

# Indexación y posicionamiento de los repositorios en motores de búsqueda

XIV Workshop REBIUN de Proyectos Digitales  
VI Jornadas de OS-Repositorios  
Los horizontes de los repositorios  
Córdoba, 11 a 13 de marzo de 2015

Antonio Juan Prieto Jiménez



**CRUE**

**REBIUN**

Red de Bibliotecas Universitarias



**UNIVERSITAT POLITÈCNICA DE CATALUNYA**  
**BARCELONATECH**

Servei de Biblioteques, Publicacions i Arxius

# Antes de empezar



Lea las instrucciones



de este Medicamento y



consulte al Farmacéutico



CRUE

REBIUN

Red de Bibliotecas Universitarias



UNIVERSITAT POLITÈCNICA DE CATALUNYA  
BARCELONATECH

Servei de Biblioteques, Publicacions i Arxius

# Este taller...



Repasa conceptos de indexación en motores de búsqueda



Gran parte basado en DSpace



Para más información, consulte las fuentes oficiales



**CRUE**

**REBIUN**

Red de Bibliotecas Universitarias



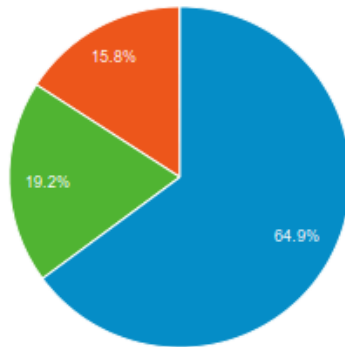
**UNIVERSITAT POLITÈCNICA DE CATALUNYA**  
**BARCELONATECH**

Servei de Biblioteques, Publicacions i Arxius

# Importancia motores de búsqueda

Sesiones por Tipo de tráfico

■ organic ■ referral ■ direct ■ twitter ■ elevator ■ Otras



Estadísticas de acceso Año/Mes

Fecha	Visitas	Downloads
2015-02	39086	108563
2015-01	89776	183771
2014-12	104026	232332
2014-11	78816	235795
2014-10	87203	246439
2014-09	78515	177420
2014-08	42274	129194
2014-07	111303	140282
2014-06	72144	149946
2014-05	43102	189136
2014-04	46035	174803
2014-03	45816	171566
2014-02	36237	177569
2014-01	55015	130171



CRUE

REBIUN

Red de Bibliotecas Universitarias



UNIVERSITAT POLITÈCNICA DE CATALUNYA  
BARCELONATECH

Servei de Biblioteques, Publicacions i Arxius

- Conceptos generales de indexación
- Google scholar
- Metatags HTML
- Herramientas webmaster y analíticas
- “Futuro?”: Datos estructurados



# Conceptos generales de indexación

---

- Software actualizado
- Sitemaps
- Robots.txt
- Últimos registros añadidos
- Redirecciones de página de descarga
- ¿Se utiliza el protocolo OAI?
- Sobre posicionamiento



# Software actualizado

- **Primera recomendación: software actualizado**
- Cambios / mejoras en cada versión

Nombre tiquet JIRA DSpace	
<a href="https://jira.duraspace.org/browse/DS-####">https://jira.duraspace.org/browse/DS-####</a>	
<i>Aplicado DSpace #.#</i>	

- Mejora
- Bug
- Nueva funcionalidad



CRUE

REBIUN

Red de Bibliotecas Universitarias



UNIVERSITAT POLITÈCNICA DE CATALUNYA  
BARCELONATECH

Servei de Biblioteques, Publicacions i Arxius

- Método para informar a motores de búsqueda del listado de páginas a rastrear a través de un listado de enlaces del sitio
- Permite rastreos más inteligentes
- Metadatos adicionales:
  - Última actualización
  - Frecuencia de modificación
  - Importancia
- Principalmente en XML, aunque se puede usar otros formatos (HTML, texto, ...)
- Si contiene muchas enlaces (+ 50.000) o pesa mucho (+ 10 MB) se utiliza un índice y se reparte en varios enlaces
- Mas información: <http://www.sitemaps.org>



- Informar a motores de búsqueda:
  - Herramientas de Webmaster
  - Inclusión en robots.txt

Sitemap: <http://my.dspace.url/sitemap>

Sitemap: <http://my.dspace.url/htmlmap>

**Robots.txt should include a SiteMap entry**

<https://jira.duraspace.org/browse/DS-1936>

*Aplicado DSpace 5.0*

# Sitemaps (DSpace)

- Desde versión 1.5
- Activar en cron:

```
# Regenerate sitemaps at 6:00 AM local time each morning  
0 6 * * * [dspace]/bin/dspace generate-sitemaps
```

- Genera dos formatos de sitemaps:
  - HTML Sitemaps: [dspace.url]/htmlmap
  - Google (XML) Sitemaps: [dspace.url]/sitemap
- Google XML comprimido
- HTML Sitemaps incluido en el pie de la interfaz DSpace

```
<a href="/htmlmap"></a>
```



CRUE

REBIUN

Red de Bibliotecas Universitarias



UNIVERSITAT POLITÈCNICA DE CATALUNYA  
BARCELONATECH

Servei de Biblioteques, Publicacions i Arxius

# Sitemaps XML (ejemplo)

## http://upcommons.upc.edu/sitemap

```
-<sitemapindex>
  -<sitemap>
    <loc>http://upcommons.upc.edu/sitemap?map=0</loc>
    <lastmod>2015-03-02T05:34:18Z</lastmod>
  </sitemap>
  -<sitemap>
    <loc>http://upcommons.upc.edu/sitemap?map=1</loc>
    <lastmod>2015-03-02T05:34:18Z</lastmod>
  </sitemap>
</sitemapindex>
```

