El Thesaurus de la UB en LOD (Linked Open Data/Dades obertes vinculades). Carme Masagué (20 juny 2018)

1 El Thesaurus de la UB en LOD (Linked Open Data/Dades obertes vinculades)

Us presentem una iniciativa del CRAI de la UB on participa la Unitat de Procés Tècnic.

Es tracta de la publicació del Thesaurus de la UB en LOD (Linked Open Data/Dades obertes vinculades)

Hem adaptat l'experiència que la Rosa Fabeiro del CRAI-UB i el Pep Casals (Nubilum) varen presentar a les Quinzenes 15 Jornades Catalanes d'Informació i Documentació el mes passat (maig 2018) per il·lustrar el projecte.

2 Què és el *Thesaurus de la Universitat de Barcelona*?

És una eina que facilita la consulta per matèries perquè:

- normalitza la forma d'accedir a les matèries
- identifica sense ambigüitat el terme buscat
- agrupa els documents del catàleg a partir de les formes acceptades

- és un vocabulari controlat i estructurat amb relacions d'equivalència, jeràrquiques i associatives

- s'utilitza per a la indexació i recuperació per matèries de tots els recursos d'informació del CRAI-UB

- està redactat en català i és multilingüe amb equivalències en castellà, anglès i francès

- és multidisciplinari (inclou tots els camps d'interès de la Universitat de Barcelona)

- incorpora classificació per categories que són la base dels Microtesaurus UB.

- està desenvolupat per la Unitat de Procés Tècnic del CRAI de la UB.

3 Cronologia del Thesaurus de la UB (1)

- El 1992 vàrem publicar la 1a edició en paper del *Thesaurus de la UB* en paper que contenia 8.500 termes.
- El 1998 el vàrem publicar en línia.

4 Cronologia del Thesaurus de la UB la evolució (2)

- 2005: els termes del Thesaurus es classifiquen en 29 Microtesaurus temàtics.
- Des del 2006 : El Tesaurus de la UB s'utilitza com a únic vocabulari controlat per a la indexació dels documents s (segons l'esquema de metadades Dublin Core) en tots els repositoris digitals propis i consorciats.

- 2007: s'incorpora el Microtesaurus de Noms geogràfics.
- el 2009 es va incorporar al catàleg, la cerca per gènere i forma.

5 Cronologia del Thesaurus de la UB (3)

Com a iniciatives més interessants podem destacar que :

- 2014: Iniciativa de publicar el Thesaurus UB en dades obertes vinculades (LOD) en format SKOS amb el programari de software lliure *TemaTres.*

  El projecte es va aturar per manca de termes equivalents del tesaurus multilingüe i per que aquesta tecnologia no permetia poder gestionar les dades des del nostre Sistema Integrat de Gestió de Biblioteca (SIGB).

En aquell moment teníem 26.500 descriptors amb el 58 % d'equivalències multilingües

- 2018: el maig de 2018 amb la col·laboració de l'empresa Nubilum i amb la seva tecnologia Coeli hem transformat les dades del Tesaurus al format SKOS per a la seva publicació en dades obertes vinculades.
  La vinculació amb altres tesaurus, llistes i vocabularis especialitzats de diferents àmbits, nacionals i internacionals i en diferents llengües fa possible l'accés a aquestes bases de dades i fitxers d'autoritats de matèries per facilitar el seu ús i reutilització. Ara encara estem en període de proves per a la publicació del tesaurus en dades obertes vinculades i amb versió multilingüe. Si tot va bé, esperem publicar-lo abans d'aquest estiu.

Actualment tenim 27.506 descriptors i 80% de termes amb equivalències multilingües

- 20.900 són descriptors temàtics
- 6.390 són descriptors geogràfics
- 216 són descriptors de gènere/forma

6 El Web semàntic

Ja hem vist les iniciatives més importants en l'evolució del tesaurus. Paral·lelament el web també esta evolucionant.

L'any 2008 hi ha un salt tecnològic des de la versió 2.0 a la 3.0 també anomenada "web semàntic" o "web de les dades".

El web semàntic tracta de permetre l'accés a les dades, fent-les disponibles en formats llegibles per màquina, és a dir, metadades en format RDF, XML, JSON i connectant-les emprant URIs (Identificador Uniforme de Recursos) facilitant així a les persones i les màquines recopilar les dades, vincular-les, i enriquir-les per fer tot tipus de coses permeses per una llicència, com per exemple la (Creative Commons Public Domain Dedication) CCO que és la llicència aplicada a les dades, totalment oberta i enfocada a afavorir la seva reutilització.

7 **Què són les dades obertes vinculades? (també anomenades, LOD, o dades obertes enllaçades)**

Vincular dades és essencial per connectar el web semàntic.

Linked Open Data és un projecte desenvolupat pel Consorci del World Wide Web que es basa en la idea d'afegir al web metadades semàntiques i ontològiques, és a dir, amb informació significativa i estructurada, amb l'objectiu de crear un medi universal per a l'intercanvi d'informació.

