MEMÒRIES DE LA SOCIETAT CATALANA DE TERMINOLOGIA, 8

Judit Feliu i Mireia Trias (cur.)

VIQUIPÈDIA I TERMINOLOGIA

scatterm
SOCIETAT CATALANA DE TERMINOLOGIA
Filial de l'Institut d'Estudis Catalans

Institut d'Estudis Catalans

Viquipèdia i terminologia

SOCIETAT CATALANA DE TERMINOLOGIA
FILIAL DE L'INSTITUT D'ESTUDIS CATALANS

MEMÒRIES DE LA SOCIETAT CATALANA DE TERMINOLOGIA, 8

JUDIT FELIU I MIREIA TRIAS
(curadores)

Viquipèdia i terminologia

BARCELONA, 2021

Carrer del Carme, 47. 08001 Barcelona

Primera edició: setembre del 2021

Text revisat lingüísticament per la Unitat d'Edició del Servei Editorial de l'IEC

Disseny de la coberta: Zink Comunicació SL

Fotografia de la coberta: Una de les gàrgoles del pati de la Casa de Convalescència, seu de l'Institut d'Estudis Catalans. Autor: Agustí Espallargas (Societat Catalana de Terminologia)

Compost per Fotocomposició gama, s. l.

ISBN: 978-84-9965-605-2

Taula

Abreviacions emprades pels autors

ACL	Association for Computational Linguistics
AENOR	Asociación Española de Normalización y Certificación
AMS	American Mathematical Society
API	*application programming interface* ('interfície de programació d'aplicacions')
APLE2	«Neologismos generales y neologismos especializados» (projecte de l'IULA)
CAT	candidat a terme
cat.	categoria
CD	coeficient de domini
CIAP	classificació internacional d'atenció primària
CIM	classificació internacional de malalties
CiT	Ciències i Tecnologia (portal terminològic)
coord.	coordinador, coordinadora
CRG	Centre for Genomic Regulation
cur.	curador, curadora
DIEC2	*Diccionari de la llengua catalana*, 2a ed. (Institut d'Estudis Catalans)
DME	*Diccionari de matemàtiques i estadística*
ed.	edició; editor, editora
EMS	European Mathematical Society
EN	anglès
ES	espanyol
ESA	*explicit semantic analysis* ('anàlisi semàntica explícita')
et al.	*et alii* ('i altres')
EWN	EuroWordNet
FD	frontera de domini
FT-I	Fundació Torrens-Ibern

HTML	*hypertext markup language* ('llenguatge d'etiquetatge d'hipertext')
IEC	Institut d'Estudis Catalans
IPC	Independentistes dels Països Catalans; índex de preus al consum; *inter-process communication* ('comunicació entre processos')
ISO	International Organization for Standardization (Organització Internacional per a la Normalització)
IULA	Institut de Lingüística Aplicada
IUPAC	International Union of Pure and Applied Chemistry (Unió Internacional de Química Pura i Aplicada)
IUPAP JWP	International Union of Pure and Applied Physics Joint Working Party
K	mil
LaTex	Lamport TeX
LC	longitud de camins
LMC	longitud mitjana de camins
M	milió
Mc	moscovi
MCR	Multilingual Central Repository
MeSH	*medical subject headings* ('paraules clau de termes mèdics')
MSCT	«Memòries de la Societat Catalana de Terminologia» (col·lecció)
NC	nombre de camins
NG	sistema Niemann-Gurewych
Nh	nihoni
NJ	nom-adjectiu
NPN	nom-preposició-nom
núm.	número
Og	oganessó
OLIF	*open lexicon interchange format* ('format d'intercanvi de lèxics oberts')
OPUS	Open Parallel Corpus
p.	pàgina
p. ex.	per exemple
pH	potencial d'hidrogen
PLN	processament del llenguatge natural
PNN	professor no numerari
RDF	*resource description framework* ('marc de descripció de recursos')
SCATERM	Societat Catalana de Terminologia
SCB	Societat Catalana de Biologia
SIAM	Society for Industrial and Applied Mathematics
SNOMED-CT	*systematized nomenclature of medicine - clinical terms* ('nomenclatura sistematitzada de medicina - termes clínics')
SPARQL	*simple protocol and RDF query language* ('protocol simple i llenguatge d'interrogació RDF')
SQL	*structured query language* ('llenguatge d'interrogació estructurat')

t	terme
TERMCAT	Centre de Terminologia de la Llengua Catalana
Ts	tennes
UOC	Universitat Oberta de Catalunya
UPC	Universitat Politècnica de Catalunya
URL	*uniform resource location* ('localitzador uniforme de recursos')
UV	Universitat de València
v.	volums
vol.	volum
W3C	World Wide Web Consortium
WN	WordNet
WND	WordNet Domains
WP	Wikipedia (Viquipèdia)
XML	*extensible markup language* ('llenguatge d'etiquetatge extensible')
XWND	Extended WordNet Domains
YAGO	Yet Another Great Ontology
YATE	Yet Another Term Extractor

Presentació

Miquel-Àngel Sànchez Fèrriz
President de la SCATERM[1]

Aquest volum número 8 de la col·lecció «Memòries de la Societat Catalana de Terminologia» (MSCT), com els tres números anteriors, representa més aviat continuïtat que no pas canvi. I, tal com es va fer amb els quatre números anteriors, es publica en format digital obert, seguint la tendència —ara consolidada— que vam iniciar des de la represa de la publicació de la col·lecció, mantenint, però, la possibilitat de preveure tirades impreses per encàrrec.