## http://upcommons.upc.edu/sitemap?map=0

```
-<urlset>
  -<url>
    <loc>http://upcommons.upc.edu/handle/123456789/9</loc>
  </url>
  -<url>
    <loc>http://upcommons.upc.edu/handle/123456789/1</loc>
  </url>
  -<url>
    <loc>http://upcommons.upc.edu/handle/123456789/2</loc>
  </url>
  -<url>
    <loc>http://upcommons.upc.edu/handle/123456789/5</loc>
  </url>
  -<url>
    <loc>http://upcommons.upc.edu/handle/123456789/10</loc>
  </url>
  -<url>
    <loc>http://upcommons.upc.edu/handle/123456789/11</loc>
  </url>
  ...
  ...
  -<url>
    <loc>http://upcommons.upc.edu/handle/2099.1/19292</loc>
    <lastmod>2013-10-16T04:15:16Z</lastmod>
  </url>
  -<url>
    <loc>http://upcommons.upc.edu/handle/2099.1/20759</loc>
    <lastmod>2014-12-15T04:30:15Z</lastmod>
  </url>
  -<url>
    <loc>http://upcommons.upc.edu/handle/2099.1/17246</loc>
    <lastmod>2013-10-15T04:15:14Z</lastmod>
  </url>
  -<url>
    <loc>http://upcommons.upc.edu/handle/2117/116</loc>
    <lastmod>2014-05-06T13:16:28Z</lastmod>
  </url>
```



# Sitemaps HTML (ejemplo)

<http://upcommons.upc.edu/htmlmap>

- [sitemap 0](#)
- [sitemap 1](#)
- [sitemap 2](#)
- [sitemap 3](#)
- [sitemap 4](#)
- [sitemap 5](#)
- [sitemap 6](#)
- [sitemap 7](#)
- [sitemap 8](#)
- [sitemap 9](#)

...



- <http://upcommons.upc.edu/handle/2099/7215>
- <http://upcommons.upc.edu/handle/2099/9490>
- <http://upcommons.upc.edu/handle/2099/10498>
- <http://upcommons.upc.edu/handle/2099.1/5054>
- <http://upcommons.upc.edu/handle/2099.1/5055>
- <http://upcommons.upc.edu/handle/2099.1/10872>
- <http://upcommons.upc.edu/handle/2099.1/5056>
- <http://upcommons.upc.edu/handle/2099.1/7929>
- <http://upcommons.upc.edu/handle/2099.1/5057>
- <http://upcommons.upc.edu/handle/2099.1/5058>
- <http://upcommons.upc.edu/handle/2099.1/6723>
- <http://upcommons.upc.edu/handle/2099.1/5059>
- <http://upcommons.upc.edu/handle/2099.1/5060>
- <http://upcommons.upc.edu/handle/2099.1/9510>
- <http://upcommons.upc.edu/handle/2099.1/5061>
- <http://upcommons.upc.edu/handle/2099.1/6724>
- <http://upcommons.upc.edu/handle/2099.1/8018>
- <http://upcommons.upc.edu/handle/2099.1/6725>
- <http://upcommons.upc.edu/handle/2099.1/8019>
- <http://upcommons.upc.edu/handle/2099.1/8020>
- <http://upcommons.upc.edu/handle/2099.1/10843>
- <http://upcommons.upc.edu/handle/2099.1/7944>



CRUE

REBIUN

Red de Bibliotecas Universitarias



UNIVERSITAT POLITÈCNICA DE CATALUNYA  
BARCELONATECH

Servei de Biblioteques, Publicacions i Arxius

# Sitemaps: extensiones

- Posibilidad de añadir nuevos campos. Ejemplo vídeos:

```
<urlset xmlns=http://www.sitemaps.org/schemas/sitemap/0.9 xmlns:video="http://www.google.com/schemas/sitemap-video/1.1">
  <url>
    <loc>http://www.example.com/videos/pagina\_destino\_un\_video.html</loc>
    <video:video>
      <video:thumbnail_loc>http://www.example.com/thumbs/123.jpg</video:thumbnail_loc>
      <video:title>Barbacoas en verano</video:title>
      <video:description>Alkis te muestra cómo conseguir que los filetes queden perfectamente hechos siempre.
    </video:description>
      <video:content_loc>http://www.example.com/video123.flv</video:content_loc>
      <video:player_loc allow_embed="yes" autoplay="ap=1">
        http://www.example.com/vidoplayer.swf?video=123</video:player_loc>
      <video:duration>600</video:duration>
      <video:expiration_date>2009-11-05T19:20:30+08:00</video:expiration_date>
      <video:rating>4.2</video:rating>
      <video:view_count>12345</video:view_count>
      <video:publication_date>2007-11-05T19:20:30+08:00</video:publication_date>
      <video:family_friendly>yes</video:family_friendly>
      <video:restriction relationship="allow">IE GB US CA</video:restriction>
      <video:gallery_loc title="Cooking Videos">http://cocina.example.com</video:gallery_loc>
      <video:price currency="EUR">1,99</video:price>
      <video:requires_subscription>yes</video:requires_subscription>
      <video:uploader info="http://www.example.com/users/grillymcgrillerson">JuanFernández</video:uploader>
      <video:live>no</video:live>
    </video:video>
  </url> </urlset>
```

- <https://support.google.com/webmasters/answer/80472?hl=es>



# Archivo Robots.txt

---

- Permite indicar a los “robots” que contenido no quieres que se indexe
- Establecer tiempo mínimo entre accesos
- Útil para minimizar carga del servidor
  - Evitar indexar páginas de búsqueda, de soporte (estadísticas de uso), costosas de procesar...
- Problema: no obliga a nada, el robot se lo puede saltar
- **Importante!** Situado en la raíz del dominio:
  - <http://dominio/robots.txt>



# Archivo Robots.txt (ejemplo)

---

*User-agent: \**

*# Disable access to Discovery search and filters*

*Disallow: /discover*

*Disallow: /search-filter*

*# This should be the FULL URL to your HTML Sitemap.*

*# Make sure to replace "[dspace.url]" with the value of your 'dspace.url' setting in your dspace.cfg file.*

*Sitemap: http://[dspace.url]/htmlmap*

*# If you have configured DSpace (Solr-based) Statistics to be publicly accessible,*

*# then you likely do not want this content to be indexed*

*# Disallow: /statistics*



# Archivo Robots.txt (ejemplo)

---

*# Uncomment the following line ONLY if sitemaps.org or HTML sitemaps are used  
# and you have verified that your site is being indexed correctly.*

*# Disallow: /browse*

*# You also may wish to disallow access to the following paths, in order  
# to stop web spiders from accessing user-based content:*

*# Disallow: /advanced-search*

*# Disallow: /contact*

*# Disallow: /feedback*

*# Disallow: /forgot*

*# Disallow: /login*

*# Disallow: /register*

*# Disallow: /search*

*Crawl-delay: 1*



**CRUE**

**REBIUN**

Red de Bibliotecas Universitarias



**UNIVERSITAT POLITÈCNICA DE CATALUNYA**  
**BARCELONATECH**

Servei de Biblioteques, Publicacions i Arxius



# Archivo Robots.txt

**Add more default blocks for certain spiders in robots.txt**

<https://jira.duraspace.org/browse/DS-2335>

*Aplicado DSpace 5.0*

Ejemplo:

<http://en.wikipedia.org/robots.txt>

<http://raw.githubusercontent.com/DSpace/DSpace/master/dspace-jspui/src/main/webapp/robots.txt>



**CRUE**

**REBIUN**

Red de Bibliotecas Universitarias



**UNIVERSITAT POLITÈCNICA DE CATALUNYA**  
**BARCELONATECH**

Servei de Biblioteques, Publicacions i Arxius

# Archivo Robots.txt

- ¿Bloquear via robots.txt archivos txt autogenerados?

