El Thesaurus de la Universitat de Barcelona : balanç d'un projecte

Carme Masagué (cmasague@ub.edu), Rosa Fabeiro (rfabeiro@ub.edu)

Unitat de Procés Tècnic.
Centre de Recursos per a l'Aprenentatge i la Investigació de la UB.

Introducció

Ja han passat 21 anys des de que el gener de 1992 es va publicar la primera edició del *Thesaurus de la Universitat de Barcelona*, i 20 anys des de que es va començar a utilitzar com a llenguatge d'indexació a la Universitat de Barcelona.

És doncs un bon moment per fer balanç d'un projecte arriscat en el seu dia però que s'ha anat consolidant al llarg d'aquest temps, tot evolucionant i adaptant-se al ritme de les necessitats de la nostra organització per guanyar en funcionalitat i aplicacions, fins a convertir-se en una eina molt útil i en sintonia amb els reptes del futur.

Per celebrar les dues dècades de creació i d'aplicació d'aquesta eina de classificació alfabètica de matèries de la Universitat de Barcelona farem un viatge a través del temps per explicar breument com va néixer, com es va aplicar i com ha anat creixent i evolucionant al llarg d'aquests anys fins a l'actualitat i quines són les perspectives de futur que preveiem.

La classificació alfabètica de matèries a la Universitat de Barcelona: inicis 1967-

Durant l'època franquista i fins a finals dels anys 70, la classificació alfabètica de matèries en la Universitat de Barcelona es feia en llengua castellana.
A partir de la publicació l'any 1967 de la primera edició *de Lista de encabezamientos de materia para bibliotecas* de Carmen Rovira i Jorge Aguayo, la "*Lista*" va ser ràpidament adoptada i difosa dins l'àmbit bibliotecari espanyol per a la classificació de matèries en els catàlegs de les biblioteques públiques i universitàries. Va ser així com va començar a utilitzar-se a la Universitat de Barcelona.
La "*Lista*", basada en la *Library of Congress Subject Headings* (*LCSH*) de la Biblioteca del Congrés de Washington, havia nascut dins el "Programa de Fomento de Bibliotecas y de la Bibliografía" de la Unión Panamericana que tenia com a objectiu obtenir una llista de termes en castellà de caràcter general i extensa que permetés la uniformització dels encapçalaments de matèria en les biblioteques de diferents països de parla castellana per dur a terme projectes cooperatius de catalogació.

Els anys 80s : normalització lingüística i automatització de la Biblioteca de la Universitat de Barcelona

Amb l'arribada de la democràcia i un cop restablerta la Generalitat de Catalunya un dels temes claus d'acció del govern català va ser iniciar el procés de normalització lingüística i cultural. Recuperar el català com a llengua pròpia de Catalunya era una tasca pendent que l'Estatut de Catalunya (1979) va recollir a l'article 3 i que la "Llei 7/1983, de 18 d'abril, de normalització lingüística a Catalunya" va desenvolupar posteriorment.
Diferents institucions iniciaven projectes per fomentar i normalitzar l'ús de la llengua catalana. Entre les diverses iniciatives, dins l'àmbit bibliotecari on va participar la Universitat de Barcelona, destaca l'acord que l'any 1982 es va signar amb l'Institut Català de Bibliografia (ICB), on es va decidir incorporar encapçalaments i subencapçalaments de matèria en català als 25.000 registres bibliogràfics existents en aquell moment en la Biblioteca de la Universitat de Barcelona. A banda d'incorporar les matèries en català, la Universitat va mantenir també els encapçalaments i subencapçalaments de matèries en llengua castellana que s'havien assignat seguint la *Lista de encabezamientos de materia para bibliotecas* de C. Rovira i J. Aguayo (1967 ; 1969 ; 1970).

Posteriorment, l'Institut Català de Bibliografia va publicar la *Llista d'encapçalaments de matèria en català*, (edició preliminar de l'any 1988 i edició preliminar actualitzada de l'any 1991) traducció i adaptació al català de la *Lista de encabezamientos de materia para bibliotecas* de C. Rovira i J. Aguayo i de la *Library of Congress Subject Headings.* La Universitat de Barcelona va utilitzar la "Llista d'encapçalaments" per a la classificació alfabètica de matèries des del 1988 fins l'any 1992.