La continuïtat en la forma és present una vegada més en les idees i en els objectius que orienten els continguts des del primer número de la col·lecció: presentar als socis i als especialistes interessats —i d'un temps ençà, *urbi et orbi*— els textos de caràcter científic que generen les activitats de la Societat Catalana de Terminologia (SCATERM).

La primera secció d'aquest volum, «Estudis i projectes», aplega el material textual de la darrera jornada organitzada per la SCATERM —la XVI Jornada, celebrada el 30 de maig de 2019 i dedicada a la terminologia de la Viquipèdia—, raó per la qual el volum duu el títol *Viquipèdia i terminologia*. En concret, agrupa la conferència a càrrec de Jorge Vivaldi sobre la utilitat de la Viquipèdia com a recurs terminològic; la ponència de Pau Cabot, sobre la fiabilitat de la Viquipèdia; la ponència de Ramon Garriga juntament amb Toni Hermoso, en què presenten l'àgora terminològica Viquiterm, i la taula rodona en què van participar Joan de Solà-Morales, Jordi Cuadros i Toni Hermoso sobre les experiències de tres societats filials de l'Institut d'Estudis Catalans (la Societat Catalana de Matemàtiques, la Societat Catalana de Química i la Societat Catalana de Biologia) en la Viquipèdia.

La segona secció, «Crònica del curs 2018-2019», dona compte de les activitats científiques que la SCATERM va dur a terme en ocasió de la XVI Jornada.

1. Miquel-Àngel Sànchez Fèrriz va ser president de la SCATERM fins al juny del 2020.

Seguim així, un any més, els mandats estatutaris d'«afavorir la difusió de la terminologia en llengua catalana en els àmbits científics i tècnics, […] i promoure la provisió i l'intercanvi d'informació sobre les activitats terminològiques entre usuaris i professionals de la terminologia».

Finalment, reitero la crida que feia Jaume Martí en la presentació del número 2 d'aquesta col·lecció en el sentit d'oferir la possibilitat de publicar en aquest mitjà «treballs que, per llur extensió, no podrien tenir cabuda dins les nostres publicacions periòdiques, *Terminàlia* i *Butlletí de la Societat Catalana de Terminologia*».

Agraeixo en nom de la SCATERM la dedicació que, tant les persones que s'han ocupat de la curadoria efectiva del volum —Judit Feliu, la nostra anterior vicepresidenta i responsable de Publicacions, i Mireia Trias, la nostra anterior secretària i actual vocal de Publicacions— com totes les que hi han col·laborat —sense anar més lluny, també l'Equip Directiu anterior—, han tingut en l'elaboració i l'edició d'aquest número, així com en l'organització de les activitats que hi són recollides.

Amb la satisfacció de constatar la perseverança de la col·lecció, que el 2020 va fer deu anys, espero que aquest volum continuï amb la tasca de difondre al màxim possible les activitats de la SCATERM.

Estudis i projectes

Judit Feliu i Mireia Trias (cur.)
Viquipèdia i terminologia
Barcelona: Institut d'Estudis Catalans, 2021, p. 15-47 DOI: 10.2436/15.2503.02.64

Viquipèdia: un recurs útil per a la terminologia?

Jorge Vivaldi Palatresi
Institut de Lingüística Aplicada (Universitat Pompeu Fabra)

1. Introducció

El projecte Viquipèdia (WP, de l'anglès Wikipedia) es pot considerar un dels més reeixits pel que fa a la recopilació de coneixement. Des dels seus inicis, l'any 2001,[1] fins ara, el nombre de lectors ha anat augmentant vertiginosament. Actualment, està disponible en més de tres-centes llengües[2] i el nombre de lectures de pàgines en tot el món es compten per milions cada dia.[3] Aquesta fita ha sigut possible gràcies a la gestió de la Fundació Wikimedia i a milers de persones que hi han contribuït i contribueixen amb el seu coneixement i el seu temps.

L'aproximació que segueix aquest desenvolupament, a diferència dels tradicionals, és que es tracta d'un projecte obert. Això significa que qualsevol persona que disposi d'un ordinador i de connexió a Internet pot fer-hi una contribució, ja sigui escrivint un article o modificant-ne un de ja existent. La filosofia de la Viquipèdia és que si una comunitat treballa conjuntament en el contingut d'un article o d'un àmbit, aquest millorarà amb el temps. D'aquesta manera es podria dir que un article mai no està acabat, ja que potencialment pot ésser modificat en qualsevol moment. En general, el projecte presenta una resposta molt ràpida (i superior a la de qualsevol altre recurs similar) als esdeveniments més rellevants. Aquest mode de treball fa possible que hi hagi articles vandàlics, és a dir, amb informació dubtosa o bé que responen a altres interessos. Viquipèdia ha anat modi-

1. En la pàgina «Viquipèdia en català» es poden consultar moltes dades sobre els orígens d'aquest recurs.
2. Vegeu estadístiques actualitzades a *https://en.wikipedia.org/wiki/List_of_Wikipedias*.
3. Vegeu les estadístiques globals a *https://stats.wikimedia.org/EN/*.

ficant la política d'admissió d'articles i de revisions per tal de protegir-se d'aquests problemes.

