CiT, portal de terminologia de ciències i tecnologia

Introducció

Des de la Declaració de Berlín sobre l'accés lliure al coneixement de les ciències i de les humanitats (2003) (*http://bid.ub.edu/15vela.htm*) —a la qual s'adheriren en el seu moment tant la Generalitat de Catalunya com l'Institut d'Estudis Catalans (IEC)—, la Secció de Ciències i Tecnologia (SECCT) treballa en aquesta direcció. Aquesta secció de l'IEC entén que una de les maneres contemporànies d'obrir els seus fons als estudiosos és aprofitar els avantatges de les tecnologies de la informació i la comunicació. Alhora, aquesta secció també està molt sensibilitzada amb la idea que determinades dades de les institucions públiques o privades haurien de ser de lliure accés per a tothom per ser reutilitzades sense restriccions d'ús ni mecanismes de control, d'acord amb el moviment internacional conegut com a *open data* o *dades obertes*.

En aquest context, la SECCT emprengué la creació d'una plataforma informàtica amb la finalitat de posar a l'abast de la comunitat científica la terminologia de ciències i tecnologia generada per l'IEC, per les seves societats filials o per altres institucions vinculades, perquè s'hi pogués accedir de manera lliure, immediata i sense restriccions (*open access*), i que estigués protegida per llicències Creative Commons. Recentment, aquesta plataforma s'ha fet pública mitjançant el portal CiT (*http://cit.iec.cat*).

El portal CiT

El portal CiT (Terminologia de Ciències i Tecnologia) és un web que aplega obres de terminologia de ciències (en un sentit ampli) i de tecnologia. Està concebut com una estructura web amb diferents mòduls, cadascun dels quals correspon a una eina. Actualment, n'hi ha tres d'operatives (vegeu la figura 1):

- *BiblioCiT* és una biblioteca en línia d'obres terminològiques que poden ser consultades individualment fent servir opcions de cerca avançada i aplega actualment unes 170.000 unitats terminològiques catalanes, procedents de trenta-quatre obres.
- *CercaCiT* és un motor de cerca avançat que permet consultar totes les obres de la BiblioCiT simultàniament o per àrees temàtiques. Actualment, s'hi poden consultar al voltant de 140.000 unitats terminològiques catalanes.
- *ContextCiT* és una eina que permet trobar contextos d'ús dels termes apareguts en les revistes especialitzades recollides en l'Hemeroteca Científica Catalana (HCC).

La plataforma tecnològica està construïda sobre una base de dades relacional i un entorn de consulta web basat en pàgines dinàmiques.

BiblioCiT

La biblioteca en línia BiblioCiT inclou obres de naturalesa diversa, ja que conté diccionaris, diccionaris enciclopèdics, vocabularis, nomenclatures, terminologies de manuals científics, terminologies universitàries i obres de rellevància terminològica. Així, per exemple, obres complexes com el *Diccionari de geologia* presenten articles lexicogràfics que poden tenir les parts següents: entrada, categoria gramatical, llengua d'origen, valoratius lingüístics, àrees temàtiques, definició, accepcions, informació enciclopèdica, sinònims, termes relacionats, equivalències a altres llengües, etimologia, notes i il·lustracions; mentre que un vocabulari d'estructura simple acostuma a tenir l'entrada, la categoria gramatical i les equivalències a altres llengües.

El fet que les obres siguin de naturalesa diversa fa que cadascuna tingui una estructura pròpia, que s'ha respectat per tal de conservar-ne l'autoria, la identitat i la integritat.