Per altra banda, l'any 1985, la Universitat de Barcelona iniciava un gran projecte: l'automatització de la Biblioteca de la Universitat de Barcelona.

En aquells anys la Universitat de Barcelona s'esforçava per millorar les infraestructures i l'equipament bibliotecari augmentant les superfícies dedicades a biblioteca, incorporant mobiliari, eines funcionals i donant suport al treball en la tasca de la seva informatització. L'objectiu principal era el de crear un catàleg únic per poder accedir a tots els fons que estaven repartits entre la "Biblioteca central" i les biblioteques de departaments i facultats.

L'accés electrònic al catàleg únic mitjançant ordinadors remots agilitzaria enormement la descripció i l'accés als fons de les biblioteques. El nou sistema informatitzat facilitaria les cerques i evitaria als usuaris un munt de desplaçaments de biblioteca en biblioteca. Enrere quedaven les fitxes de cartolina, els calaixos de fusta i el treball manual dels bibliotecaris. El canvi cap a l'automatització havia estat no només tecnològic sinó també normatiu i estructural.

L'any 1985 es va iniciar la catalogació centralitzada dels fons de la Biblioteca de la Universitat de Barcelona per crear el catàleg únic, però ben aviat van sorgir problemes amb la indexació dels documents a causa del seu nivell d'especialització.

Els anys 90s: Llista o Thesaurus?

A finals de l'any 1990, el Servei de Catalogació de la UB i la comunitat universitària ja veien com els termes de la "Llista d'encapçalaments" no eren suficients per a la indexació dels seus fons. El problema principal era la manca de termes tècnics i especialitzats en diverses àrees temàtiques i encara que l'edició preliminar actualitzada de l'any 1991 de la *Llista d'encapçalaments de matèria en català* va augmentar el nombre de termes pensant en les necessitats de les biblioteques especialitzades, es continuava detectant aquesta mancança terminològica mentre el nombre de propostes de nous termes en el Servei de Catalogació s'anava multiplicant.

Per aconseguir unes classificacions més exactes i específiques, els catalogadors van començar a consultar la *Library of Congress Subject Headings* (LCSH). En molts documents en llengua anglesa trobar el terme en la LCSH era una tasca senzilla. Un cop trobats els encapçalaments i subencapçalaments en anglès, es buscaven els equivalents en català i en castellà en diccionaris generals i/o especialitzats, en l'*Enciclopèdia Catalana,* i en altres obres de referència. Aquesta tasca però, de vegades era més complicada i requeria més temps, (pensem que en aquells anys no teníem per exemple la quantitat d'eines que tenim ara accessibles en Internet). Al llarg del procés d'adaptació dels termes de la *LCSH* es va iniciar també un procés de modulació d'alguns termes que deixaven entreveure prejudicis de caire sexista, ètnic o religiós o d'altres amb punts de vista no gaire objectius.

Per portar un control dels termes provinents de la LCSH que no constaven en la "Llista d'encapçalaments" es redactaven unes fitxes amb tota la informació recopilada: epígrafs en anglès, en català i castellà, referències creuades, etc.
L'ús de la *LCSH* i el número de fitxes anava en augment. Era evident que la manca d'un vocabulari extens i especialitzat en català i també en castellà estava creant problemes en la indexació i alentia la catalogació dels documents en el Servei de Catalogació.

Per altra banda les noves tecnologies venien acompanyades de noves eines: nous formats, nous estàndards, i nous llenguatges de classificació com per exemple els tesaurus. Calia estudiar si els tesaurus podien ser una alternativa per a la solució al problema de la indexació.