Reports that Google Scholar is sometimes linking to DSpace extracted text (\*.pdf.txt) files instead of original PDF 

<https://jira.duraspace.org/browse/DS-1387>

*Sin resolver*



CRUE

REBIUN

Red de Bibliotecas Universitarias



UNIVERSITAT POLITÈCNICA DE CATALUNYA  
BARCELONATECH

Servei de Biblioteques, Publicacions i Arxius

# Últimos registros añadidos

- Facilitar desde la página principal del repositorio un enlace a un listado de los últimos registros añadidos
- Facilita el rastreo de los registros nuevos

**Add a way for harvesters to find recently added items (request from Google) **

<https://jira.duraspace.org/browse/DS-1482>

*Fixed DSpace 4.0*



**CRUE**

**REBIUN**

Red de Bibliotecas Universitarias



**UNIVERSITAT POLITÈCNICA DE CATALUNYA  
BARCELONATECH**

Servei de Biblioteques, Publicacions i Arxius

# Redirecciones de páginas de descarga

---

- Se desaconseja re-direccionar la página de descarga a alguna página intermedia.
- Posible problemas de ser marcado como “hackeo” de contenido (cloaking) → Daña posicionamiento en índices de buscadores
- Ejemplos pasados en DSpace:
  - Capturar accesos en Google analytics  
<http://comments.gmane.org/gmane.comp.db.dspace.user/27728>



CRUE

REBIUN

Red de Bibliotecas Universitarias



UNIVERSITAT POLITÈCNICA DE CATALUNYA  
BARCELONATECH

Servei de Biblioteques, Publicacions i Arxius

# ¿Se utiliza el protocolo OAI?

---

- Acceso a los últimos registros añadidos o modificados
- Metadatos descriptivos de los recursos a indexar



**CRUE**

**REBIUN**

Red de Bibliotecas Universitarias



**UNIVERSITAT POLITÈCNICA DE CATALUNYA**  
**BARCELONATECH**

Servei de Biblioteques, Publicacions i Arxius

# ¿Se utiliza el protocolo OAI?

- **Generalmente NO.**
- No hay forma fiable de determinar la dirección base del servidor OAI-PMH de un repositorio
- No hay un estándar o forma predecible de llegar a la pantalla de descripción del registro/archivo, haciendo la indexación y presentación de resultados difícil
- Normalmente solo se ofrece acceso a metadatos en formato Dublin Core simple, un subconjunto de los metadatos disponibles
- NOTA: En 2008, Google oficialmente anunció que [retira el soporte para OAI-PMH sitemaps](#)



# Sobre posicionamiento

---

- Gran parte parte basado en ser enlazados
- Mejorar la indexación
- ¿Uso del Handle?
- Páginas indexadas? (robots.txt)
- Adaptación a móviles:
  - <http://www.whatsnew.com/2015/02/26/google-quiere-seguir-mejorando-la-calidad-de-los-resultados-de-las-busquedas-moviles/>
  - <http://googlewebmastercentral.blogspot.com.es/2015/02/finding-more-mobile-friendly-search.html>



- Search Engine Optimization

<https://wiki.duraspace.org/display/DSDOC5x/Search+Engine+Optimization>

- Jira DSpace

<https://jira.duraspace.org/browse/DS>





- Funcionalidades
- Guía de inclusión
- Formularios: petición recolección y contacto
- Cambios de sistema / enlaces y redirecciones



# Funcionalidades

---

- Agrupación de resultados
- Ordenación por relevancia / fecha
- Extracción de citas
- Generar citaciones
- Perfil de autores, bibliometria, etc.
- No API ?



**CRUE**

**REBIUN**

Red de Bibliotecas Universitarias



**UNIVERSITAT POLITÈCNICA DE CATALUNYA**  
**BARCELONATECH**

Servei de Biblioteques, Publicacions i Arxius

# Guía de inclusión

---

- Recomendaciones para diferentes entes: autores individuales, universidades, revistas
- Contenido:
  - Principalmente documentos académicos: journal papers, conference papers, technical reports, or their drafts, dissertations, pre-prints, post-prints, or abstracts
  - NO: magazine articles, book reviews and editorials
- Documentos de más de 5MB (libros / grandes tesis) subidos a Google Books
- Acceso al texto completo o abstract visible (escrito por el autor)
- Evitar otros bloqueos: autenticación, anuncios, ...



CRUE

REBIUN

Red de Bibliotecas Universitarias



UNIVERSITAT POLITÈCNICA DE CATALUNYA  
BARCELONATECH

Servei de Biblioteques, Publicacions i Arxius

# Guía de inclusión II

---

- Formatos de archivo: html o pdf (buscable) (?)
- Navegación:
  - Como mucho 10 enlaces html desde la página principal
  - Preferiblemente listar por fecha de publicación
  - Listar los añadidos en las 2 últimas semanas (sites grandes)
- Comportamiento frente códigos HTTP: 5xx, 4xx, 301
- No bloquear en Robots.txt



CRUE

REBIUN

Red de Bibliotecas Universitarias



UNIVERSITAT POLITÈCNICA DE CATALUNYA  
BARCELONATECH

Servei de Biblioteques, Publicacions i Arxius

# Guía de inclusión: indexación

---

- Cada artículo o abstract en su pdf o página html
- Metatags (Dublin Core como última alternativa)
  - Obligatorio Título, Autor (mínimo uno), Fecha
  - Autores:
    - Ambos formatos
    - Excluir afiliación, titulación, etc.
  - En general, incluir la información que incluirías en una citación
- Recomendaciones específicas sobre como organizar la información sin uso de metatags
- Enlazar todas las versiones del texto completo (y en el mismo subdirectorío del abstract)
- Marcar la sección de referencias con una cabecera “References” / “Bibliography” y listarlas / numerarlas (“<ol>”)