FIGURA 1. Pàgina d'inici del portal CiT

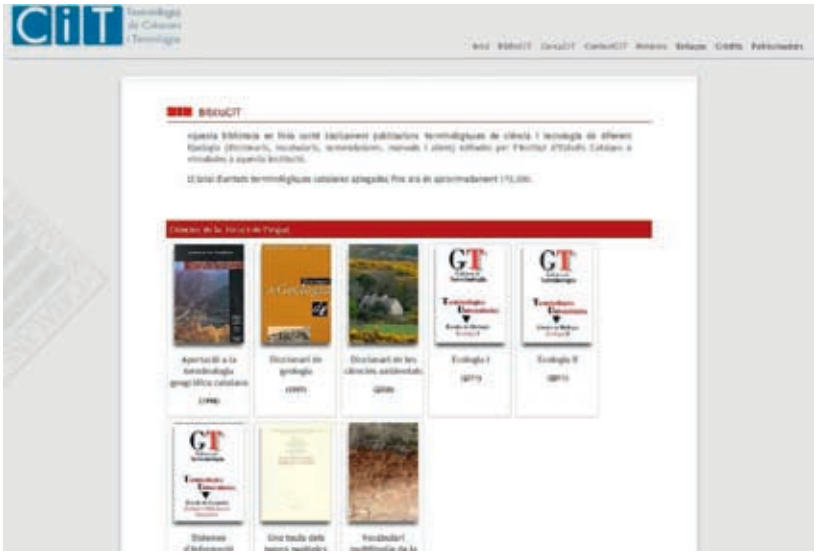

Figura 2. Pàgina d'inici de BiblioCiT

Per aquest motiu, els camps en els quals es poden fer consultes no són idèntics en totes les obres. Per exemple, el *Vocabulari multilingüe de la ciència del sòl* permet la cerca per àrea temàtica (tots els termes en tenen una o més d'una), però no la cerca per definició, ja que no n'hi ha; mentre que l'*Aportació geogràfica a la terminologia catalana* no disposa de cerca per àrees temàtiques, però sí que s'hi poden fer cerques en la definició.

Les obres es presenten agrupades en «prestatges» segons l'àmbit d'especificitat a què pertanyen: «Ciències de la Terra i de l'espai», «Ciències de la vida i la salut», «Ciències exactes i experimentals», «Ciències socials i humanes» i «Enginyeria i tecnologia» (vegeu la figura 2).

La major part de les obres de la BiblioCiT han estat publicades originàriament en paper per l'IEC; per altres institucions, com la Universitat de les Illes Balears, o per editorials de prestigi reconegut, com ara Enciclopèdia Catalana, Barcino o Reverté. Tot i això, també conté obres inèdites, com ara el *Diccionari de les ciències ambientals*, publicat fins ara únicament en línia en la plataforma CiT.

CercaCiT

El motor de cerca CercaCiT permet la cerca transversal de totes les obres recollides en la BiblioCiT. Es tracta d'un motor de cerca avançat que permet consultar totes les obres de manera simultània, però també de manera selectiva i amb criteris de cerca diversos.

Figura 3. CercaCiT: exemple de cerca que contingui el terme *sòl* en les obres de «Ciència de la Terra i de l'espai»

A diferència del que passa a la BiblioCiT, al CercaCiT s'han homogeneïtzat les fitxes de les unitats terminològiques, de tal manera que totes tenen la mateixa estructura i es mostren els mateixos camps d'informació: entrada, categoria gramatical, títol de l'obra de què prové l'entrada i àrea temàtica. A més, també apareixen, si n'hi ha, les abreviacions, la definició, les equivalències, els exemples d'ús, els sinònims i els termes relacionats.

El motor de cerca permet fer dos grans tipus de seleccions: indicar l'obra o obres en què es vol fer la consulta, d'una banda, i indicar en quin camp o camps volem que es faci la cerca i com s'ha de fer, de l'altra. Quant a la tria d'obres, es pot cercar en totes, especificar-ne una de concreta o decantar-se per un dels blocs temàtics d'obres de la BiblioCiT. Pel que fa a la segona selecció, es pot especificar en quins camps es fa la consulta (entrada, definició, equivalències, sinònims, exemples i termes relacionats o en tots els camps alhora) i quin grau de coincidència hi ha amb la cadena de caràcters introduïda (coincident, començada per, acabada per o que conté). A més, per mitjà d'un menú desplegable, es pot fer una cerca per àrees de coneixement, cosa que permet delimitar-la més que quan se seleccionen unes quantes obres (vegeu la figura 3).

ContextCiT

Mentre que les dues eines anteriors se centren en el significat i les relacions lingüístiques dels termes (de sinonímia, d'equivalència, etc.), el ContextCiT és un cercador avançat que troba termes en el seu context. El corpus sobre el qual treballa és el que formen les revistes recollides en la HCC de l'IEC (*http://revistes.iec.cat*), que recull més de 17.000

articles. Els textos especialitzats són una font molt rica en unitats terminològiques, ja que les difonen i, fins i tot, les fixen. Per aquest motiu i aprofitant que bona part de les revistes de l'HCC estan digitalitzades en format PDF —algunes en PDF original i d'altres escanejades—, amb l'ajuda d'un cercador de Google personalitzat, es fan consultes en els documents, els resums i les metadades de la HCC. Actualment, però, la cerca es limita a aquells documents que incorporen sistemes de reconeixement de caràcters (OCR).

Treball futur

El programa CiT és un projecte viu que, a banda de continuar publicant obres en línia i actualitzant les que ja inclou, té previst crear noves eines.

