

# Enriquecimiento de registros mediante la API de OpenAire

16º Workshop de REBIUN de Proyectos Digitales  
7ª Jornadas OS-Repositorios  
11º Coloquio Internacional de Ciencias de la Documentación  
Salamanca 25, 26 y 27 de octubre de 2017

Antonio Juan Prieto Jiménez

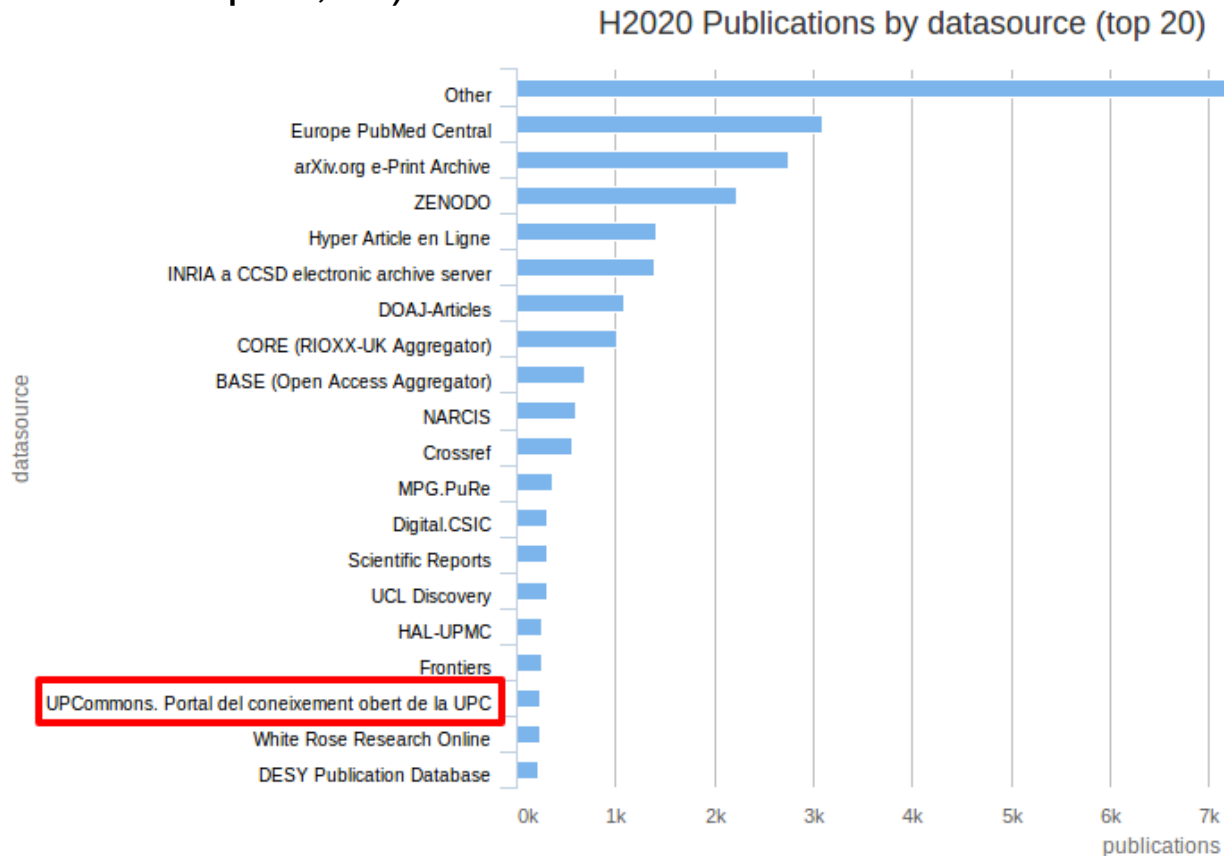


UNIVERSITAT POLITÈCNICA  
DE CATALUNYA  
BARCELONATECH

*Campus d'Excel·lència Internacional*

# UPC y OpenAire

- ~ 85.000 documentos
- 78 proyectos del programa H2020 UPC  
( + 59 Barcelona Supercomputing Center, + 43 Institut de Ciències Fotòniques, ...)



