



La experiencia de la Universitat Politècnica de Catalunya con la API de OpenAire

Toni Prieto - antonio.juan.prieto@upc.edu

Tres experiencias con OpenAire: taller online de repositorios de REBIUN
29 de octubre de 2019

Organizado por REBIUN y la Universidad de Las Palmas de Gran Canaria

Guión

1. Cómo utilizamos las APIs de OpenAire en la UPC:
 - a. Proyectos Comisión Europea
 - b. OAI-PMH: registros enriquecidos
2. Nuestra experiencia con la validación en la versión 4 de les guidelines (centrado en DSpace)

Contexto (I): 2017

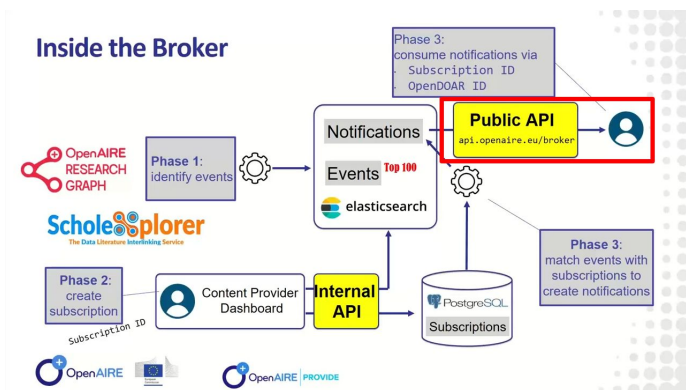
- En la UPC empezamos a utilizar las APIs de OpenAire en 2017
- Presentamos la experiencia en el 16º Workshop de REBIUN de Proyectos Digitales:

Prieto, Toni. **Enriquecimiento de registros mediante la API de OpenAire**. Ponència presentada a la Jornada Ecosistemas del Conocimiento Abierto (ECA 2017). Salamanca. Octubre del 2017 <<http://hdl.handle.net/2117/109379>>

- En el mismo Workshop se presentaban los avances con el nuevo Dashboard de OpenAire

Contexto (II): Actualidad y futuro

- OpenAire continua con el desarrollo y mejoras en el Dashboard.
- Se trabaja en nuevas formas de acceder a la información enriquecida:



Acceso en bloque via API a los eventos del broker

https://www.slideshare.net/OpenAIRE_eu/8th-content-providers-community-call

4SCIENCE
Share your knowledge

Enrich local data via the OpenAIRE Graph

Integration of the Notification Broker Service in DSpace 7 and DSpace-CRIS 7. This proposal is about the development of two data products for the repository community, built on top of the OpenAIRE Graph and the OpenAIRE Broker Service to enrich the data locally available and identify new data of interest reducing the friction of deposit

Integración automática

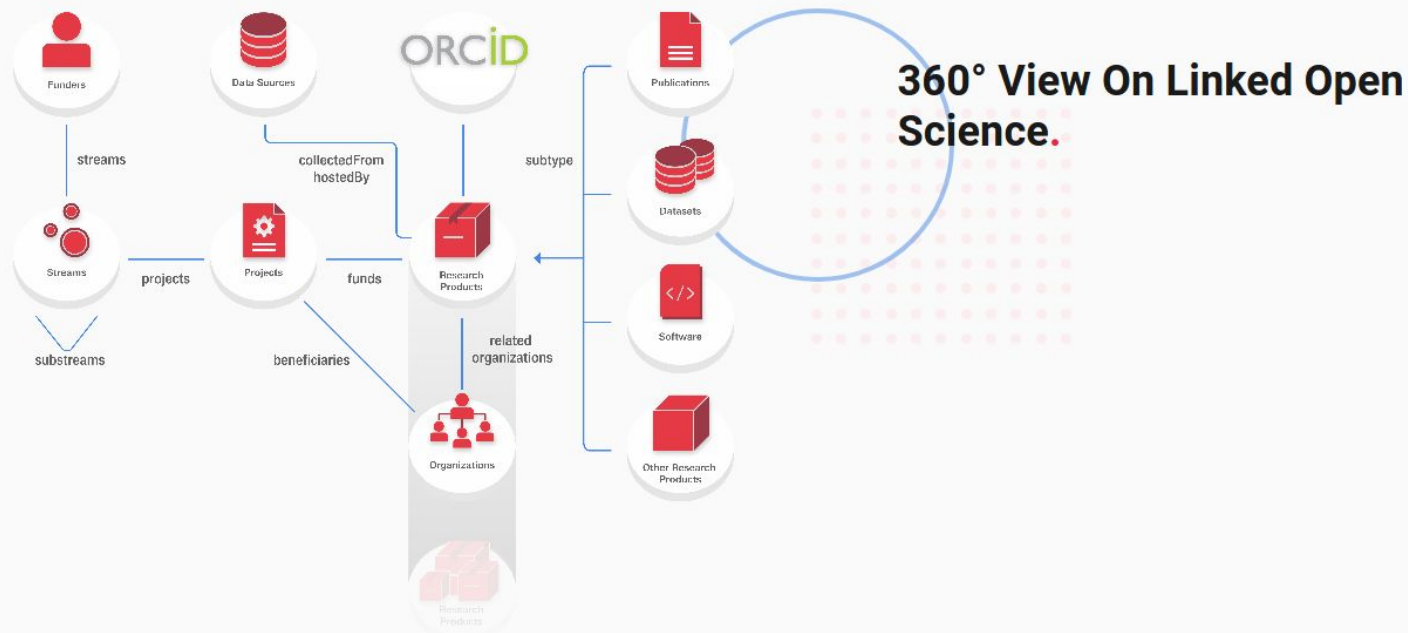
<https://www.openaire.eu/open-innovation-in-openaire>

