7 post grant APCs Pilot

- e-infrastructure for monitoring impact of OA mandates and research projects
- OpenAIRE guidelines for metadata exchange
- Zenodo Repository for the deposition of research products

Human Network

50 Partners: EC countries, data centers, universities, libraries,

By Donatella Castelli and Alessia Bardi, September 2015, bardi@isti.cnr.it
OpenAIRE data model:
a view from the moon
Populating the OpenAIRE information space: making sense out of heterogeneity

OpenAIRE Guidelines (http://guidelines.openaire.eu):

- Literature Repositories (enriched qualified DC, in cooperation with SHARE-US, La Referencia-South America, JISC-UK),
- Data Archive (enriched qualified DataCite),
- CRIS systems (OpenAIRE/CERIF-XML)
## OpenAIRE data sources

(September 2015)

<table>
<thead>
<tr>
<th>Datasource typology</th>
<th>Number of datasource</th>
<th>Type of objects</th>
</tr>
</thead>
<tbody>
<tr>
<td>Journal Platform</td>
<td>5549</td>
<td>publications</td>
</tr>
<tr>
<td>Publication Repository</td>
<td>510</td>
<td>publications</td>
</tr>
<tr>
<td>Data Repository</td>
<td>38</td>
<td>datasets and publications</td>
</tr>
<tr>
<td>Aggregator of Publication Repositories</td>
<td>7</td>
<td>publications</td>
</tr>
<tr>
<td>Publication Catalogues</td>
<td>9</td>
<td>publications</td>
</tr>
<tr>
<td>Aggregator of Data Repositories</td>
<td>1</td>
<td>datasets</td>
</tr>
<tr>
<td>Aggregator/Publisher of Journals</td>
<td>4</td>
<td>publications</td>
</tr>
<tr>
<td>Entity Registry (data sources offering</td>
<td>11</td>
<td>data sources (publication repositories, data repositories),</td>
</tr>
<tr>
<td>authoritative lists of entities)</td>
<td></td>
<td>projects, funders</td>
</tr>
<tr>
<td>CRIS systems</td>
<td>0 (the first CRIS</td>
<td>publications, datasets, projects, persons</td>
</tr>
<tr>
<td></td>
<td>systems will be</td>
<td></td>
</tr>
<tr>
<td></td>
<td>aggregated by the</td>
<td></td>
</tr>
<tr>
<td></td>
<td>end of 2015)</td>
<td></td>
</tr>
</tbody>
</table>
OpenAIRE information space numbers (September 2015)

• 12M+ publications (de-duplicated)
• 200,000+ links publication-project from 5 funders
• 9,000+ datasets linked to publications or projects
• 34,000+ organizations (de-duplicated)
Opening up the OpenAIRE information space

- OpenAIRE APIs, programmatic access: http://api.openaire.eu
  - OAI-PMH (production ready)
  - REST Search (production ready)
  - Linked Open Data (delivery 2015, 2016)

- OpenAIRE Services
  - Portal search and statistics (production ready)
  - Literature Repository Notification Broker Service (delivery 2016)
  - Data–Literature Interlinking Service (BETA ready, integration in OpenAIRE to be planned)
Linked Open Data

• Specified an RDF vocabulary
  • Based on the OpenAIRE entity-relationship data model
  • Skipping complicated and unnecessary attributes of entities
  • Reused common vocabularies, e.g. CERIF, FOAF, BIBO, DC, ...

• Data conforming to LOD best practices to be published in BETA end of 2015
  • Entity descriptions downloadable from URIs + dumps + SPARQL endpoint

• Integration in OpenAIRE data provision workflow in 2016

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of input rows</td>
<td>20,985,097</td>
</tr>
<tr>
<td>Number of output rows</td>
<td>~</td>
</tr>
<tr>
<td>Time of indexing(s)</td>
<td>1.043</td>
</tr>
<tr>
<td></td>
<td>655,000,000</td>
</tr>
</tbody>
</table>
OpenAIRE RDF Graph: integration into the cloud

Identified Datasets to be interlinked to: DBLP, CiteSeer, CEUR-WS.org, and others
Literature Repository Notification Broker Service

- Literature repository subscriptions based on configurable criteria of publication-repository “closeness”
  - From repositories (and other data sources) to institutional repositories
  - For example, from journals or thematic repositories to institutional repositories

Service design in cooperation/alignment with SHARE US, JISC-UK, La Referencia (South America)
Literature Repository Notification Broker Service

Publication-repository “closeness”, relationships:

• publication-author-organization-repository
  • Inferred by mining PDF for affiliations and deduplicating organizations

• publication-author-repository
  • Inferred from occurrences of author articles within repositories

• publication-project-organization-repository
  • Inferred from deduplicating organizations
Data–Literature Interlinking Service

• In cooperation with RDA/WDS Publishing Data Service Working Group, to create: “an open, freely accessible, web based service that enables its users to identify datasets that are associated with a given article, and vice versa”

• Major scholarly communication stakeholders involved at different levels: providers, users, requirements
PDS–WG Stakeholders