CRUE

REBIUN

Red de Bibliotecas Universitarias



UNIVERSITAT POLITÈCNICA DE CATALUNYA  
BARCELONATECH

Servei de Biblioteques, Publicacions i Arxius

# Guía de inclusión III

---

- Tiempos de incorporación:
  - Nuevas: varias veces por semana
  - Actualizaciones: 6-9 meses
- Búsqueda “site:dominio” no representa el número total de registros indexados
  - Solo busca la versión principal del documento
  - Cálculo estimado buscando sobre una porción del índice



CRUE

REBIUN

Red de Bibliotecas Universitarias



UNIVERSITAT POLITÈCNICA DE CATALUNYA  
BARCELONATECH

Servei de Biblioteques, Publicacions i Arxius

# Guía de inclusión: formularios

---

- Solicitar recolección :
  - <http://www.google.com/support/scholar/bin/request.py>
- Contacto:
  - <https://support.google.com/scholar/contact/general?hl=en>



**CRUE**

**REBIUN**

Red de Bibliotecas Universitarias



**UNIVERSITAT POLITÈCNICA DE CATALUNYA**  
**BARCELONATECH**

Servei de Biblioteques, Publicacions i Arxius

# Cambios de sistema / enlaces

---

- Evitar errores 404, o redirecciones a página principal → muchos errores pueden borrar el contenido del índice
- Redireccionar todas las páginas indexadas retornando HTTP 301 (moved permanently) a la nueva localización (no página intermedia)
- Mantener un tiempo (al menos 12 meses) hasta actualización índice.
- Pruebas en desarrollo:
  - Contacto con ellos un mes antes
  - Posibilidad de probar recolección en entorno de pruebas



CRUE

REBIUN

Red de Bibliotecas Universitarias



UNIVERSITAT POLITÈCNICA DE CATALUNYA  
BARCELONATECH

Servei de Biblioteques, Publicacions i Arxius



- Inclusion Guidelines - Google Scholar  
<http://scholar.google.com/intl/es/scholar/inclusion.html#overview>
- Videos
  - To Disappear, or Not to Disappear: How to Avoid Dropping Out of Search - Darcy Dapra  
<https://www.youtube.com/watch?v=mP5DuqqBMu0>
  - To Disappear, or Not to Disappear: How to Avoid Dropping Out of Search - Questions  
<https://www.youtube.com/watch?v=2MtME9mSfqs>

# Metatags html: Sumario

---

- Descripción
- Highwire Press
- Autor y fechas: cambios recientes en DSpace
- Referencia a página descriptiva y de descarga
- Datos más concretos de la citación: UPC
- Otros usos



CRUE

REBIUN

Red de Bibliotecas Universitarias



UNIVERSITAT POLITÈCNICA DE CATALUNYA  
BARCELONATECH

Servei de Biblioteques, Publicacions i Arxius

# Descripción Metatags

---

- Etiquetas html en el encabezado de las páginas web
- Incluyen metadatos de referencia sobre la página
- Aportan información estructura y útil a los buscadores
  
- Dublin Core como estándar
- En el ámbito de los repositorios / publicaciones se utilizan otros más específicos:
  - Highwire Press
  - PRISM
  - Eprints
  - BE Press
  - ...?



- Plataforma de publicación compañía HighWire Press
- Opción destacada en los ejemplos de la documentación de Google Scholar
- Implementado en DSpace
  
- Metadatos
  - Nombre con prefijo *citation\_*
  - Tabla completa?
  - Metadatos específicos por tipo de documento

# Highwire Press: tabla I

HP Metadata	DC
citation_author	DC.creator
citation_date / citation_publication_date	DC.issued
citation_title	DC.title
citation_publisher	DC.publisher
citation_keywords	DC.subject
citation_language	DC.language



CRUE

REBIUN

Red de Bibliotecas Universitarias



UNIVERSITAT POLITÈCNICA DE CATALUNYA  
BARCELONATECH

Servei de Biblioteques, Publicacions i Arxius

# Highwire Press: tabla II

HP Metadata	DC
citation_doi	"DC.identifier.doi"
citation_pmids	"DC.identifier.pmid"
citation_issn	"DC.identifier.issn"
citation_isbn	"DC.identifier.isbn"

HP Metadata	DC
citation_volume	"DC.citation.volume"
citation_issue	"DC.citation.issue"
citation_firstpage	"DC.citation.spage"
citation_lastpage	"DC.citation.epage"



# Highwire Press: tabla III

HP Metadata	Tipo	DC
citation_conference_title	Congresos	DC.relation.ispartof
citation_journal_title	Revistas	DC.relation.ispartof
citation_inbook_title	Capítulo de libro	DC.relation.ispartof
citation_technical_report_number	Reports	
citation_technical_report_institution	Reports	DC.publisher
citation_dissertation_name	Tesis	DC.title
citation_dissertation_institution	Tesis	
citation_patent_country	Patente	
citation_patent_number	Patente	



# Highwire Press: enlace archivo y abstract

HP Metadata	Descripción	DC
citation_abstract_html_url	Enlace página descriptiva (metadatos)	
citation_fulltext_html_url	Enlace texto completo en html	
citation_pdf_url	Enlace texto completo	DC.identifier

- citation\_pdf\_url
  - ¿Qué sucede si hay más de un archivo?
  - no solo PDF:

Store link to "primary bitstream" in citation\_pdf\_url for Google Scholar (request from Google) 

<https://jira.duraspace.org/browse/DS-1483>

*Aplicado DSpace 4.0*



CRUE

REBIUN

Red de Bibliotecas Universitarias



UNIVERSITAT POLITÈCNICA DE CATALUNYA  
BARCELONATECH

Servei de Biblioteques, Publicacions i Arxius



# HighWire Press: configuración DSpace

- En DSpace desde la versión 1.7

Provide metatags used by Google Scholar for enhanced indexing 

<https://jira.duraspace.org/browse/DS-396>

*Aplicado DSpace 1.7*



CRUE

REBIUN

Red de Bibliotecas Universitarias



UNIVERSITAT POLITÈCNICA DE CATALUNYA  
BARCELONATECH

Servei de Biblioteques, Publicacions i Arxius

# HighWire Press: configuración DSpace

- Activar en dspace.cfg

```
google-metadata.enable = true
```

- Mapeo de metadatos a través de un archivo de configuración:

```
[dspace]/config/crosswalks/google-metadata.properties
```