# Registros enriquecidos

- OpenAire crea registros enriquecidos de:
  - La fusión de registros de diferentes fuentes
  - Aplicar procesos sobre el texto completo





- La API ofrece estos registros enriquecidos vía OAI
  - set específico “enriched”
- Más otras posibles consultas:
  - Lista de proyectos
  - Consultas selectivas HTTP

## OpenAIRE API documentation

26 May 2017

---

### Table of Contents

1. Overview
2. Bulk access via OAI-PMH
3. Bulk access to projects  
DSpace/ePrints
4. Selective access with HTTP API  
Common API parameters  
Publications  
Datasets  
Projects

- Aprovechar esta información para enriquecer nuestros registros:
  - Añadir proyectos europeos no informados
  - Recuperar versiones en abierto
  - Duplicados
  - ...



## Pero antes...

- Hacer más visible la información sobre los proyectos asociados
- Mejorar la información que podíamos mostrar de los proyectos pasando del código mínimo al completo

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



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



# Proyectos: UI

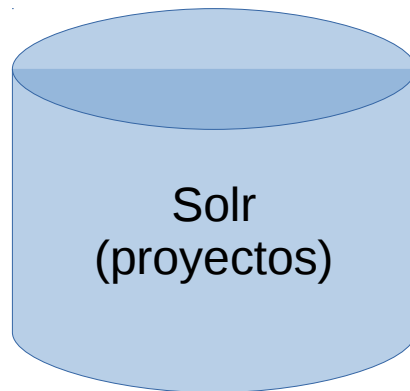
- Valores existentes → Tarea de curación de DSpace que obtiene el valor completo de la API de OpenAire
- Nuevos valores → Mejoras en la interfaz de entrada y edición:

**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** 2020/12345  
info:eu-repo/grantAgreement/EC/FP7/234313/EU/Optimisation for low Environmental Noise impact AIRcraft/OPENAIR  
info:eu-repo/grantAgreement/EC/FP7/232366/EU/Trap for the novel control of horse-flies on open-air fields/TABANOID  
info:eu-repo/grantAgreement/EC/H2020/643410/EU/Open Access Infrastructure for Research in Europe 2020/**OpenAIRE2020**

# Registro proyectos

- Carga local en Solr de los proyectos FP7 y H2020 obtenidos en bloque de la API de openaire



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/ <b>OPENAIRE</b>            | 2020/12345 |
| info:eu-repo/grantAgreement/EC/FP7/234313/EU/Optimisation for low Environmental Noise impact AIRcraft/OPENAIR              |            |
| info:eu-repo/grantAgreement/EC/FP7/232366/EU/Trap for the novel control of horse-flies on open-air fields/TABANOID         |            |
| info:eu-repo/grantAgreement/EC/H2020/643410/EU/Open Access Infrastructure for Research in Europe 2020/ <b>OpenAIRE2020</b> |            |

- Futuro, poder incluir también proyectos financiados por planes estatales



# Visualización de proyectos

Filtra

## Proyecto de la Comisión Europea

EFFINET - Efficient Integr **Buscar**  
Real-time Monitoring and C  
of Drinking Water Networks  
FP7-318556) (54)

HYCON2 - Highly-complex  
networked control systems  
FP7-257462) (2)

ISENSE - Making Sense of  
Nonsense (EC-FP7-270428

EMBOCON - Embedded  
Optimization for Resource  
Constrained Platforms (EC  
FP7-248940) (1)

WIDE - Decentralized and  
Wireless Control of Large-S  
Systems (EC-FP7-224168)

Mostrando items 1-10 de 54



**On the assessment  
applied to drinking v**  
Grosso Pérez, Juan  
Carlos; Puig Cayuel  
Texto en actas de c  
Acceso abierto  
Water systems are  
influences such as  
paper, two different



**A distributed comm:  
networks**   
Tedesco, Francesc  
Vicenç (Institute of  
Texto en actas de c  
Acceso restringido  
This paper proposes  
the operational cont  
to this kind of mana

## On the assessment of tree-based and chance-constrained predictive control approaches applied to drinking water networks



Grosso Pérez, Juan Manuel   
Maestre Torreblanca, José María  
Ocampo-Martínez, Carlos   
Puig Cayuela, Vicenç