Claire Austin
David Arctur
Amir Aryani (ANDS)
Geoff Bilder (CrossRef)
Timea Biro
Adrian Burton (ANDS) - co-chair
Ian Bruno (CCDC)
Sarah Callaghan
David Carlson
Jamus Collier (PANGAEA)
Suenje Dallmeier-Thiessen
Tim DiLauro
Ingrid Dillo
Rorie Edmunds
Janine Felden
Harkan Grudd
Siddeswara Guru
Laure Haak (ORCID)
John Helly
Francisco Hernandez
Simon Hodson
Richard Kidd (RSc)
Hylke Koers (Elsevier) - co-chair
Paolo Manghi (OpenAire)
Haralampos Marmanis
Caroline Martin
Jo McEntyre (EMBL – EBI)
Yolanda Meleco
Sheila Morrissey
Lyubomir Penev
Mohan Ramamurthy
Howard Ratner
Nigel Robinson (Thomson Reuters)
Sergio Ruiz (DataCite)
Uwe Schindler (PANGAEA)
Johanna Schwarz (Springer)
Martina Stockhause
Carly Strasser
Eefke Smit (STM)
Jonathan Tedds
Joachim Wackerow
Juanle Wang
Hua Xu
Eva Zanzerkia
Carol Goble
Jeffrey Grethe
System development and operation: OpenAIRE and PANGAEA

Core Data Model

Information Space
- PID resolution
- De-duplicating
- Harmonizing

Links collection

Data Sources

Examples:
- Pairs of DOIs
- DataCite records
- PANGAEA records
The Service (BETA)
http://dliservice.research-infrastructures.eu

Powered by:

- OpenAIRE D-NET software
- PANGAEA search engine
## DLI Links providers

<table>
<thead>
<tr>
<th>Content provider</th>
<th>Contributed links</th>
<th>Referred objects</th>
<th>Referred publications</th>
<th>Referred datasets</th>
<th>Referred objects of unknown typology</th>
</tr>
</thead>
<tbody>
<tr>
<td>OpenAIRE Resolver</td>
<td>0</td>
<td>7579</td>
<td>7579</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Datasets in DataCite</td>
<td>4444966</td>
<td>701126</td>
<td>28259</td>
<td>572337</td>
<td>100530</td>
</tr>
<tr>
<td>Cambridge Crystallographic Data Centre</td>
<td>1025640</td>
<td>733642</td>
<td>227736</td>
<td>505513</td>
<td>393</td>
</tr>
<tr>
<td>OpenAIRE</td>
<td>20682</td>
<td>10800</td>
<td>7579</td>
<td>3160</td>
<td>61</td>
</tr>
<tr>
<td>Elsevier</td>
<td>138976</td>
<td>73233</td>
<td>7383</td>
<td>65849</td>
<td>1</td>
</tr>
<tr>
<td>Thomson Reuters</td>
<td>48466</td>
<td>28540</td>
<td>4203</td>
<td>24326</td>
<td>11</td>
</tr>
<tr>
<td>Australian National Data Service</td>
<td>19086</td>
<td>6183</td>
<td>463</td>
<td>3545</td>
<td>2175</td>
</tr>
<tr>
<td>PubMed</td>
<td>0</td>
<td>7576</td>
<td>7576</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>System Deduction</td>
<td>2436020</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>PANGAEA</td>
<td>775128</td>
<td>283768</td>
<td>11722</td>
<td>248080</td>
<td>23966</td>
</tr>
<tr>
<td>DataCite Resolver</td>
<td>0</td>
<td>19035</td>
<td>0</td>
<td>19035</td>
<td>0</td>
</tr>
<tr>
<td>RCSB</td>
<td>175648</td>
<td>134444</td>
<td>44843</td>
<td>88503</td>
<td>98</td>
</tr>
<tr>
<td>CrossRef</td>
<td>0</td>
<td>300850</td>
<td>300850</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Lessons learnt
Challenges encountered in OpenAIRE

• Homogeneity of description: “they all want it, but it is not for free”
  • Guidelines are hard to be adopted: lack of funds, legacy content, old practices, multiple guidelines, discipline/community requirements, etc.

• Mining: lack of access to the payload (articles, datasets, experiment)
  • Open Access is definitely taking off, but still it is not the rule
  • PDFs are not often available

• Interlinking: products of science live in disconnected
What’s missing?

• Literature repositories find it hard to keep up with the demands of Open Scholarship in the digital era
  • Supporting interlinking with other products of science, e.g. datasets and experiments (e.g. software, services, workflows), alternative products
  • Supporting modern ways to export/exchange content (metadata and files) with publishers, other repositories, or third-party services
  • Advocate and support novel forms of research products and evaluation, e.g. readers feedback, downloads, etc.
What’s missing?

- Literature: cross-community
- Datasets: community-oriented
- Experiments: community or research groups oriented

Scholarly Communication workflows
Scientific Reward practices

Literature

Established

Research data

Experiments

Community e-infrastructure

Research data

Experiments

Community e-infrastructure

Community e-infrastructure
What is missing?

• The “infrastructural” elements necessary to enable Open Scholarship are mostly missing

• Beyond the PDF (sharing and citation): lack of metadata standards for several research-related entities: e.g. experiments (e.g. software, workflows, services), authors, funding, projects, etc.

• Contextualization: Lack of standards for interlinking datasets, articles, and experiments across community and tech boundaries

• Identification: Coexistence of several persistent identifier schemes

• Scientific reward: Lack of evaluation schemes for
We need scholarly communication infrastructures

- Foster and boost collective approach (community building and trust): round table of policy makers, funders, publishers, research organizations, data centers, scientists, general public and companies
  - OA mandates, OA data pilot
- Policies: making synergies on how science should be openly shared, communicated, reused, and evaluated cross-community and community level
  - OpenAIRE guidelines, OA mandates and data pilot advocacy
- Services: can provide service infrastructure support where it
Questions?

Contacts:

• Paolo Manghi, paolo.manghi@isti.cnr.it