B. ANÁLISIS Y RECUPERACIÓN DE INFORMACIÓN

B.1. Dublin core, metadatos y vocabularios

Por **Eva Méndez**

Méndez, Eva. "Dublin core, metadatos y vocabularios". En: *Anuario ThinkEPI*, 2007, pp. 61-64.

"DC es un buen estándar porque es simple, extensible e interoperable. El ser norma ISO ha facilitado su uso en contextos corporativos y en algunos ECMS (enterprise content management system)"

"Ahora hablan de 'vocabularios' todas las comunidades que desarrollan sus sistemas y servicios de información para la Web"

EN EL BALANCE sobre el vertiginoso mundo de la WWW encontramos un conjunto de estándares, un montón de buenas prácticas, experiencias y la credibilidad explícita de que el futuro de la Web cuenta con los bibliotecarios y profesionales de la información, y los metadatos han tenido y tienen mucho que ver con esto.

Cualquiera de los lectores de EPI sabe qué son los metadatos: "datos sobre los datos", información estructurada y descriptiva sobre los recursos electrónicos para mejorar, entre otras cosas, la recuperación de información.

En sus orígenes, un término atractivo y de moda, pero torpe: metadatos, al que acompaña una definición igualmente poco acertada y sibilina[1]: "datos sobre los datos" que no nos llevaba a más que a dudar incluso de nuestra sempiterna catalogación. Sin embargo, con el paso del tiempo, tras ríos de tinta y miles de bytes de experiencia en la descripción y recuperación de objetos de información digital de índole diversa, la solidez de la tendencia de los metadatos y del propio Dublin core ha quedado demostrada y nadie pone en duda su utilidad y necesidad.

Ahora hablamos de calidad de metadatos y cómo crear metadatos que se puedan compartir, pero su uso es indiscutible en todos los proyectos de descripción, preservación y acceso a la información digital.

El Dublin core, como casi todo en la WWW, tiene el origen en dos circunstancias:
–Por un lado, en un hecho real, una necesidad y una coyuntura informativa. En este caso, la imposibilidad de catalogar la Web a través del formato MARC, que habían evidenciado, ya en 1995, proyectos como *InterCat* de OCLC[2] o *Catriona* en el Reino Unido[3].
–Por otro, un cúmulo de casualidades y buenas intenciones de un grupo humano que con el tiempo se ha convertido en anécdota. En el caso del Dublin Core, la anécdota comienza en los pasillos del *2º Congreso de la Web* en Chicago[4] donde se identificó la necesidad de un núcleo básico de metadatos para ayudar la recuperación de contenidos en la Web. Ese grupo de personas a las que se unieron unos pocos más (hasta 52), que se han definido de forma jocosa como "*geeks, freaks, and people with sensible shoes* (tipos raros, frikis y gente con zapatos cómodos), se reunirían en marzo de 1995 en Dublin (Ohio) y crearían un modelo básico para la

descripción de recursos electrónicos que comenzaría a llamarse Dublin core.

Desde 1995, el Dublin core (DC) en particular, pero también la tendencia teórica de la metainformación como método para mejorar la recuperación en entornos Web, se ha hecho indiscutible. Si bien no ha llegado a ser (todavía) el mecanismo por excelencia para la recuperación de información en sistemas de búsqueda all-the-web, vinculado a la codificación html de metaetiquetas, como se vaticinaba en el año 97-98 en las mejores épocas de Altavista, sí se ha asociado a sistemas finitos de recuperación semántica, haciendo un dupla excepcional con RDF (*resource description framework*).

A lo largo de esta década "prodigiosa" para la descripción y recuperación de recursos de información digital, DC evolucionaría dando pasos certeros hacia una estandarización formal e internacional:

–En 1998 se crea la primera versión del *Dcmes* (*Dublin core metadata element set*) versión 1.0 y la correspondiente *Request for comments* (*RFC2413*[5]).

–En 1999 se publica la segunda versión del Conjunto de elementos DC (*Dublin core element set*, versión 1.1).

–En el año 2000 se convierte en una recomendación europea, a través del *CWA13874* (*CEN workshop agreement*).

–En 2001 es un estándar americano *ANSI-NISO Z39.85*.