- Ejemplo:

<https://github.com/DSpace/DSpace/blob/master/dspace/config/crosswalks/google-metadata.properties>

```
google.citation_title = dc.title
```

```
google.citation_publisher = dc.publisher
```

```
google.citation_author = dc.author | dc.contributor.author | dc.creator
```



# Fechas: cambios en DSpace

- Problemas con la asignación automática de la fecha de publicación (dc.date.issued)

**"dc.date.issued" is often incorrectly set (reported from Google)**

<https://jira.duraspace.org/browse/DS-1481>

*Aplicado DSpace 4.0*

**DSpace should no longer assign "dc.date.issued=[today]" when date field is missing**

<https://jira.duraspace.org/browse/DS-1745>

*Aplicado DSpace 4.0*

**Find a way to report on existing, possibly inaccurate "dc.date.issued" values**

<https://jira.duraspace.org/browse/DS-1822>

*Sin resolver*



# Autores: cambio en DSpace

- Separación de los autores en diferentes tags

<b>Google Scholar author metadata tags incorrect</b>
<a href="https://jira.duraspace.org/browse/DS-2309">https://jira.duraspace.org/browse/DS-2309</a>
<i>Aplicado DSpace 5.0</i>

NOTA: En DSpace 5, el campo google.citation\_authors ha cambiado a google.citation\_author en el archivo de configuración



# Datos de la citación

**dc.identifier.citation**

Clavero Campos, Javier [et al.]. FUTUR: el nou portal de la producció científica de la Universitat Politècnica de Catalunya. "Item: revista de biblioteconomia i documentació", 2013, núm. 57, p. 145-156.

- Disponer de los campos propios de la citación separados (título de revista, número, volumen, página inicial, página final,.. ) permite configurar-los para mostrar en su respectivo metatag:

<b>upcommons.citation.published</b>	true
<b>upcommons.citation.publicationName</b>	Item: revista de biblioteconomia i documentació
<b>upcommons.citation.number</b>	57
<b>upcommons.citation.startingPage</b>	145
<b>upcommons.citation.endingPage</b>	156



CRUE

REBIUN

Red de Bibliotecas Universitarias



UNIVERSITAT POLITÈCNICA DE CATALUNYA  
BARCELONATECH

Servei de Biblioteques, Publicacions i Arxius

# Datos de la citación

## Tipus de document:

Article de revista  
Audiovisual  
Capítol de llibre  
Comunicació de congrés  
Concurs d'arquitectura i urbanisme  
Llibre

Seleccioneu el tipus de document. Per seleccionar més d'un valor de la llista, mantingueu premuda la tecla Control

## Camps citació:

Genera la citació

Para comunicaciones o textos en actas de congreso, introduce el nombre del congreso

Título de la publicación (revista / título de congreso / título de libro en capítulos de libro)

Item: revista de biblioteconomia i documentació

Volumen

Número

57

Pág. Ini

145

Pág. fi

156

Lugar de publicación

Edición

Otros

Marqueu si cal generar la citació i ompliu els camps necessaris per completar-la

Clavero Campos, Javier [et al.]. FUTUR: el nou portal de la producció científica de la Universitat Politècnica de Catalunya. "Item: revista de biblioteconomia i documentació", 2013, núm. 57, p. 145-156. [Actualitza/torna a crear citació](#)



CRUE

REBIUN

Red de Bibliotecas Universitarias



UNIVERSITAT POLITÈCNICA DE CATALUNYA  
BARCELONATECH

Servei de Biblioteques, Publicacions i Arxius

# Datos de la citación

```
<meta content="2013" name="citation_publication_date">
<meta content="0214-0349" name="citation_issn">
<meta content="2015-03-06" name="citation_online_date">
<meta content="http://upcommons.upc.edu/bitstream/2117/47760/1/2013%20-%20Cacho%
20Figueras%20et%20al.%20-%20FUTUR%20el%20nou%20portal%20de%20la%20produccio%20cientifica%20de%
20la%20Universitat%20Politecnica%20de%20Catalunya.pdf" name="citation_pdf_url">
  <meta content="Col·legi Oficial de Bibliotecaris-Documentalistes de Catalunya"
name="citation_publisher">
  <meta content="Clavero Campos, Javier; Martı́nez, Dı́dac; Prieto
Jiménez, Antonio Juan; Rovira, Anna; Serrano-Muñoz, Jordi" name="citation_authors">
  <meta content="FUTUR: el nou portal de la producció científica de la
Universitat Politècnica de Catalunya" name="citation_title">
  <meta content="Item: revista de biblioteconomia i documentació;"
name="citation_journal_title">
  <meta content="57" name="citation_issue">
  <meta content="145" name="citation_firstpage">
  <meta content="156" name="citation_lastpage">
  <meta content="Avaluació de la recerca; Difusió de la recerca;
Visibilitat; Accés obert; Producció científica; Currículums dels
investigadors; Universitat Politècnica de Catalunya; FUTUR; Article"
name="citation_keywords">
```



- Además de facilitar la recolección es útil para:
  - Exportar a gestores de referencias:

Mendeley: [Easy one-click addition of papers from Highwire Press, BMC, PLoS, Arxiv and more](#)

- Link resolver (OpenURL)
  - No con metatags, configuración específica



- Arlitsch, Kenning, and Patrick S. O'Brien. "Invisible institutional repositories: Addressing the low indexing ratios of IRs in Google Scholar." *Library Hi Tech* 30.1 (2012): 60-81. [https://jira.duraspace.org/secure/attachment/13020/Invisible\\_institutional.pdf](https://jira.duraspace.org/secure/attachment/13020/Invisible_institutional.pdf)



# Herramientas de Webmaster

---

- Funcionalidades
  - Estado índice: páginas indexadas, etc.
  - Estadísticas de búsquedas, % clicks
  - Enlaces (internos / externos)
  - Recomendaciones web / móvil
  - Eliminar URLs
  - Rastreo: errores / estadísticas
  - Seleccionar “mejores horas”
  - Explorar como Googlebot / Bing (enviar a índice)
  - Gestionar Sitemaps
  - Comprobar robots.txt
  - Alertas: Problemas de seguridad, errores, etc.
  - Ver datos estructurados