L'èxit aconseguit ha atret recercadors de molts àmbits. En conseqüència, existeix una literatura abundant en què s'analitzen diversos aspectes d'aquest projecte des de punts de vista diferents. Cal mencionar, entre altres, els treballs de revisió de la literatura existent de [1] i [2]. També existeixen dues plataformes que recullen treballs de diferents tipus: WikiLit[4] i WikiPapers.[5]

L'impacte de la Viquipèdia abasta diferents aspectes del saber que inclouen la representació del coneixement, la sociologia o l'educació, entre altres. Aquesta presentació se centrarà en aspectes relatius a l'explotació de la Viquipèdia com a font de coneixement per a projectes de processament del llenguatge natural (PLN).

Després d'aquesta introducció, en l'apartat 2, estudiarem amb cert detall l'estructura interna de la Viquipèdia. A continuació, en l'apartat 3, analitzarem breument una qüestió complexa i controvertida com és la credibilitat. En l'apartat 4, mostrarem com es poden fer consultes sistemàtiques en aquest recurs i continuarem amb l'apartat 5, en què analitzarem algunes qüestions a tenir en compte quan es fan consultes informàtiques a la Viquipèdia. En l'apartat 6, presentarem molt breument algunes aplicacions d'aquest recurs en diferents aspectes del PLN, entre els quals s'analitzaran amb cert detall algunes aplicacions de l'àmbit de la terminologia. Finalment, presentarem algunes conclusions que es poden extreure d'aquest treball.

2. Estructura de la Viquipèdia

Un usuari qualsevol pot consultar molt fàcilment qualsevol article de la Viquipèdia. Una observació atenta de qualsevol pàgina revelarà l'existència d'una gran quantitat d'informació associada. Aquest fet hauria de fer pensar que aquest recurs disposa d'una estructura capaç i prou complexa per respondre a la majoria de les necessitats d'informació de qualsevol usuari. En aquest apartat es presenta amb cert detall l'estructura de dades associada a aquest recurs.

En primer lloc, cal assenyalar que la unitat d'informació de la Viquipèdia és l'article. L'usuari, quan fa una consulta, a través d'una pàgina web, rep una pàgina com a resposta. Diferenciem *article* de *pàgina* en el sentit que l'article conté estrictament un text amb l'explicació enciclopèdica d'una unitat d'informació. La pàgina, en canvi, conté, a més de l'article, altra informació relativa a aquesta unitat, com ara les categories a les quals està vinculada, la traducció a altres llengües, bibliografia rellevant, pàgines web on es pot ampliar la informació, etc.

4. *http://wikilit.referata.com/wiki/Main_Page*.
5. *http://wikipapers.referata.com/wiki/Main_Page*.

En el text de cada article hi ha alguna paraula (o grup de paraules) que serveixen també com enllaços a altres articles de la mateixa llengua. En cada article hi ha, de mitjana, quinze enllaços d'aquest tipus. L'autor de l'article, d'acord amb la guia d'estil, escull quines són la paraula o les paraules que considera necessari associar a un enllaç per tal de facilitar la comprensió de l'article. Es tracta, doncs, d'una informació que, en potència, és semànticament rellevant. Aquests enllaços són unidireccionals i es denominen *enllaços de sortida*. De la mateixa manera, cada pàgina és apuntada per un cert nombre de paraules d'altres pàgines, el conjunt de les quals se solen denominar *enllaços d'entrada*. El conjunt d'articles i els seus enllaços formen un graf dirigit.[6]

Cada article té assignades una o més categories mitjançant el que s'acostuma a denominar *enllaços categorials*. Aquestes categories es poden veure com a classes que tenen associades una sèrie d'instàncies, que en aquest cas són pàgines. Al mateix temps, una categoria està vinculada a altres categories mitjançant enllaços anomenats *supercategories* (categories que trobem quan recorrem el graf cap al *top*) o *subcategories* (resta de categories). Encara que no ho siguin sempre, és freqüent considerar aquests enllaços com a taxonòmics.

Podem veure el conjunt de les categories com a un altre graf dirigit (en [3] trobareu una interessant anàlisi dels dos grafs). Els nusos d'aquest graf són les categories associades a cada pàgina i els enllaços són els vincles entre aquestes categories. Segons [4] les categories es poden classificar de la manera següent:

1. Categories de contingut (*content categories*): categories destinades a ajudar l'usuari a trobar articles segons atributs d'aquests. Es poden dividir en aquests subgrups:

— Categories de temes (*topic categories*). Per exemple, la «Categoria: Catalunya» conté els articles relacionats amb el tema *Catalunya*.

— Categories de conjunts (*set categories*): categories que indiquen una classe, normalment en plural. Per exemple, la «Categoria: Automòbils» conté els articles relacionats amb el tema *automòbils*.

2. Categories de projecte o de servei (*project categories*): categories destinades a l'organització interna del projecte i que són utilitzades per editors o per eines automàtiques. En són exemples les categories ocultes, els esborranys d'articles, els articles que necessiten neteja o ampliació, etc.

A més dels enllaços ja mencionats, una pàgina de la Viquipèdia pot contenir altres tipus d'informació i enllaços, com ara:

6. En general, es pot dir que un *graf* és una representació abstracta d'un conjunt d'objectes (o nodes); alguns parells d'aquests objectes estan connectats per arestes. En aquest cas, els objectes són les pàgines i les arestes són enllaços que connecten paraules amb pàgines. Es diu que un graf és dirigit quan les arestes que uneixen els nodes tenen una orientació definida.