Tres experiencias con OpenAire: taller online de repositorios de REBIUN

Contexto (III): OpenAire Research Graph

- Generado a partir de información de varias fuentes:
 - Registros oficiales: proyectos, investigadores, organizaciones, data sources
 - Repositorios y otras fuentes
- Diferentes procesos:
 - Limpieza
 - Validación
 - De-duplicación
 - Inferencia de información
- Ofrece (entre otras cosas):
 - Registro de proyectos (Comisión Europea y otros)
 - Registros enriquecidos de los repositorios recolectados

How it works



<https://graph.openaire.eu/>

APIs de OpenAire

Bulk access

OpenAIRE Research Graph Dumps
OAI-PMH
Bulk access to projects

Selective access via HTTP API

Publications
Research data
Software
Other Research Products
Projects

Linked Open Data

Documentation
OpenAIRE LOD Ontology
OpenAIRE RDF dump
SPARQL endpoint

<https://develop.openaire.eu/>

Uso de “Bulk access to projects”

- Recuperar el inventario completo de proyectos de la Comisión Europea (FP7 y H2020)
 - <http://api.openaire.eu/projects/dspace/H2020/ALL/ALL>
 - <http://api.openaire.eu/projects/dspace/FP7/ALL/ALL>
- Indexamos en un sistema propio para facilitar la selección:

Número de l'acord de subvenció (FP7, OpenAire, H2020, Llei de la Ciència, etc.):

info:eu-repo/grantAgreement/EC/FP7/246686/EU/Open Access Infrastructure for Research in Europe/**OPENAIRE**

info:eu-repo/grantAgreement/EC/FP7/283595/EU/2nd-Generation Open Access Infrastructure for Research in Europe/**OPENAIREPLUS**

info:eu-repo/grantAgreement/EC/H2020/643410/EU/Open Access Infrastructure for Research in Europe 2020/**OpenAIRE2020**

info:eu-repo/grantAgreement/EC/H2020/731011/EU/**OpenAIRE/**OpenAIRE-Connect

info:eu-repo/grantAgreement/EC/H2020/777541/EU/**OpenAIRE** Advancing Open Scholarship/**OpenAIRE-Advance**

- Esperamos poderlo hacer también con proyectos ofrecidos por Recolecta

Proyectos: guardar formato completo

Formato mínimo vs formato completo:

info:eu-repo/grantAgreement/Funder/FundingProgram/ProjectID



info:eu-repo/grantAgreement/Funder/FundingProgram/ProjectID/[Jurisdiction]/[**ProjectName**]/[**ProjectAcronym**]

A la espera de DSpace 7.x que espera incluir objetos específicos para recoger la información de los proyectos

Filtros y visualización

Proyecto de la Comisión Europea

ROMOL - Riding on Moore's Law (EC-FP7-321253) (77)

COMP-DES-MAT - Advanced computational design of materials (EC-FP7-320815) (69)