# Herramientas de Webmaster

---

- Añadir sitio:
  - Archivo en la raíz del dominio
  - Añadiendo Metatag
  - Código de Google analytics
  - ...
- Google: <https://www.google.com/webmasters/tools/home?hl=es>
- Bing / Yahoo: <http://www.bing.com/toolbox/webmaster>



CRUE

REBIUN

Red de Bibliotecas Universitarias



UNIVERSITAT POLITÈCNICA DE CATALUNYA  
BARCELONATECH

Servei de Biblioteques, Publicacions i Arxius

# Eliminación de URL (Google)

- Herramientas para borrar del índice y la cache:

- Panel del sitio
- Mensajes del sitio
- ▶ Aspecto de la búsqueda ⓘ
- ▶ Tráfico de búsqueda
- ▼ Índice de Google
  - Estado de indexación
  - Palabras clave de contenido
  - Eliminación de URL**

## Eliminación de URL

Utiliza **robots.txt** para especificar cómo deben rastrear tu sitio los motores de búsqueda o solicita que se **elimine tu URL** ). Solo el propietario de un sitio y los usuarios con permisos completos pueden solicitar que se elimine una UR

Crear una nueva solicitud de eliminación

Introduce la URL que quieras eliminar (distingue entre mayúsculas y minúsculas).

**URL:**

**Motivo:**

El propietario del sitio ha eliminado la página de los motores de búsqueda o la ha bloqueado.



CRUE

REBIUN

Red de Bibliotecas Universitarias



UNIVERSITAT POLITÈCNICA DE CATALUNYA  
BARCELONATECH

Servei de Biblioteques, Publicacions i Arxius

# Actualizar URL (Google)

## Explorar como Google

http://upcommons.upc.edu/

Escritorio

OBTENER

OBTENER Y PROCESAR

Deja la URL en blanco para recuperar la página principal. Las solicitudes pueden tardar unos minutos en procesarse.

Ruta	Tipo de robot de Google	Procesamiento solicitado	Estado	Fecha:
<a href="#">/handle/2117/13854</a>	Escritorio		✓ Completo	8/3/15 14:35

Enviar al índice



CRUE

REBIUN

Red de Bibliotecas Universitarias



UNIVERSITAT POLITÈCNICA DE CATALUNYA  
BARCELONATECH

Servei de Biblioteques, Publicacions i Arxius

# Eliminación/actualización de URL

- <https://www.bing.com/webmaster/tools/content-removal>

## Eliminación de contenido

Utiliza la herramienta de Eliminación de contenido para informar a Bing sobre una página web que ya fue eliminada por el administrador web pero que aún se muestra en los resultados de búsqueda, o sobre una página en caché con contenido desactualizado. Nota: si eres el administrador web, utiliza mejor la [herramienta de URL bloqueadas](#). [Más información](#)

### URL de contenido

### Tipo de eliminación

## Historial de envíos

FECHA	URL	TIPO DE ELIMINACIÓN	ESTADO	RESPUESTA HTTP
-------	-----	---------------------	--------	----------------



CRUE

REBIUN

Red de Bibliotecas Universitarias



UNIVERSITAT POLITÈCNICA DE CATALUNYA  
BARCELONATECH

Servei de Biblioteques, Publicacions i Arxius

# Herramientas analíticas

- Google analytics
  - Procedencia tráfico
  - Palabras de búsqueda (comportamiento)
- Capturar descarga de archivos:

**Record bitstream downloads as Google Analytics events** 

<https://jira.duraspace.org/browse/DS-2088>

*Aplicado DSpace 5.0*

**Add an XMLUI aspect to report Google Analytics stats** 

<https://jira.duraspace.org/browse/DS-2108>

*Aplicado DSpace 5.0*



**CRUE**

**REBIUN**

Red de Bibliotecas Universitarias



**UNIVERSITAT POLITÈCNICA DE CATALUNYA  
BARCELONATECH**

Servei de Biblioteques, Publicacions i Arxius

# “Futuro?” : Datos estructurados

---

- Datos estructurados
- Iniciativa schema.org
- Ejemplos generales
- Ejemplos en “datos bibliográficos”



CRUE

REBIUN

Red de Bibliotecas Universitarias



UNIVERSITAT POLITÈCNICA DE CATALUNYA  
BARCELONATECH

Servei de Biblioteques, Publicacions i Arxius



# Datos estructurados

---

- Introducir información semántica en el contenido de las páginas web
- Información más comprensible para “robots”
- Diferentes formas:
  - Microformatos
  - RDFa
  - Microdata (HTML5)
  - JSON-LD (JSON for Linking Data)



- Sencillas convenciones (conocidas como **entidades**)
- Describen un tipo concreto de información:
  - una opinión, un evento, un producto, una empresa o una persona). Cada entidad tiene sus propias **propiedades**.
- Utilizando tags y atributos html existentes: “class”, “rel”, “rev”
- Ejemplos: hAtom, hCalendar, hCard (adre, geo), hReview, ...

# Microformatos (ejemplos bibliográficos)

- hCite

```
<span class="hcite">  
  <span class="creator vcard"><span class="fn">Apellido, Nombre</span></span>,  
  <span class="title">Título de la publicación.</span>  
  In <span class="container hcite">  
    <abbr class="type" title="Journal">J.</abbr><abbr class="title" title="Aerospace medicine">Aersp.  
    Med.</abbr>  
    <span class="volume">45</span>  
    <span class="issue">10</span>  
    <abbr class="date-published" title="101974">Oct, 1974</abbr>  
  </span>, pages <span class="page">1115-36</span>.  
</span>
```

- COinS, <http://ocoins.info/>:

```
<span class="Z3988" title="ctx_ver=Z39.88-  
2004&amp;rft_val_fmt=info%3Aofi%2Ffmt%3Akev%3Amtx%3Adc&amp;rft_id=http%3A%2F%  
2Fhdl.handle.net%2F2099.4%2F1570&amp;rft_id=b12346354&amp;rfr_id=info%3Asid%2Fdsp  
ace.org%3Arepository&amp;rft.creator=Violette%2C+H.&amp;rft.date=2014-12-  
16T12%3A13%3A19Z&amp;rft.date=2014-12-16T12%3A13%3A19Z....>Contenido ...</span>
```