— URL externs: adreces web que són potencialment interessants en relació amb l'article que s'està visualitzant (p. ex., «grip» → «canal Salut específic de la Generalitat de Catalunya»).

— InterViqui: article, en una altra llengua, que presumiblement és equivalent a l'article que s'està visualitzant.[7]

— Altres articles de la Viquipèdia que amplien la informació amb temes fortament relacionats amb la pàgina que s'està visualitzant (p. ex., «Galileo Galilei» → «termòmetre de Galileu»).

— Bibliografia de referència. Pot estar indicada en forma de referència bibliogràfica o bé mitjançant un URL.

— Informació potencialment rellevant, sovint fa referència a URL.

— Registre d'autoritat. Indicació de les fonts d'informació utilitzades per a la confecció de l'article.

La figura 1 mostra esquemàticament l'estructura global de la Viquipèdia. La part dreta mostra esquemàticament una pàgina qualsevol i les relacions més rellevants que té associades, mentre que la part esquerra reflecteix la posició de cadascuna de les categories associades a aquesta pàgina i com aquestes es relacionen amb altres categories del graf de categories.

Totes les categories del graf de categories tenen associades almenys una pàgina. En la figura 1, l'article de la pàgina que es detalla en el graf de la dreta té associades altres pàgines («pàgina a», «pàgina b»…) cadascuna de les quals té una estructura semblant a la pàgina que es mostra. És a dir, cadascuna d'elles té associades categories i es relacionen amb altres pàgines. El mateix succeeix amb la resta d'informacions.

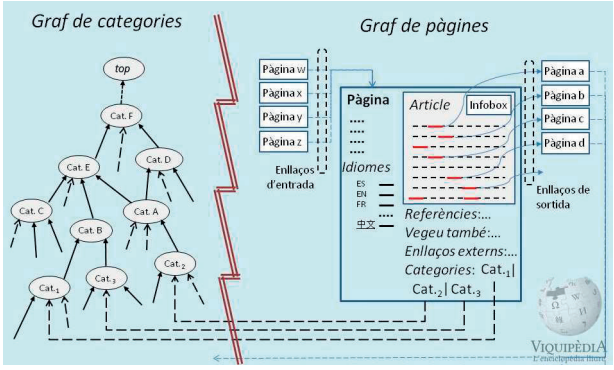


FIGURA 1. Estructura de pàgines i categories de la Viquipèdia.
FONT: Elaboració pròpia.

7. Cal comentar que molts articles són una traducció adaptada i/o retallada de l'article equivalent en llengua anglesa.

A la Viquipèdia existeixen també alguns altres tipus d'informacions de gran utilitat però que no es fan evidents a primer cop d'ull:

1. Pàgines de redirecció (*redirect pages*): informacions emmagatzemades a l'estructura de la Viquipèdia que permeten resoldre casos de:

— Sinonímia. Es redirigeix cap a la pàgina principal. Per exemple: «febres tercianes» → «malària», o «bicicle» → «velocípede».

— Equivalències. Mots que es consideren equivalents. Per exemple: «ronyons» o «sistema renal» → «ronyó».

— Desplegament de sigles no ambigües. Per exemple: «SCATERM → Societat Catalana de Terminologia».

— Correspondència entre un adjectiu relacional i un nom. Per exemple: «hepàtic» → «fetge», o «apical» → «àpex»).

— Correcció d'alguns errors tipogràfics més comuns que pot cometre l'usuari. Un exemple d'aquesta situació és quan l'usuari interroga aquest recurs amb les paraules «ronyo» o «ronyò» i el sistema mostra directament la pàgina amb la grafia correcta de «ronyó».

2. Pàgines de desambiguació (*disambiguation pages*). Es tracta de casos d'homonímia; és a dir, de pàgines amb títols molt semblants o semànticament ambigües. L'exemple clàssic d'aquesta situació és el mot «banc», que en la Viquipèdia incorpora sis lectures. Un altre exemple és l'entrada de «cosa», que inclou, entre altres, una referència a l'objecte, però també a l'organització Cosa Nostra o al personatge Juan de la Cosa. Aquestes pàgines s'utilitzen també per al desplegament de sigles ambigües com ara «IPC», per a la qual s'indiquen els tres significats possibles.

Finalment, si el que se cerca és informació terminològica hi ha una peça d'informació que és molt rellevant: la *infobox* (o infotaula). Aquesta informació normalment es mostra en forma d'una taula que apareix en l'angle superior dret de l'article i serveix per mostrar-ne (en forma de parells atribut-valor) un resum dels aspectes més rellevants. Sovint inclou fotografies, esquemes, etc. La informació acumulada a les *infoboxes* són una peça fonamental d'informació que són captades per la DBpedia i altres usuaris especialitzats.

Una qualitat de les *infoboxes* és que les dades que s'hi inclouen faciliten la comparació entre articles semblants. Per exemple, en medicina totes les malalties tenen informacions comunes, com ara: especialitat que la tracta, símptomes, medicació, part del cos afectada, causa, efectes, codis de classificació (CIM, CIAP), recursos externs que l'analitzen (MeSH, MedicinePlus, SNOMED-CT), etc. D'aquesta manera és fàcil i ràpid trobar informació pròpia d'una malaltia i comparar-la amb la d'altres. Cada branca de la ciència té definides les seves peces d'informació específiques. Les eines de desenvolupament incorporen unes plantilles que faciliten la tasca a l'editor. La figura 2 mostra un exemple de dues malalties i les respectives *infoboxes*.