Aquestes informacions addicionals, que són les que descriuen el contingut, el significat i les relacions de les dades, s'han de proporcionar de manera formal per tal que puguin ser avaluades automàticament per màquines de processament. Amb això es pretén ampliar la interoperabilitat dels sistemes informàtics i reduir la mediació dels operadors humans en els processos intel·ligents de flux d'informació.

Les dades obertes poden ser utilitzades, reutilitzades i redistribuïdes lliurament per qualsevol persona i estan disponibles sota una llicència oberta.

A la vostra esquerra de la diapositiva podeu veure un esquema dels 4 "Principis de les dades vinculades":

1. Emprar URIs per a identificar els recursos publicats en el web.
2. Aprofitar el HTTP de la URI per a que la gent pugui trobar i consultar aquests recursos.
3. Proporcionar informació útil sobre el recursos
4. Incloure enllaços a altres URIs relacionades amb les dades contingudes en el recurs, de manera que es potenciï el descobriment d'informació en el web.

En el centre, en color blau, teniu un resum dels "fonaments bàsics" per a vincular les dades:

1 les URI (Identificador Uniforme de Recursos) per identificar els recursos;

2 el RDF (Marc de Descripció de Recursos) format per a representar i descriure els recursos;

3 i el SPARQL que és el llenguatge normalitzat per interrogar i actualitzar dades expressades en RDF .

A la vostra dreta tenim la famosa tassa amb les 5 estrelles que ens indica, que necessitem per publicar les dades vinculades obertes.

Com a valor afegit, podem veure un vídeo d'Europeana que explica molt bé que són les dades obertes vinculades.

8 **El CRAI de la UB amb la col·laboració de l'empresa Nubilum i amb la seva tecnologia Coeli s'ha sumat al projecte de dades en accés obert amb la publicació del nostre Thesaurus UB.**

Ja hem vist que en els últims 20 anys ha hagut una important evolució tecnològica dins del web i que des del CRAI hem anat adaptant les nostres eines segons s'ha anat produint aquests canvis. Així ara presentem una nova iniciativa, la de publicar el Thesaurus en accés obert per millorar la seva accessibilitat, interoperabilitat i visibilitat. El THUB es podrà utilitzar per altres institucions i per altres usuaris perquè es publicarà en accés obert. El thesaurus s'enriquirà amb els enllaços a altres catàlegs i vocabularis.

9 . **El Thesaurus de la UB en LOD**

Podem veure un petit exemple de com es presentarà el tesaurus (ara està encara en fase proves i en construcció) i quines són les

Novetats en la funcionalitat i les seves possibilitats

- El tesaurus s'integra com un recurs més del CRAI.
- L'experiència d'usuari ens mostra:
    - estructura multilingüe completa, no només a nivell de traducció del terme sinó a nivell global.
    - cerca predictiva, que ofereix als usuaris els termes que s'aproximen a la seva construcció de cerca segons els termes usats al tesaurus.
    - incorpora les facetes a partir de l'estructura dels microtesaurus.
    - afegeix una navegació gràfica a través de constel·lacions de nodes.
- Es potencia la interoperabilitat a través de diferents EndPoints:
    - API Rest amb sortida JSON
    - Linked Open Data en RDF (SPARQL)
- Vinculació a múltiples eines de descoberta no només Cercabib.
- Ajuda a simplificar i optimitzar els processos d'edició i revisió.

10 **Representació gràfica del Thesaurus**

Aquí tenim la representació gràfica del Thesaurus. La constel·lació de nodes sembla que ens ha sortit amb estructura de galàxia.

Fent un símil podríem dir que en el centre tenim les estrelles compactes que equivalen als temes amb jerarquies més completes i en la part exterior les estrelles més solitàries que equivalen als termes més nous que s'han d'anar vinculant, relacionant, associant, etc...

11 Què ens queda pendent?

- Completar les equivalències en castellà, anglès i francès. Actualment tenim el 80% de termes i ens falta el 20 % (majoritàriament noms geogràfics)

Les vinculacions SKOS a altres vocabularis ens permeten agilitzar els processos a l'hora d'establir correspondències.

- Completar els nodes que falten en l'estructura jeràrquica: Hi ha un 12,6% de termes que encara no estan associats a un descriptor genèric però que podrien relacionar-se amb termes d'alguna branca que ja està parcialment desplegada.

Les noves funcionalitats ens han permès visualitzar de forma més clara les mancances en les jerarquies i facilitar-nos la identificació de possibles relacions i l'anàlisi de l'estructura del tesaurus.

- Ampliar els horitzons: Les institucions interessades en el tesaurus començaran a usar-lo? Els usuaris interns perceben la millora?...

12 Més informació:

Fabeiro, Rosa; Casals, Pep: El Thesaurus de la Universitat de Barcelona: 25 anys en constant evolució.

Iniciatives de dades obertes vinculades en les Biblioteques

Datos enlazados en la BNE

Dades vinculades en la Bibliothèque nationale de France

Dades vinculades en la British Library

Dades vinculades en la Deutsche Nationalbibliothek

Dades vinculades de la Library of Congress (Linked Data Service)

Dades vinculades dels Getty Vocabularies (Linked Open Data)

13 Moltes gràcies per la vostra atenció!