**Tipo de documento:** Texto en actas de congreso

**Fecha de publicación:** 2014

**Editor:** International Federation of Automatic Control (IFAC)

**Condiciones de acceso:** Acceso abierto

Proyecto de la Comisión Europea

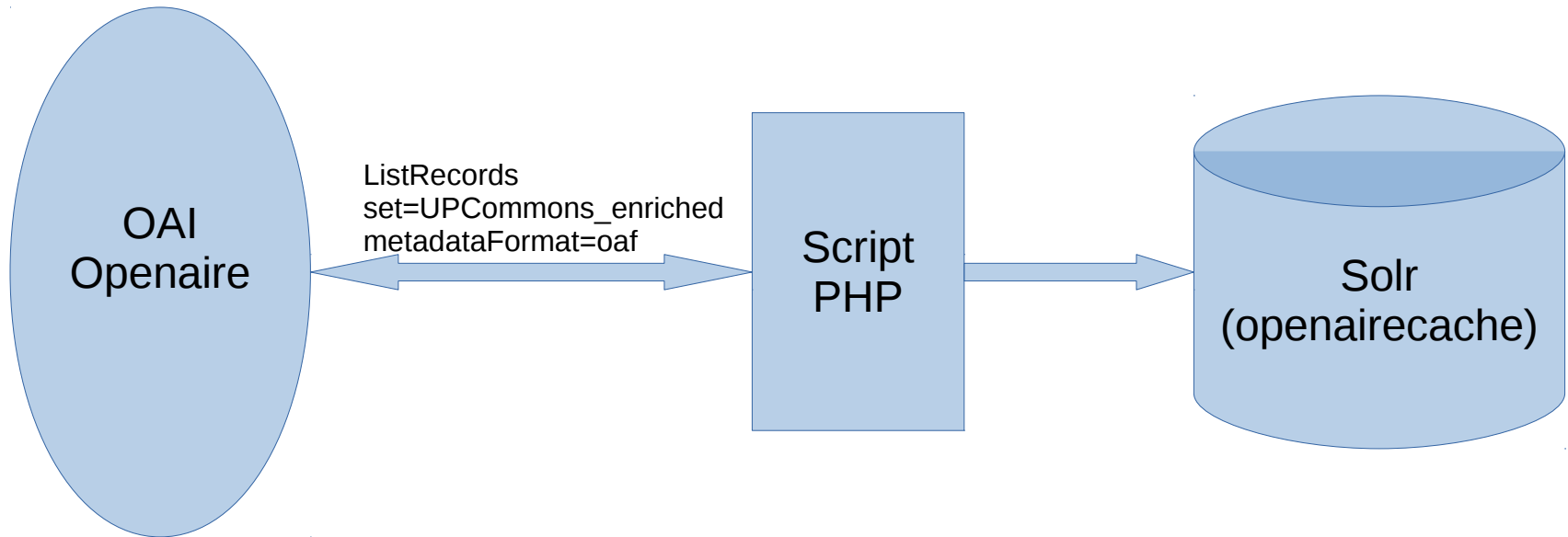
[HYCON2 - Highly-complex and networked control systems \(EC-FP7-257462\)](#)  
[EFFINET - Efficient Integrated Real-time Monitoring and Control of Drinking Water Networks \(EC-FP7-318556\)](#)

Resumen

Water systems are a challenging problem because of their size and exposure to uncertain influences such as the unknown demands or the meteorological phenomena. In this paper, two different stochastic programming approaches are assessed when controlling a drinking water network: chance-constrained model predictive control (CC-MPC) and tree-based model predictive control (TB-MPC). Under the former approach, the disturbances are modeled as stochastic variables with non-stationary uncertainty description, unbounded support and quasi concave probabilistic distribution. A deterministic equivalent... [\[+\]](#)

Ver/Abrir

[1506-On-the-Assessment-of-Tree-Based-and-Chance-Constrained-Predictive-Control-Approaches-applied-to-Drinking-Water-Networks.pdf \(420.0Kb\)](#)



- 59.948 registros enriquecidos de 84.448 registros recolectados
- Utilizando formato propio oaf (vs oai\_dc). Ofrece información sobre valores de confianza



## 4 objetivos

---

- 4 tareas de curación:
  - Incluir proyectos europeos no informados
  - Encontrar duplicados
  - Localizar otras fuentes abiertas para documentos en acceso restringido
  - Incluir referencias bibliográficas