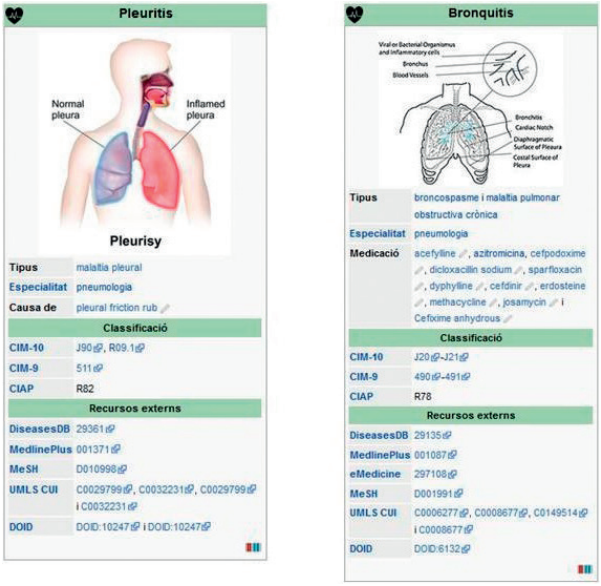


FIGURA 2. Exemples d'*infoboxes* en medicina.
FONT: Viquipèdia.

3. CREDIBILITAT DE LA VIQUIPÈDIA

Des de la seva creació el 2001, s'han dut a terme nombrosos treballs per estudiar i valorar els continguts de la Viquipèdia.[8] La majoria d'ells la comparen amb altres fonts tradicionals subjectes a revisió com ara l'*Encyclopædia Britannica* o *Encarta*. En general, els resultats indiquen que la qualitat dels articles és comparable a la que es troba en les enciclopèdies tradicionals però variable en relació amb recursos especialitzats.

Els editors disposen d'eines que els permeten veure la història de cada article i la corresponent pàgina de discussió. Els usuaris, en canvi, malgrat comptar amb accés a aquesta pàgina, no solen fer-ho per una qüestió de temps i també perquè se'ls pot fer difícil comprovar fins a quin punt la informació d'un article determinat és fiable. Com a conseqüència, l'usuari s'exposa a articles de qualitat variable.

S'ha de tenir en compte que la Fundació Wikimedia garanteix que tots els articles són revisats per experts de totes les branques de la ciència, incloent-hi els articles amb un alt contingut tècnic. Cada idioma té els seus propis mecanismes de control de qualitat. Els millors solen portar una marca de qualitat que es concreta en el fet que pertanyen a una categoria específica: «featured articles», en anglès; «artí-

8. Per a més informació, podeu consultar la pàgina *https://en.wikipedia.org/wiki/Reliability_of_Wikipedia*.

culos destacados», en espanyol, i «articles de qualitat», en català.[9] Hi ha una sèrie de criteris que ha de complir un article per ser considerat d'aquest tipus: que estigui ben redactat, que sigui complet, neutral, etc. Aquesta categoria sol ser oculta, és a dir, no està visible com les altres categories que tenen associades totes les pàgines.

Com ja s'ha comentat, l'origen de la informació de la Viquipèdia no és l'habitual en aquest tipus de recursos; és a dir, la producció per part d'un grup d'experts. Potencialment, la informació pot ésser afegida per qualsevol persona que compleixi unes normes d'edició establertes (i actualitzades a mesura que se'n veu la necessitat). Això pot crear i crea dubtes seriosos entre els experts sobre la fiabilitat de la informació que es troba en aquest recurs.

Per exemple, en [5] s'afirma que els articles sobre malalties cardiovasculars presenten errors per omissió.[10] D'altra banda, en [6] s'ha fet un estudi per tal d'identificar les tendències en l'ús de la WP com una referència en publicacions científiques amb comitè de revisió entre els anys 2002 i 2015. La conclusió és que troben citacions a la WP en revistes d'impacte i en articles produïts per acadèmics d'institucions rellevants.

4. Accessibilitat

La manera d'accedir a la Viquipèdia més coneguda i àmpliament utilitzada és mitjançant un navegador web.[11] Tota la informació que es mostra en una pàgina qualsevol (vegeu l'apartat 2) de la WP està formatada com una pàgina web tradicional.[12] Però la seva utilitat no seria completa si no fos possible accedir-hi també d'una forma automatitzada. Només d'aquesta manera el contingut d'aquest recurs pot ésser utilitzat, per exemple, per les aplicacions típiques del PLN.

L'extracció del text contingut en una pàgina de la WP pot fer-se molt fàcilment amb un buscador web i un *parser*. L'estructura regular de les pàgines permet la utilització d'aquest procediment. Tot i això, aquesta tècnica no sempre és satisfactòria, en particular si es vol augmentar el ventall d'aplicacions.

Afortunadament, hi ha múltiples aplicacions informàtiques que faciliten l'accés a la Viquipèdia. Existeixen llibreries que permeten accedir-hi utilitzant di-

9. En la WP en català hi ha 793 pàgines que es consideren articles de qualitat. Consulteu la pàgina «articles de qualitat» per saber-ne més detalls.