Encara que els primers tesaurus havien nascut a principis del segle XX per a la gestió d'informes tècnics i altre tipus de documentació especialitzada com un enfocament alternatiu a les classificacions precoordinades, no va ser fins a meitat del segle passat, amb l'arribada de l'automatització a les biblioteques i als centres de documentació, quan van agafar més embranzida (Uniterm (1955), Thesaurus of Engineering and Scientific Terms (1967), Thesaurofacet (1961), etc.). Alguns professionals del món de la informació van preferir crear els seus propis vocabularis i treballar assignant conceptes als documents en lloc de temes. Amb aquest sistema no calia establir estructures temàtiques precoordinades. Fins que un usuari no feia una cerca en l'ordinador els conceptes no s'organitzaven o combinaven de cap manera. Havia nascut la indexació postcoordinada i els tesaurus agilitzaven les tasques i milloraven l'accés a la documentació.

Tornant al Servei de Catalogació, i per solucionar el problema de la indexació es va plantejar la necessitat d'avaluar els avantatges i inconvenients d'utilitzar una llista o un

tesaurus en el nou entorn automatitzat i analitzar si amb un nou llenguatge s'agilitzarien els processos. A més, s'havien de preveure les possibilitats d'aplicació i ús d'aquestes eines a curt i llarg termini avaluant també la creació d'un tesaurus propi.

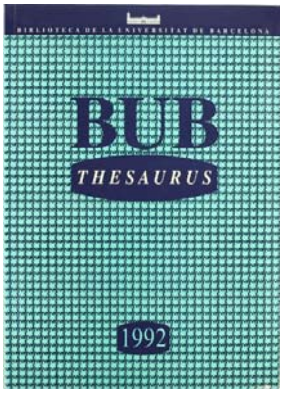
Seria molt llarg analitzar tots els pros i els contres de les llistes i els tesaurus però en línies generals els aspectes avaluats van ser els següents:

- **Ús**. - A nivell d'ús, les llistes són eines que han nascut pensades per a ser utilitzades en sistemes manuals i els tesaurus en sistemes automatitzats.
- **Complexitat.-** Les llistes s'organitzen a partir de temes representats per entrades principals o encapçalaments que poden portar subencapçalaments o entrades secundàries, subdivisions sota noms geogràfics, de llengües o subdivisions comunes, etc., és a dir, tenen una estructura més complexa que els descriptors d'un tesaurus i estableixen normes i regles sintàctiques entre les diferents nocions. En els tesaurus els conceptes poden associar-se lliurement entre sí.
- **Recuperació**. - Les llistes són llenguatges precoordinats, és a dir coordinen els diferents temes, en el moment de la descripció del document, abans de l'emmagatzematge dels documents i els tesaurus són llenguatges postcoordinats, combinen els conceptes en el moment de la cerca documental, no abans, per això tenen més possibilitats en l'anàlisi documental.
- **Formació**. - La complexitat de la sintaxi pre-coordinada de les llistes d'encapçalaments requereix un ampli període de temps per a la formació dels catalogadors i per a la validació posterior de les indexacions. La formació en la indexació amb descriptors és més àgil.
- **Presentació.** - De cara a l'usuari la presentació alfabètica dels índexs de matèria és molt més clara i fàcil de consultar amb l'ús de descriptors que amb encapçalaments i subencapçalaments.
- **Especialització**. - Els tesaurus utilitzen un llenguatge més especialitzat i tècnic que el de les llistes.
- **Exhaustivitat.** - Els tesaurus permeten realitzar una descripció més exhaustiva del contingut temàtic dels documents, augmentat d'aquesta manera les possibilitats de recuperació des de diversos punts de vista.

Finalment, es va concloure que els tesaurus eren les eines que donaven més garanties d'exhaustivitat, pertinença i rapidesa i es va decidir adoptar aquest tipus de llenguatge com a vocabulari controlat pel catàleg de la UB. En no existir cap tesaurus en la nostra llengua capaç de cobrir la multidisciplinarietat dels nostres fons vam decidir iniciar la construcció d'un tesaurus propi.

En aquella època sovint ens preguntaven com és que canviàvem la "Llista" per un tesaurus?
Nosaltres ens preguntàvem: Perquè continuen emprant un sistema manual en un sistema automatitzat?