EFFINET - Efficient Integrated Monitoring and Control of Distribution Networks (EC-FP7-318556)

METRO-HAUL - METRO Haul 5G Application-aware optical edge storage, compute and network (EC-H2020-761727) (56)

INTELLACT - Intelligent observation and execution of Actions and Models (EC-FP7-269959) (49)

... Muestra más

Buscar

Proyecto de la Comisión Europea: ROMOL - Riding on Moore's Law (EC-FP7-321253) x

Mostrando ítems 1-10 de 77



Efficiency
Barredo F
Planas, M
Artículo.
Acceso al
Moore's L



Efficiency analysis of modern vector architectures: vector ALU sizes, core counts and clock frequencies

Barredo Ferreira, Adrián
Cebrián González, Juan Manuel
Valero Cortés, Mateo
Casas Guix, Marc
Moreto Planas, Miquel

Tipo de documento: Artículo

Fecha de publicación: 2020-03

Condiciones de acceso: Acceso abierto

Proyecto de la Comisión Europea

ROMOL - Riding on Moore's Law (EC-FP7-321253)

Resumen

Moore's Law predicted that the number of transistors on a chip would double approximately

OAI-PMH : registros enriquecidos

- OpenAire ofrece un set específico con los registros enriquecidos de cada repositorio.
- Estos registros los descargamos de este OAI-PMH y los indexamos localmente.
- Creamos tareas de curación de DSpace específicas que utilizan esta información para:
 - ✓ Encontrar posibles duplicados
 - ✓ Localizar fuentes en abierto en otros repositorios
 - ✓ Enlazar proyectos Europeos no informados en nuestro repositorio
 - ✓ Recuperar identificadores no incluídos (DOIs, Pubmed, ...)
 - ✓ Referencias

1) Posibles duplicados

- Número importante de duplicados reportados (primeras pruebas unos 1000)
- Grado de fiabilidad diverso:
 - Depende de información enviada (más fiable con identificadores asociados)
 - Casos especiales:
 - Secciones de revistas (editorial, ...)
 - Trabajos finales de carrera: co-autorías o proyectos recurrentes.
- **Centrados en casos específicos** a resolver manualmente:
 - Duplicados generados por documentos de diferentes fuentes (CRIS vs entidades vinculadas)
 - Documentos con el mismo archivo (comprobación extra por checksums)

2) Localizar fuentes en abierto

- Inicialmente un listado de unos 1000 documentos
- A comprobar manualmente:
 - Posibles falsos duplicados
 - Sustitución manual
- Diversa fiabilidad: priorizar fuentes “fiables” en la comprobación
- Principales fuentes:
 - arXiv, Zenodo, HAL, Digital.CSIC, DOAJ, ...

3) Enlazar proyectos

- Aprox. 1000 referencias a proyectos europeos añadidos.
- Gran porcentaje recuperados por **procesos de inferencia en texto completo** de OpenAire.
- Consideramos el valor de confianza asociado:
 - Los añadimos automáticamente a partir de un nivel de confianza

4) Recuperar identificadores

- Útil para incluir identificadores inicialmente no utilizados:
 - Ejemplo con pubmed ID
- DOI : Importante como opción para conectar con otras APIs.
 - Posibles errores: identificador vs relacionado
- Posibilidad de ampliar la experiencia a ORCIDs.
 - También la API Crossref como fuente

5) Referencias

- Incluidas mediante proceso de inferencia a través de Grobid (suponemos)
 - Resultado dispares según experiencia propia
- Posibles alternativas:
 - API Crossref
 - Open citations?

¿Porque utilizamos la API?

- Experiencia previa a la existencia del actual Dashboard.
- Posibles “ventajas”:
 - Posibilidad de extraer información para trabajar/filtrar fuera del dashboard.
 - Posibilidad de usar información local al hacer las comprobaciones (ejemplo archivos iguales en duplicados).
 - Poder realizar cambios directos en el repositorio (ejemplo incluir relaciones proyectos).