10. Cal destacar que aquestes mancances no fan referència a l'existència de la pàgina mateixa sinó al contingut de l'article. Es troben a faltar figures i taules per clarificar algunes pàgines; també hi ha deficiències pel que fa referència a la fisiopatologia, els mecanismes, l'enfocament diagnòstic i el pla de gestió.

11. Segons la companyia britànica YouGov, la WP és el setè lloc web més popular al Regne Unit. Segons les estadístiques proporcionades per Wikimedia Statistics, en l'últim any, les pàgines en català han tingut 210,78 milions de visualitzacions a tot el món.

12. En realitat, el format utilitzat no és l'HTML com en una pàgina web convencional, sinó un de molt semblant que s'anomena *wiki markup language*.

versos llenguatges de programació (Python, Perl, Java, Javascript, etc.) tant des d'estacions de treball com des de dispositius mòbils.

Malgrat les formes d'accés a la Viquipèdia que acabem d'esmentar, si es necessita un accés repetitiu i àgil, aquests mètodes no són eficients. En [3] s'aborda aquest problema des d'un altre punt de vista. Els autors proposen la creació d'un programari que permet convertir la descàrrega completa (o *dump*) d'aquest recurs al format SQL[13] i carregar-lo posteriorment a una base de dades convencional. D'aquesta manera, el temps d'accés a qualsevol consulta disminueix, fent possible l'ús intensiu d'aquest recurs com el que es necessita en les aplicacions que es descriuen en l'apartat 6. L'avantatge que representa la velocitat d'accés es veu mitigada per la necessitat de reproduir el procés de descàrrega i conversió a base de dades cada vegada que es considera necessari actualitzar el recurs.

Una altra opció és la utilització d'eines com ara el Wikipedia Miner Toolkit [7], un programari de lliure disposició que permet integrar l'accés a la WP en aplicacions pròpies. La idea és compartir algoritmes i codi en lloc de recursos. Inclou, entre altres coses, una API de Java que permet accedir i explorar les categories, pàgines i redireccions de la WP. S'hi inclouen també programaris per al processament dels *dumps*, mesures de similitud entre pàgines, serveis web, etc.

DBpedia [8] és un projecte col·laboratiu per a l'extracció d'informació estructurada i multilingüe de la WP, Viquidata i Viquimèdia Commons per fer-la accessible lliurement en la web utilitzant les tecnologies de la web semàntica i dades enllaçades.[14] La informació s'emmagatzema mitjançant l'estàndard RDF[15] mentre que per a les consultes a la base de dades s'utilitza l'SPARQL.[16] Des de fa pocs anys, aquestes tecnologies permeten, a través de llibreries específiques i en diversos llenguatges de programació, una altra forma d'accés molt ràpid i eficient a la Viquipèdia (juntament amb d'altres recursos que complementen la informació disponible).

5. Utilització de la Viquipèdia

Els diferents mètodes per accedir a la Viquipèdia que s'han mostrat en l'apartat 4 permeten un accés àgil i eficaç per al desenvolupament d'importants projec-

13. El contingut de la WP i altres recursos relacionats estan disponibles en format XML i poden descarregar-se lliurement des de *http://download.wikipedia.org*. Aquest recurs se sol actualitzar almenys un cop al mes.

14. Per al català només existeix una versió en preparació a *http://ca.dbpedia.org*.

15. L'RDF és un marc per a la representació de recursos a la web que s'ha dissenyat per ser utilitzat exclusivament entre ordinadors. Utilitza l'XML i forma part de la W3C's Semantic Web Activity.

16. L'SPARQL és un llenguatge d'interrogació per a grafs RDF i un protocol per accedir a aquests grafs dissenyat pel W3C RDF Data Access Working Group. Un graf RDF és un conjunt de tripletes. Una tripleta consisteix en un subjecte, un predicat i un objecte que conformen un fet complet. Un conjunt de tripletes enllaçades forma un graf de coneixement o graf RDF.

tes com els que es mencionaran en l'apartat 6. De totes maneres, és important tenir en compte les qüestions següents:

— Tipus de graf. El graf de categories no és una taxonomia, encara que és molt convenient i útil considerar-lo com a tal. La denominació mateixa de les relacions entre categories ja dona a entendre que aquest enllaç no sempre indica una relació d'hiponímia/hiponímia. Com ja s'ha mencionat en l'apartat 2, existeixen les categories de servei, que s'utilitzen per gestionar i estructurar el recurs o bé d'altres de caràcter enciclopèdic. Per exemple: agrupar esborranys; articles incomplets o que necessiten revisió; classificació de temes per any, país, regió, etc. Aquest fet no es pot deixar de tenir en compte quan s'explora aquest graf.

— Atribució de categoria a pàgines. Aquesta assignació pot respondre més a criteris enciclopèdics que taxonòmics. Per exemple, certes categories poden ésser assignades per establir informació de tipus (p. ex., «Enrico Fermi» → «físics teòrics»), de nacionalitat (p. ex., «Pau Casals» → «violoncel·listes catalans»), activitat («José de San Martín» → «militars argentins»), etc.

— Circularitat. Ambdós grafs presenten el problema de la formació de cicles. La figura 3 mostra un exemple real on es veu com les categories «ciències socials», «sociologia» i «societat» formen un cicle. La conseqüència és que, si no es prenen les precaucions oportunes, es bloqueja el recorregut d'una part del graf.