–Y por fin en el año 2003 se convierte en la norma *ISO 15836*, revestido no sólo de una formalización como norma apta para la industria, sino con la solidez que otorga la interdisciplinariedad y la internacionalización a cualquier proyecto en y para la Web, y se ha convertido en poco tiempo en una de las infraestructuras operacionales para la Web semántica.

–En la actualidad se están estableciendo estándares nacionales de la norma *ISO*, aunque de momento sólo Finlandia y España (a través del *CT50* de *Aenor*) están realizando esta tarea de traducción y adaptación donde se define la semántica de los 15 elementos de DC.

El éxito de Dublin core y de la utilización y adopción de sus elementos se debe, a mi juicio, a varias razones (seguro que hay más):

–Define una semántica precisa pero es sintáctico-independiente. Con esto me refiero a que no depende de una sintaxis de codificación particular, ni HTML, ni XML, ni RDF, sino de todas ellas[6].

–Se ha adoptado internacionalmente y sus elementos y semántica asociada están traducidos a más de 20 idiomas[7].

–Es un estándar de propósito general, no depende de ningún dominio informativo, pero se adapta a las distintas comunidades de información Web y se lleva muy bien con otros esquemas de metadatos de propósito específico, sirviendo de "piedra Rosetta" para la representación de las relaciones entre elementos (*crosswalk*).

–Es el modelo de metadatos clave en sistemas y servicios de información digital, como por ejemplo para la iniciativa de archivos abiertos (*OAI*[8]), o servicios comerciales como *Connexion* de *OCLC*[9]. También ha sido adoptado por distintos gobiernos en sus proyectos de e-Gov, por ejemplo en: Australia, Canadá, Dinamarca, Finlandia, Irlanda, Nueva Zelanda y Reino Unido, y por supuesto, por los más emblemáticos proyectos de bibliotecas digitales y de gestión del patrimonio digital.

–Tiene una gran validez como estándar porque es simple, extensible e interoperable. El hecho de ser una norma *ISO* lo convierte en válido para la industria y así lo demuestra su uso en contextos corporativos y el protagonismo que ha adquirido en algunos sistemas *ECMS* (*enterprise content management system*).

– Pero sobre todo creo que el éxito de Dublin core se debe a la generosidad de las personas que conforman esta comunidad, a la creencia de todas ellas en que una Web mejor es posible y al espíritu abierto, global e independiente de la *DCMI* que hace que sea una iniciativa viva que se adapta a las necesidades de distintos tipos de usuarios o distintos tipos de información, creando más términos de metadatos[10], más perfiles de aplicación, o simplemente adaptando el uso de los elementos a un fin particular.

Y en los próximos 10 años, qué

Hacer análisis sobre la realidad en pretérito perfecto (simple o compuesto, ya que

algunos de los hechos que relato en esta nota son de antes-de-ayer) es bastante fácil, sobre todo cuando se ha sido testigo de esa realidad. Sin embargo, hablar del futuro siempre produce vértigo, aunque habría que hacer también análisis prospectivos. En este sentido, siempre me gusta parafrasear a **Alan Kay** (no soy la única que lo hace, cuando se trata de aventurar el futuro tecnológico): "La mejor forma de predecir el futuro... es inventarlo". Sin embargo, en este caso no es necesario inventarlo. Dublin core es sobre todo un "estándar por seducción" y seguirá evolucionando y seduciendo a comunidades y dominios informativos en la Web y junto a él, teniendo siempre en cuenta este estándar, se crearán y evolucionarán otros vocabularios.

Vocabularios

Antes sólo hablábamos de vocabularios los bibliotecarios y profesionales de la información (y no todos, porque a algunos no les gustaba el término de **Lancaster**[11]) para referirnos a los lenguajes controlados de indización; sin embargo ahora hablan de "vocabularios" todas las comunidades que desarrollan sus sistemas y servicios de información para la Web. Vocabularios son: esquemas de metadatos de índole diversa formados por conjuntos de elementos descriptivos (aquí podría poner muchas TLAs *–three letter acronyms*, en algunos casos con alguna letra más–, que no harían más que atestarnos la cabeza); son también vocabularios las ontologías emanadas de la *Iniciativa de la web semántica* del *W3C*[12], son vocabularios, por supuesto, aunque redefinidos en lenguajes de representación formal (por ej. *SKOS*[13]) todos los sistemas de organización del conocimiento que se fundan en estructuras de espacio-valor como los tesauros y las clasificaciones; y lo son también las taxonomías o vocabularios de los ámbitos de información corporativa.