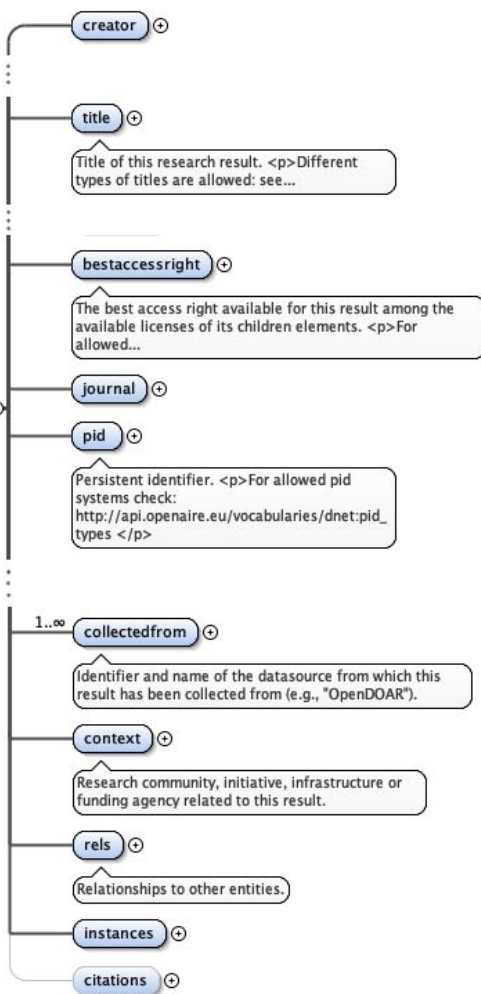
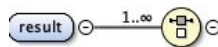
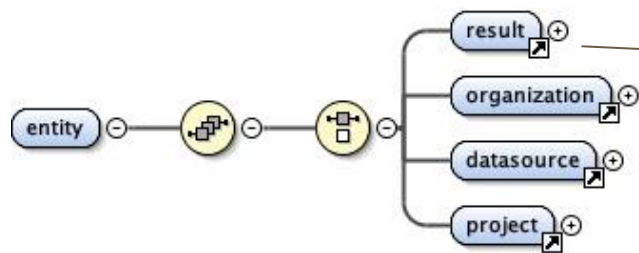
Cómo acceder a la información de vuestro set

- Listar los set disponibles:
 - http://api.openaire.eu/oai_pmh?verb=ListSets
- Buscar el nombre del set enriquecido de vuestro repositorio (acabado en _enriched):

```
▼<set>  
  <setSpec>UPCommons__Portal_del_coneixement_obert_de_la_UPC_enriched</setSpec>  
  <setName>UPCommons__Portal_del_coneixement_obert_de_la_UPC_enriched</setName>  
  ▼<setDescription>
```

- Recuperar los registros enriquecidos en bloques de 100 con consultas a “ListRecords” con el format oaf (OpenAire Format):
 - http://api.openaire.eu/oai_pmh?verb=ListRecords&metadataPrefix=oaf&set=UPCommons__Portal_del_coneixement_obert_de_la_UPC_enriched

Formato OAF



<https://www.openaire.eu/schema/latest/doc/oaf.html>

Nuestra experiencia con la validación Guidelines v4. para repositorios de literatura

- Tema complejo, seguramente daría para otra sesión, solo unos comentarios:
- Novedades (<https://guidelines.openaire.eu/en/latest/literature/index.html>):
 - Utilización de un esquema basado en Dublin Core, Datacite y esquema propio (oaire)
 - Incluye esquemas de identificación de autores, organizaciones, financiadores,
 - Uso de vocabularios COAR
- Nuestra experiencia:
 - Experiencia previa con la validación del repositorio de datos
 - Validados aunque no recolectado (?)

Objetos complejos en DSpace

- Las nuevas guidelines requieren informar sobre “objetos complejos” para **mejorar su identificación**:
 - Autores con información relacionada (identificadores ORCID)
 - Proyectos en formato estructurado con identificadores
- Soluciones diferentes en función de la implementación local
- Solución natural con software que pueda almacenar esta información:
 - DSpace: Entidades en DSpace7 / DSpace-CRIS

Implementaciones (DSpace)

- Versión DSpace 7.x facilitará la implementación:
 - Sistema de entidades
- Versiones previas (5x, 6x):
 - Implementación de 4Science (pendiente añadir en nuevas release):
 - <https://www.4science.it/en/2020/05/22/dspace-5-6-extension-now-available-to-support-orcid-and-new-openaire-guidelines/>
 - <https://github.com/DSpace/DSpace/pull/2709> (cambios en xsl y núcleo de indexación)
 - <https://github.com/DSpace/DSpace/pull/2859> (en futuro 6.4, solo cambios núcleo de indexación)

¡Gracias!