- Extensiones del XHTML para introducir información semántica propuestas por W3C
- Generalización de los atributos de las etiquetas meta y link de HTML:
  - Typeof, about, rel, rev, href, resource, property, content, datatype
- RDFa Lite 1.1 (simplificación)

```
<p xmlns:dc="http://purl.org/dc/elements/1.1/"  
  about="http://www.example.com/books/wikinomics">  
  In his latest book <em property="dc:title">Wikinomics</em>,  
  <span property="dc:creator">Don Tapscott</span> explains deep changes in  
  technology, demographics and business. The book is due to be published in <span  
  property="dc:date" content="2006-10-01">October 2006</span>.  
</p>
```

# Microdata (HTML5)

---

- Semántica en HTML5: header, nav, article, section, footer, ...
- Microdata permite incluir más información semántica al contenido en HTML5:
- Atributos:
  - Itemscope: sección donde se anida la información
  - Itemtype: tipo
  - Itemid: identificador
  - Itemprop: propiedad
  - Itemref: referencia



- Provee de una colección de esquemas/vocabularios para marcar HTML y hacerlo comprensible a la mayoría de motores de búsqueda
- Colaboración entre Google, Microsoft, y Yahoo! (+ Yandex)
- Sobre tres estándares:
  - Microdatos (preferido)
  - RDFa
  - JSON-LD
- Beneficios:
  - Crear **fragmentos enriquecidos**
  - Otros futuras...

# Schema.org (ejemplo)

```
<div itemscope itemtype = "http://schema.org/Movie">  
  <h1 itemprop="name">Avatar</h1>  
  <div itemprop="director" itemscope itemtype="http://schema.org/Person">  
    Director: <span itemprop="name">James Cameron</span> (born <span  
    itemprop="birthDate">August 16, 1954)</span>  
  </div>  
  <span itemprop="genre">Science fiction</span>  
  <a href=" ../movies/avatar-theatrical-trailer.html" itemprop="trailer">Trailer</a>  
</div>
```



CRUE

REBIUN

Red de Bibliotecas Universitarias



UNIVERSITAT POLITÈCNICA DE CATALUNYA  
BARCELONATECH

Servei de Biblioteques, Publicacions i Arxius

- Todo de tipo Thing (4 atributos: name, description, url y image)
- Tipos más específicas (sus atributos y los de las clases más genéricas)
- Las más comunes:
  - Creative works: [CreativeWork](#), [Book](#), [Movie](#), [MusicRecording](#), [Recipe](#), [TVSeries](#) ...
  - Embedded non-text objects: [AudioObject](#), [ImageObject](#), [VideoObject](#)
  - [Event](#)
  - [Organization](#)
  - [Person](#)
  - [Place](#), [LocalBusiness](#), [Restaurant](#) ...
  - [Product](#), [Offer](#), [AggregateOffer](#)
  - [Review](#), [AggregateRating](#)



- **Más es mejor**, pero marcar solo contenido visible siempre que sea posible
- Ciertos atributos se pueden definir como solo texto o un objeto completo (ya sea del definido en el esquema o alguno de sus “descendientes”).
- Utilizar la propiedad *url* para dirigir a página con más información sobre el objeto definido

- Etiquetas/atributos específicos para definir algunos valores de forma más comprensible para “robots”:
  - time[datetime], link[href], meta[content]  
<time datetime="2011-05-08T19:30">May 8, 7:30pm</time>
- Enumerations: vocabularios controlados para ciertos valores (Ej.: InStock)
- Utilización de link (href) para enlaces no visibles
- Utilización de la etiqueta meta (content) para introducir información de contenido no marcable
- Mecanismos para extender el vocabulario
- Más info: <http://schema.org>

# Schema.org: Recursos

---

- URL validadores:
  - Google: <https://developers.google.com/structured-data/testing-tool/>
  - Bing: <http://www.bing.com/toolbox/markup-validator>
  - Yandex: <https://webmaster.yandex.com/microtest.xml>
  - Structured Data Linter: <http://linter.structured-data.org/>
- Asistente para el marcado:
  - <https://www.google.com/webmasters/markup-helper/u/0/>



# Schema.org: fragmentos enriquecidos

## Entradas AC/DC | AC/DC - Entradas para Conciertos y ...

[www.viagogo.es](http://www.viagogo.es) > Entradas Conciertos > Hard Rock/Metal ▾

Encuentra las entradas de **AC/DC** que buscas en viagogo, el sitio de venta de entradas más grande el mundo. Selecciona un evento para ver los tipos de ...

10 de abr. - 12 de abr. [AC/DC, Jack White ...](#) Empire Polo Field, Indio ...

17 de abr. - 19 de abr. [AC/DC, Jack White ...](#) Empire Polo Field, Indio ...

mar., 5 de may. [AC/DC](#) Gelredome, Arnhem, Holanda

## Los pilares de la tierra, 1 - Goodreads

[www.goodreads.com/book/show/12994251-los-pilares-de-la-tierra-1](http://www.goodreads.com/book/show/12994251-los-pilares-de-la-tierra-1) ▾

★★★★★ Valoración: 4,2 - 648 votos

Los **pilares de la tierra, 1** has 648 ratings and 27 reviews. Aranzazu said: No le pongo 10 estrellas porque no las hay! es un libro precioso donde encuent...

## Flamenquines caseros cordobeses - Recetas de rechupete



[www.recetasderechupete.com/flamenquines...paso.../9505/](http://www.recetasderechupete.com/flamenquines...paso.../9505/) ▾

25 min

Cómo preparar unos deliciosos **flamenquines** de filetes de cerdo, rellenos de jamón y con un toque crujiente de rebozado. Los más famosos de Córdoba.