En definitiva, esquemas de elementos descriptivos o esquemas de valores temáticos orientados a la representación de materias (que los anglófonos diferencian entre *schemAs* y *schemEs* y se quedan tan anchos)..., pero en cualquier caso, metadatos con vocación de estándares, vocabularios para definir la información sobre la información para hacer útiles los datos, con una representación más semántica que permita una recuperación más inteligente. Todos estos vocabularios y estándares se desarrollarán a la par que lo hará el Dublin core pero, me consta, nunca en contra de él y compartiendo estructuras de codificación en RDF[14].

El futuro son pues más metadatos, más vocabularios (los de **Lancaster** también, pero redefinidos y codificados en *RDF/SKOS/OWL*), que quizás algún día hagan realidad prototipos de búsqueda y metadatos para la Web semántica como *Swoogle*[15], un proyecto de buscador semántico de la *Universidad de Maryland*, EUA). Quizás estoy soñando... pero recordemos que *Yahoo!* hace 10 años era también un proyecto universitario. Si queréis hablamos otro día de *Swoogle*...

Notas:

1. **Jacques Ducloy** [et al.]. *Les metadonnées et le catalogage des documents numériques*: *http://biblio-fr.info.unicaen.fr/rencontres98/minutes/metadonnees/texte.html*

2. Internet Cataloging Project: *http://digitalarchive.oclc.org/da/ViewObject.jsp?objid=0000003519*

3. *CATaloguing and Retrieval of Information Over Networks Applications* (Catriona). No existe actualmente link activo de este proyecto

4. *Second International WWW Conference '94: Mosaic and the Web*: *http://archive.ncsa.uiuc.edu/SDG/IT94/IT94Info.html*

5. Dublin Core Metadata for Resource Discovery (RFC2413): *http://www.ietf.org/rfc/rfc2413.txt*

6. Ver: DCMI Encoding Guidelines: *http://dublincore.org/resources/expressions*

7. Ver: DCMI Registry: *http://dublincore.org/dcregistry*

8. Open Archives Initiative: *http://www.openarchives.org*

9. *Connexion* es un sistema comercial que surge del proyecto de investigación de *OCLC* de finales de los 90's llamado *CORC* (*Cooperative Online Resources Catalog*), al que se le ha añadido el potencial del *WorldCat*. Ver: *http://www.oclc.org/connexion/*

10. Los *Metadata terms* son lo que antes se conocía como Dublin core cualificado o calificado a través de:

nuevos elementos (añadidos a los 15 que componen el DC-simple, como por ejemplo, *audience* o *accessibility*, matizaciones de los elementos, esquemas de codificación de los elementos, o términos de un vocabulario constituido ad hoc (como el Vocabulario Type). Ver *DCMI Metadata Terms*:
http://dublincore.org/documents/dcmi-terms

11. Recordemos el clásico de **F. W. Lancaster**. *Vocabulary control for information retrieval*. 2nd ed. Arlington, VA: Information Resources Press, 1986, editado (1993) y re-editado (2002, si no recuerdo mal) en español por la *Universidad de Valencia* con el título *El control del vocabulario en la recuperación de información*.

12. W3C, Semantic Web Activity:
http://www.w3.org/2001/sw

13. SKOS-Core: *Simple Knowledge Organization System*:
http://www.w3.org/2004/02/skos/

14. Aunque he hablado de soslayo de RDF, aún no he citado las especificaciones que componen este estándar de facto para la Web semántica y que podréis encontrar en:
http://www.w3.org/RDF

15. Swoogle: *http://swoogle.umbc.edu*

Eva Mª Méndez Rodríguez
Departamento de Biblioteconomía y Documentación. Universidad Carlos III de Madrid.
emendez@bib.uc3m.es

EU Research Scholar at Metadata Research Center (MRC).
School of Information and Library Sciences. University of North Carolina at Chapel Hill
http://ils.unc.edu/mrc/
emendez@email.unc.edu