El Thesaurus de la Universitat de Barcelona



L'any 1992 es va publicar el *Thesaurus de la Universitat de Barcelona*, preparat per Carme Cambrodí amb la col·laboració de M. Teresa Tarrida, amb 8.500 termes (descriptors i no descriptors) de diverses disciplines amb llurs relacions corresponents i un índex permutat (índex KWIC) per facilitar la cerca.

L'any 1993 i gràcies al disseny de programes informàtics desenvolupats en la Universitat de Barcelona, es va crear una **aplicació per a la gestió** del tesaurus que facilitava enormement les tasques de creació i sobretot de relació recíproca automàtica entre els termes. També es va iniciar la tasca de reconversió del vocabulari d'indexació.

Paral·lelament a l'inici de la reconversió es van fer sessions informatives al personal de la Biblioteca per explicar les raons del canvi del llenguatge cap a un tesaurus, les característiques principals del nou llenguatge i com es duria a terme la reconversió, així com **sessions de formació** per als catalogadors sobre la nova forma d'indexació amb descriptors del tesaurus i el nou flux de treball.

Durant els anys 1993/1997 es van **reconvertir** les entrades de matèria de 280.000 registres bibliogràfics. Moltes entrades es van processar automàticament i moltes d'altres de forma manual a causa de la seva complexitat de reconversió.

Finalment, a mitjans del 1998, la Universitat de Barcelona va posar a disposició dels seus usuaris el *Thesaurus de la UB* en línia, com a **eina de cerca** de les matèries del Catàleg Bibliogràfic del fons modern de la UB.

A partir de gener de 2005, i per tal de facilitar la cerca temàtica al Catàleg del fons modern, tots els termes del "Thesaurus" es van agrupar en 29 **Microtesaurus temàtics** i l'octubre de 2007 es va ampliar la llista dels microtesaurus a 30 amb la incorporació del **Microtesaurus de Noms Geogràfics**.

El juny del 2008, amb el canvi de sistema de VTLS cap a Millennium, es van unificar els registres de fons antic i fons modern en una única base de dades i així, el "Thesaurus" es va convertir també en l'eina de cerca de matèries del **Fons Antic de Reserva**. També es va deixar d'indexar en castellà (encara es mantenia la doble indexació en català i castellà iniciada l'any 1982) i es va començar a treballar en el projecte de tesaurus multilingüe amb la incorporació en els registres d'autoritat matèria de les equivalències en castellà, anglès i francès dels descriptors en català.

El desembre de 2009 es va incorporar al catàleg la cerca per **gènere i forma** per tal de facilitar l'accés a un conjunt de materials molt diversos en diferents suports i amb diferents característiques físiques. Fins aquell moment, l'accés a aquesta diversitat de materials havia estat limitat a la recuperació per matèries o per autors i els usuaris no havien tingut sempre accés a aquestes dades. Amb la incorporació al catàleg de la cerca per gènere/forma es va facilitar l'accés a obres de referència, material

audiovisual, documents gràfics, recursos electrònics, materials amb continguts especials per a persones amb discapacitats visuals i auditives, etc. entre d'altres.

El Thesaurus de la UB en l'actualitat : una eina de futur

L'aplicació del nostre llenguatge controlat s'ha ampliat darrerament més enllà del catàleg bibliogràfic (UB i CCUC) com a eina per a la indexació dels diferents repositoris creats en els últims anys per donar difusió tant al nostre fons patrimonial com als resultats de l'activitat docent i investigadora de la nostra comunitat.