ExtractCiT

Aquesta eina és fruit de la col·laboració entre el programa CiT i el grup IULATERM, de la Universitat Pompeu Fabra. Aquest grup ja ha desenvolupat eines automàtiques de tractament de la informació general i especialitzada, com ara un sistema d'extracció automàtica de terminologia de textos en suport digital, Terminus 2.0 (*http://terminus.iula.upf.edu/cgi-bin/terminus2.0/terminus.pl*).

Les tasques que s'han realitzat fins ara dins el programa es limiten a la constitució d'un corpus i a l'extracció de terminologia de les revistes editades per l'IEC en llengua catalana i digitalitzades a la HCC. Aquest buidatge terminològic no pressuposa l'elaboració de fitxes terminològiques completes, sinó una llista dels possibles candidats a terme acompanyats del grau de «terminologicitat», de la freqüència d'aparició, dels seus contextos d'ús i de la referència precisa —obra i localització dins l'obra.

Un cop enllestit aquest buidatge, la terminologia s'organitzarà en format de base de dades amb accés lliure, per tal que la comunitat científica disposi d'una font actualitzada de terminologia en ús (d'aquesta manera, es podran fer estudis sectorials o transversals i analitzar la vida dels termes). Pel que fa a la consulta, de manera semblant a com funciona el CercaCiT, es podrà seleccionar la revista la terminologia de la qual es vol consultar. A més, hi haurà un motor de cerca que permetrà buscar termes a totes les revistes simultàniament o només en una selecció.

Conclusions

La terminologia sempre va lligada a l'avenç científic i tecnològic, ja que n'és l'element vehicular més essencial. Per aquest motiu, vetllar per la difusió d'aquestes unitats lèxiques i fer-les accessibles en línia i de manera simple a usuaris de tota mena —des de científics a estudiants passant pels professionals de la llengua— és una tasca que correspon, en part, a la SECCT de l'IEC.

La pròpia naturalesa de les obres terminològiques catalanes, destinades a un públic especialitzat molt reduït, fa que les edicions publicades en paper siguin de tiratges també reduïts i esdevinguin descatalogades en un període de temps breu. Sovint, passa el mateix amb algunes obres científiques en català que no són estrictament terminològiques. El portal CiT dóna una nova vida a aquestes obres i a la terminologia continguda: les posa en un entorn informàtic —que els proporciona més accessibilitat, visibilitat i funcionalitat tot respectant l'autoria, la integritat i la identitat dels continguts originals— i, al mateix temps, fa que formin part d'una xarxa terminològica que s'enriqueix amb cada nova obra inclosa.

Per això, l'estructuració feta de les bases de dades consultables al portal CiT, procedents d'obres molt diverses i de concepció lexicogràfica molt variada, amb l'ajut de cercadors avançats i amb criteris de cerca diversos, permet trobar noves correlacions entre els significats dels termes i les seves relacions lingüístiques i terminològiques.

Agraïments

El programa CiT no s'hauria pogut dur a terme sense els ajuts de la Secretaria Científica de l'IEC (2008-2014) i la generosa contribució de la Fundació PuntCat. El programa CiT vol agrair el suport rebut dels serveis de l'IEC. El Servei d'Informàtica (Xavier Torrents) ha desenvolupat les eines, les bases de dades i els motors de cerca que han permès l'edició dels reculls terminològics en línia i la cerca transversal. El Servei de Correcció Lingüística (Josep M. Mestres) ha vetllat per les obres i els textos publicats en el portal CiT. El Servei de Recursos Digitals (Santi Muxach) ha tingut cura de l'aspecte visual i gràfic de la interfície del portal CiT.

Alhora, el programa CiT vol donar les gràcies al grup IULATERM —i, especialment, a M. Teresa Cabré i Mercè Lorente— pels extrets terminològics de les revistes de la HCC, punt de partida de l'ExtractCiT, i a M. Magdalena Ramon, per la seva col·laboració en l'elaboració de les edicions en línia de les terminologies universitàries de la Universitat de les Illes Balears. Finalment, també volem agrair el suport rebut d'Agustí Mayor, tècnic lingüístic de l'Ajuntament de la Vila Joiosa, coordinador del projecte (2004-2008) i actualment col·laborador.

Joana Torres i Mireia Trias
Programa CiT
Secció de Ciències i Tecnologia
Institut d'Estudis Catalans

Salvador Alegret
Secció de Ciències i Tecnologia
Institut d'Estudis Catalans