— Enllaços entre pàgines. Els enllaços entre alguns mots d'un article i una altra pàgina no tenen cap significat semàntic definit encara que el criteri d'edició sigui que han de ser rellevants per comprendre l'article. Alguns d'ells són superflus, a vegades erronis, o, fins i tot, poden referir-se a pàgines que encara no existeixen.

— Considerem, per exemple, la pàgina de «Galileo Galilei». Hi ha un enllaç amb la pàgina «països de parla catalana» que només serveix al lector per indicar-li en quin lloc geogràfic aquest científic es coneix com a «Galileu». Altres vegades, l'enllaç és manifestament inadequat. Considerem, per exemple, la pàgina d'«objecte»; la primera frase diu: «Un objecte és un ens limitat amb una funció precisa i…». La paraula «funció» té un enllaç cap a la pàgina de «funció» en el sentit utilitzat en matemàtiques però no en el sentit apropiat en aquest article. Aquestes circumstàncies dificulten la utilització d'aquest graf.

— Ubicació de les categories. La posició d'algunes categories pot causar un cert desconcert i crítiques entre els especialistes del domini. Considerem el cas de «medicina» i, dintre d'aquest domini, la categoria «veterinària». En anglès existeix la categoria *veterinary medicine* com a subcategoria de *medicine*, mentre que en català «medicina» i «veterinària» estan al mateix nivell i ambdues són subcategories de «ciències de la salut». Aquesta diferència pot ser polèmica i no és innòcua, ja que en aplicacions terminològiques pot provocar la selecció i/o validació de termes que no interessen per a l'àmbit de la «medicina». Aquest fet ens permet

recordar l'important paper que tenen els editors tant pel que fa a les pàgines com a les categories.

— Desenvolupament asimètric. El mecanisme d'ampliació d'aquest recurs fa que certes àrees del coneixement es desenvolupin més que altres.

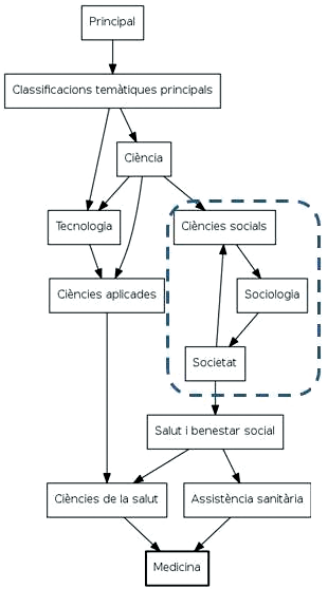


FIGURA 3. Exemple de cicles en l'estructura de categories de la Viquipèdia.
FONT: Elaboració pròpia.

Una qüestió important a tenir en compte és que aquest recurs pot ésser utilitzat també en altres llengües. Un requisit fonamental perquè això sigui possible, en la llengua que s'està considerant, és el grau de compleció de la WP i l'existència de recursos bàsics per al PLN raonablement complets.

6. APLICACIONS DE LA VIQUIPÈDIA

Des de la seva creació i malgrat les qüestions que hem plantejat en els apartats 4 i 5, la WP s'ha utilitzat i s'utilitza sovint per a consultes típicament enciclopèdiques però també ha sigut objecte de molts altres usos menys coneguts. Aquests van des de tasques que li són pròpies (com ara la detecció automàtica d'errors o l'ampliació automàtica del contingut, entre d'altres) fins a tasques relacionades amb el PLN. En els subapartats següents presentem algunes de les aplicacions més rellevants en aquest últim àmbit.

6.1. *Millora o estudi de la Viquipèdia mateixa*

Existeixen molts treballs que es dediquen a estudiar la qualitat dels articles. En [9], per exemple, es pren en consideració un centenar d'atributs lingüístics[17] per estudiar la qualitat de la WP en polonès. Aquest estudi s'ha fet sobre la base de 500.000 articles escollits aleatòriament, i el resultat és que se'n poden considerar de qualitat entre el 4 % i el 5 %.

També s'han desenvolupat programaris específics per a la detecció d'articles inadequats o vandàlics. En [10], es proposa un sistema motivat lingüísticament per a la detecció automàtica d'articles d'aquest tipus. Es fa palès que aquests articles tenen un estil propi i que es poden detectar mitjançant gramàtiques probabilístiques. En [11], es proposa un sistema amb objectius idèntics basat, però, en l'establiment de patrons i utilitzant un sistema d'aprenentatge automàtic.

Finalment, cal destacar els estudis fets dintre del domini mèdic. En primer lloc, en el treball de [12] s'estudia el temps que es triga a actualitzar la referència a articles d'impacte en aquest domini. La conclusió és que transcorre una mitjana de noranta dies a aparèixer citacions en publicacions rellevants com ara: *Cochrane Database of Systematic Reviews*, *Nature* o *The Lancet*, entre altres. En segon lloc, destaquem el treball de [13], en què es descriu el *WikiProject Medicine* i els seus punts forts i febles, i se'l compara amb projectes semblants. Es destaca la importància d'aquest recurs per al públic en general, així com per a estudiants i professionals de la salut. Finalment, destaquem el treball descrit a [14], en què s'utilitza la WP per enriquir un glossari de termes radiològics per a usuaris no especialistes. Del total de 4.090 conceptes presents al glossari, 3.063 (el 74,9 %) han trobat una correspondència a una pàgina de la WP. A més a més, de 800 conceptes escollits aleatòriament, el 51 % s'han enriquit semiautomàticament amb imatges preses de la WP.