# Resultado: proyectos añadidos

## **Objetivo:**

Recuperar los proyectos europeos no informados en el repositorio y añadirlos

- 1060 nuevos proyectos detectados (alto porcentaje: ~ 600 proyectos previamente informados)



# Resultado: proyectos añadidos

---

**Objetivo:**

Recuperar los proyectos europeos no informados en el repositorio y añadirlos

- Muchos valores inferidos por procesos de OpenAire
- Aplicado automáticamente si supera un umbral de confianza



# Resultado: duplicados

## **Objetivo:**

Detectar posibles duplicados en el repositorio

- Inicialmente muchos duplicados (1558)
- Porcentaje de duplicados mal inferidos



# Resultado: duplicados

## **Objetivo:**

Detectar posibles duplicados en el repositorio

- Segunda comprobación comparando checksum
  - 736 duplicados para comprobar y resolver manualmente
  
- Motivos diversos:
  - Documentos entrados antes de la integración con el CRIS
  - Duplicados con revistas editadas en la UPC vs producción
  - Trabajos académicos con dos autores
  - ...



# Resultado: fuentes abiertas

## **Objetivo:**

Detectar versiones en abierto en otras fuentes de publicaciones en acceso restringido en el repositorio

- Listado con 977 documentos
  
- A comprobar **manualmente**





# Resultado: fuentes abiertas

## **Objetivo:**

Detectar versiones en abierto en otras fuentes de publicaciones en acceso restringido en el repositorio

- Posibles falsos duplicados
- Diversa fiabilidad de las fuentes
  
- Principales fuentes:
  - 277 arXiv
  - 138 Digital.CSIC
  - 71 Zenodo
  - 53 DOAJ
  - 39 Europe PMC
  - ...



# Resultado: referencias

## **Objetivo:**

Recuperar referencias bibliográficas extraídas por OpenAire y mostrarlas en el repositorio

- 1462 documentos con referencias extraídas
- Actualmente probando la misma herramienta que utiliza OpenAire directamente contra el repositorio.
  - GROBID (<https://github.com/kermitt2/grobid>)

## Internet traffic and the behavior of processing workloads



Ver/Abrir

ACACES09-Ruken.pdf  
(78.16Kb)

Comparte:



Ver Estadísticas de uso

Mostrar el registro completo del ítem

Zilan, Ruken  
Verdú Mulà, Javier   
García Vidal, Jorge   
Nemirovsky, Mario  
Valero Cortés, Mateo

**Tipo de documento:** Texto en actas de congreso

**Fecha de publicación:** 2009-06

**Condiciones de acceso:** Acceso abierto

### Resumen

Nowadays, the evolution of network services provided at the edge of Internet increases the requirements of network applications. Such applications result in complexities thus, the processors need to execute more complex workloads that can deal not only with the packet header, but also with the packet payload (e.g. Deep Packet Inspection). Unlike common routing applications that show similar processing among packets, next-generation of network applications present variations in the processing procedure among packets. Thus, different traffic behaviors can produce different process patterns and... [+]

**URI:** <http://hdl.handle.net/2117/8385>

### Referencias

- [HVG05] R. Holanda, J. Verdú, J. García, and M. Valero. Performance analysis of a new packet trace compressor based on TCP flow clustering. In Proc. of ISPASS'05, pages 219–225, Austin, TX, March 2005.
- [RWW05] Ramaswamy Ramaswamy, Ning Weng, and Tilman Wolf. Analysis of network processing workloads. In Proc. of ISPASS'05, pages 226–235, Austin, TX, March 2005.
- [VGNV05] Javier Verdú, Jorge García, Mario Nemirovsky, and Mateo Valero. Architectural impact of stateful networking applications. In Proc. of ANCS'05, NY, USA, 2005. ACM.
- [ZBPS02] Y. Zhang, L. Breslau, V. Paxson, and S. Shenker. The characteristics and origins of internet flow rates, 2002.



- OpenAire es una fuente útil de información
- Procesos de extracción potentes
- Ampliable a más comprobaciones: Datasets?, ...



- Fiabilidad de los duplicados
- Es necesaria una revisión manual en algunos casos

# ¡Gracias!

antonio.juan.prieto@upc.edu



**UNIVERSITAT POLITÈCNICA  
DE CATALUNYA  
BARCELONATECH**

*Campus d'Excel·lència Internacional*