Actualment el THUB és l'eina per indexar tots els recursos dipositats a:

- al dipòsit institucional, Dipòsit Digital de la UB,
- als repositoris patrimonials Memòria Digital de Catalunya, Portal Universitat de Barcelona. Fons de reserva de la Biblioteca Virtual Cervantes
- al repositori de Revistes científiques UB
- a UBTV, el portal vídeo de la UB
- als diferents repositoris consorciats on participa la UB (RACO, TDX, RECERCAT, MDX)

Des de l'inici de cada un d'aquest projectes el CRAI de la UB ha apostat per l'ús d'aquesta eina com a llenguatge controlat, en primer lloc per considerar-la una eina idònia per a la recuperació en un entorn de cerca per paraula clau, en segon lloc, per adequar-se al format de catalogació en metadades i, en definitiva, per la decisió estratègica del CRAI de normalitzar tots els recursos d'informació de què disposa de manera centralitzada a través de la Unitat de Procés Tècnic i de manera col·laborativa i coordinada amb la resta d'unitats i grups de treball implicats en cada un dels projectes. Aquesta manera col·laborativa i transversal de treballar ens ha servit per establir procediments i fluxos de treball de valor afegit tant per a la millora de la cerca als diferents dipòsits com per a la pròpia millora del Thesaurus UB.

En aquesta línia, i com a acció de suport per a la promoció de la publicació en obert de la recerca realitzada a la Universitat de Barcelona, el CRAI va posar en marxa una prova pilot entre juny i octubre de 2012 per a la interconnexió del portal GREC per a la gestió de la RECERCA, amb el Dipòsit Digital de la UB (DDUB). De manera automàtica i a partir dels currículums dels investigadors, els articles publicats en revistes científiques que permeten l'accés obert, passen directament al Dipòsit Digital. La integració d'ambdues aplicacions inclou també el disseny d'un flux de treball automàtic que garanteix la validació, per part de bibliotecaris del CRAI, de les metadades que acompanyen als articles abans de la seva publicació al DDUB.

En aquest projecte, la **integració del Thesaurus en el GREC**, ha permès facilitar la feina d'indexació dels documents continguts als currículums dels professors per part dels bibliotecaris. D'aquesta manera, des de la pròpia aplicació del GREC, es revisa la informació, s'assignen termes del Thesaurus, s'adjunten els articles i es dipositen en el DDUB.

La utilització del Thesaurus per a la indexació en els diferents repositoris ens ha facilitat el camí per assolir el proper projecte previst que és la creació del **Thesaurus multilingüe**. A partir de 2008 es comença a treballar en aquesta direcció amb la incorporació en els registres d'autoritat de matèria de les equivalències en castellà, anglès i francès dels descriptors en català.

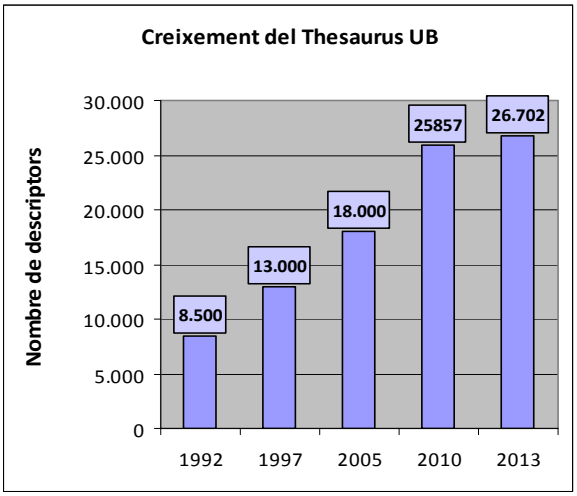
Respondre a les necessitats de termes de matèria en anglès, com a requeriment per a la recol·lecció dels registres dels repositoris, ha promogut el creixement de registres del tesaurus amb totes les equivalències lingüístiques fins al 52,3 % en l'actualitat.

En paral·lel a les tasques de construcció del Thesaurus multilingüe estem treballant en un projecte de millora de la interoperabilitat del Thesaurus en la línia del projecte LOD (Linked Open Data) estudiant els canvis tecnològics necessaris del programa actual de publicació del tesaurus en línia per tal d'oferir-lo en **format SKOS** i implementar serveis d'enllaç a dades en obert.

Creixement i dades

El THUB ha de respondre a les necessitats d'accés conceptual als continguts del fons del CRAI de la Universitat de Barcelona. Quan el fons del CRAI s'incrementa amb nous recursos sobre noves disciplines la demanda de termes d'indexació per matèries augmenta per tal de donar accés i facilitar la recuperació d'aquest material a l'usuari final. L'adscripció de centres especialitzats, la donació de fons bibliogràfics, de fons personals, de col·leccions especials, la documentació dipositada en els repositoris de la Universitat, etc. són alguns dels factors que fan créixer el tesaurus.