<https://developers.google.com/structured-data/testing-tool/>



CRUE

REBIUN

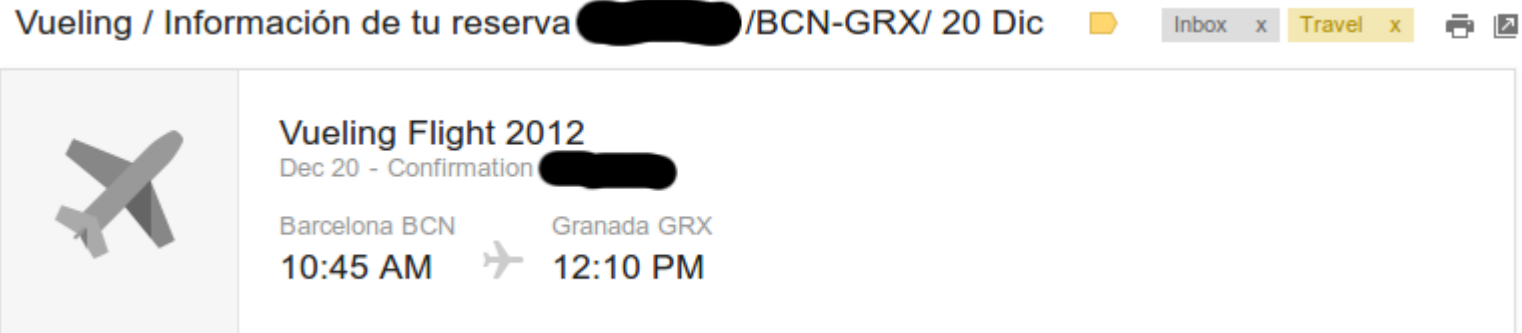
Red de Bibliotecas Universitarias



UNIVERSITAT POLITÈCNICA DE CATALUNYA  
BARCELONATECH

Servei de Biblioteques, Publicacions i Arxius

# Schema.org: etiquetado de correos



<https://developers.google.com/gmail/markup/>

- En proceso de estandarización:
- Importación directa a calendarios, Google Now, ...



CRUE

REBIUN

Red de Bibliotecas Universitarias



UNIVERSITAT POLITÈCNICA DE CATALUNYA  
BARCELONATECH

Servei de Biblioteques, Publicacions i Arxius

# Schema.org: otros usos

- Restringir un Custom search engine:

## Restringir páginas mediante tipos de esquema de schema.org

Restringir las páginas de la lista de sitios anterior a solo aquellas que contengan los tipos de esquema de Schema.org de la lista siguiente.

Puedes añadir hasta diez (10) tipos de schema.org a tu motor de búsqueda. Ten en cuenta que, cuando se añade un nodo, se incluyen automáticamente todos los elementos secundarios para que no los tengas que volver a añadir. Por ejemplo, si añades CreativeWork, no es necesario añadir Book, ImageObject, VideoObject, etc. por separado.

<https://support.google.com/customsearch/answer/4544182?hl=en>



CRUE

REBIUN

Red de Bibliotecas Universitarias



UNIVERSITAT POLITÈCNICA DE CATALUNYA  
BARCELONATECH

Servei de Biblioteques, Publicacions i Arxius

# Schema.org: Datos bibliográficos

---

- Vocabulario definido:
  
- <http://schema.org/Thing>
  - <http://schema.org/CreativeWork>
    - <http://schema.org/Article>
      - <http://schema.org/ScholarlyArticle>
    - <http://schema.org/Book>
    - <http://schema.org/Dataset>
    - ...



# Recursos en datos bibliográficos

---

- Schema bib extend community group:
  - <https://www.w3.org/community/schemabibex/>
  - [http://www.w3.org/community/schemabibex/wiki/Main\\_Page](http://www.w3.org/community/schemabibex/wiki/Main_Page)
    - Casos de uso, vocabularios propuestos, tipos de objeto, ...
- Ejemplos de implementaciones:
  - GoodReads
    - <http://www.goodreads.com/>
  - WorldCat (Explore WorldCat Linked Data)
    - <http://www.oclc.org/developer/develop/linked-data/linked-data-exploration.en.html>





# Ejemplo en Fondo antiguo UPC

<http://fonsantic.upc.edu/handle/2099.4/167>

Català Castellano English

## Fondo Antiguo de la UPC

*Patrimonio bibliográfico histórico de las Bibliotecas de la UPC*

Imprimir Twitter Facebook

Portal de Fondo antiguo de la UPC > Biblioteca del Campus del Baix Llobregat > Fons antic d'Agricultura > Ver ítem>

### A l'entorn de la recerca econòmico-agrícola

[Mostrar el registro completo del ítem](#)

<b>Título:</b>	A l'entorn de la recerca econòmico-agrícola
<b>Autor:</b>	Llovet Mont-Ros, Josep
<b>Fecha:</b>	1936
<b>Descripción:</b>	Separata de: Arxius d'Escola Superior d'Agricultura. Vol. II, fasc. IV
<b>Tema:</b>	Agricultura -- Aspectes econòmics
<b>URI:</b>	<a href="http://hdl.handle.net/2099.4/167">http://hdl.handle.net/2099.4/167</a>
<b>Enlace al catálogo:</b>	<a href="http://cataleg.upc.edu/record=b1366093">http://cataleg.upc.edu/record=b1366093</a>

### Ficheros en el ítem

Ficheros	Tamaño	Formato	Vista
<a href="#">b13660937.pdf</a>	11.74Mb	PDF	



CRUE

REBIUN

Red de Bibliotecas Universitarias



UNIVERSITAT POLITÈCNICA DE CATALUNYA  
BARCELONATECH

Servei de Biblioteques, Publicacions i Arxius

# Ejemplo geolocalización

- <http://fonsantic.upc.edu/handle/2099.4/48>

```
<span itemtype="http://schema.org/GeoCoordinates"
  itemscope="itemscope">
  <meta itemprop="latitude" content="41.9021667" />
  <meta itemprop="longitude" content="12.4539367" />
</span>
```



CRUE

REBIUN

Red de Bibliotecas Universitarias



UNIVERSITAT POLITÈCNICA DE CATALUNYA  
BARCELONATECH

Servei de Biblioteques, Publicacions i Arxius

- Ronallo, Jason. "HTML5 Microdata and Schema.org." Code4Lib Journal (2012).  
<http://journal.code4lib.org/articles/6400>
- Pilgrim, Mark. Dive Into HTML5  
<http://diveintohtml5.info/>  
<http://diveintohtml5.info/extensibility.html>



# ¡Gracias!

antonio.juan.prieto@upc.edu



**CRUE**

**REBIUN**

Red de Bibliotecas Universitarias



**UNIVERSITAT POLITÈCNICA DE CATALUNYA**  
**BARCELONATECH**

Servei de Biblioteques, Publicacions i Arxius