Per aquesta raó el creixement en nombre de descriptors, després dels primers anys de conversió ha continuat creixent de manera continuada fins arribar a les dades actuals que presentem a continuació desglossades per tipologies de termes:



26.702 descriptors dels quals

- 20.222 descriptors temàtics
- 6.265 descriptors geogràfics
- 215 descriptors de gènere/forma

13.239 no descriptors dels quals

- 10.456 no descriptors temàtics
- 2.641 no descriptors geogràfics
- 142 no descriptors de gènere/forma

Pel que fa als descriptors en català amb equivalències amb castellà, anglès i francès, són **14.125 descriptors** que representen el 52,3% del THUB.

26.702	Nombre de registres del Thesaurus
26.561	Nombre de relacions jeràrquiques
12.750	Nombre de termes relacionats
29.594	Nombre de referències
14.125	Nombre de registres multilingües

Pel que fa al nombre de consultes que rep el THUB també ha crescut de manera seqüencial. Les dades que oferim a continuació recullen les visites a l'aplicació web a partir de 2009, data en que es va separar la cerca del Catàleg d'Autoritats UB de la cerca del Thesaurus UB.

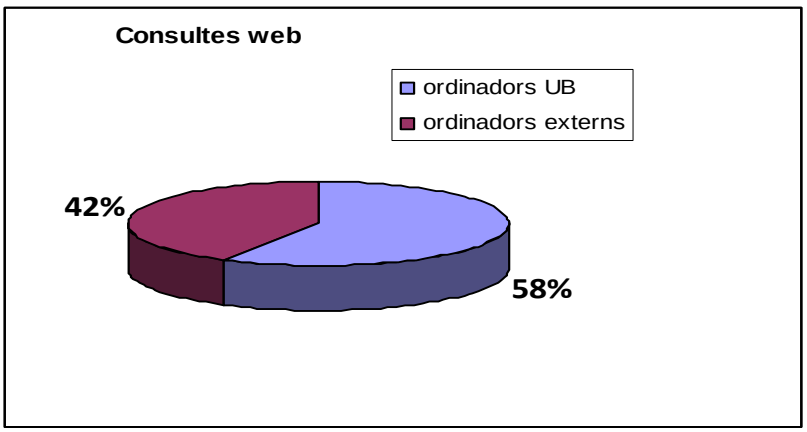
	Nombre visites	ordinadors UB	ordinadors externs
***2009**	**27.128**	**19.099**	**8.029**
2010	**56.372**	**27.075**	**29.297**
2011	**79.147**	**35.669**	****43.478**
2012	**46.229**	**33.842**	**12.387**
****2013**	**33.189**	**25.837**	**7.352**
Totals	**242.065**	**141.522**	**100.543**
Promig de consultes	48.413,00	28.304,40	20.108,60

* 2009 només des de juny per canvi de programari

**2013 dades recollides fins a maig

***Augment inesperat de consultes externes per descàrrega automàtica dels termes del THUB

En els darrers anys s'observa que el promig de consultes tant des d'ordinadors UB com des d'ordinadors externs està bastant igualat, dada que ens confirma l'ús d'aquesta eina per usuaris de fora de la nostra comunitat.



Al llarg d'aquest anys hem rebut peticions d'ús d'altres institucions que s'han interessat pel nostre Thesaurus com eina d'indexació, com per exemple el Departament de Medi Ambient de la UAB, i el Centre d'Estudis d'Opinió, òrgan autònom de la Generalitat. La nostra institució es mostra oberta a compartir aquesta eina amb altres organismes que així ho expresin.

Com a curiositat des de novembre de 2012 està recollit en el Thesaurus Portal , portal per a facilitar la recerca de vocabularis controlats i eines d'indexació, el desenvolupament d'ontologies, la informació d'alfabetització, etc.

Més informació : Thesaurus UB