6.2. *Tasques pròpies del processament de la llengua*

En una primera aproximació al gran ventall d'aplicacions de la WP, cal observar la freqüència amb què apareix en els articles presentats als congressos que organitza l'Association for Computational Linguistics (ACL).[18] Les actes d'aquests esdeveniments es recullen en una base de dades que actualment conté més de 48.000 articles. En aquest recull es poden fer cerques mitjançant una interfície web[19] en què es pot veure que els articles que d'una manera o altra mencionen la WP es compten per centenars.

17. Per exemple: nom i tipus de nom, adjectiu, verb i tipus de verb, cas, freqüència, gènere, etc.

18. L'ACL és una organització de reconegut prestigi en l'àmbit de la lingüística computacional que organitza regularment un gran nombre de congressos i tallers.

19. *https://www.aclweb.org/anthology*.

Fem, a continuació, un breu recorregut per algunes de les aplicacions més comunes que utilitzen aquest recurs dintre del PLN:

— Creació de corpus monolingües. Un corpus és una col·lecció organitzada de textos i és un requisit bàsic per a qualsevol tasca relacionada amb la lingüística de corpus, el PLN, etc. Un corpus pot ser monolingüe o plurilingüe i, en aquest últim cas, els textos que el formen poden ser paral·lels o comparables. La Viquipèdia és un conjunt de textos ben formats que constitueixen un recurs ideal per a la constitució de corpus textuals en diferents llengües i temàtiques. Com ja s'ha mencionat, els articles estan formatats en una variant del llenguatge HTML. Per tant, l'únic requeriment és l'eliminació de les marques d'aquest tipus. Existeixen força exemples de corpus compilats a partir de la Viquipèdia en moltes llengües, entre les quals el català.[20] També trobem aquest recurs formant part de corpus més grans, com ara Linguatools,[21] Sketch Engine[22] o OPUS,[23] entre altres.

— Traducció automàtica. Com ja s'ha comentat en l'apartat 2, alguns articles en una llengua poden tenir l'article corresponent en una altra llengua. En conseqüència, es poden considerar com a textos comparables i s'utilitzen com un recurs per aquests sistemes.

D'aquesta manera, es creen models per tal d'extreure, per exemple, les frases paral·leles i utilitzar-les per entrenar un sistema d'aquest tipus. Aquests treballs poden referir-se a un àmbit general però també a dominis específics.

En [15] s'extreuen les frases per a les parelles castellà-anglès, alemany-anglès i romanès-anglès i s'entrena un sistema de traducció automàtica estadístic. Un treball similar es descriu a [16]. El treball s'ha fet per a l'anglès i el francès i s'ha restringit a temes relacionats amb els Alps.

— Resum automàtic multidocument (extractiu). Aquestes eines identifiquen, amb l'ajuda de la WP, els conceptes rellevants d'un document i les frases que els contenen. Aquestes són classificades d'acord amb la importància dels conceptes que inclouen. El sistema selecciona les frases més rellevants i les comprimeix per formar el resum [17].

— Creació de taxonomies multilingües. L'objectiu aquí és la integració de diverses edicions de la WP per formar una única taxonomia [18].

— Creació d'una ontologia de domini. En [19] es proposa la creació d'una ontologia de domini en dos passos. En primer lloc es crea automàticament una ontologia utilitzant la Viquipèdia mitjançant l'ús de les relacions implícites en les

20. Disponible a *https://repositori.upf.edu/handle/10230/20050*. Aquest corpus està format per un total de 390.000 articles amb 125,6 M de paraules i ha sigut processat amb les eines de l'Institut de Lingüística Aplicada de la Universitat Pompeu Fabra (IULA).

21. *https://linguatools.org/tools/corpora/*.

22. *https://www.sketchengine.eu/*.

23. *http://opus.nlpl.eu/*.

plantilles de les pàgines, les categories i les *infoboxes*. Finalment, es proposa una manera d'aprofitar la informació inicial i un cercador per completar l'ontologia.

— Creació de bases de coneixement: la WP és la font principal d'informació per a la creació dels recursos següents: DBpedia,[24] Freebase,[25] WikiNet[26] i YAGO.[27]

— Enllaç d'entitats (*entity linking* o *Wikify*). Aquesta tasca consisteix a enriquir algunes paraules d'un text amb enllaços a pàgines de la Viquipèdia. Es pot considerar com un tipus d'etiquetatge semàntic en què l'etiquetari és el conjunt de pàgines de la WP. Per tant, s'han d'identificar els termes rellevants del text i anotar-los de manera no ambigua amb la pàgina pertinent. En relació amb aquesta tasca destaquem les propostes presentades en [20], [21] i [22].

— Anàlisi semàntica. ESA (*explicit semantic analysis*) és una representació vectorial de text (paraules o documents) que utilitza un corpus com una base de coneixement. Aquest treball utilitza cada article de la WP com una unitat de coneixement; es va iniciar amb [23] i [24], i ha sigut objecte de múltiples millores i adaptacions. El camp d'aplicació d'aquesta tècnica és el càlcul de la similitud entre paraules, frases o documents.

— Càlcul de la relació semàntica entre paraules o textos. Es tracta d'utilitzar el coneixement implícit als enllaços entre pàgines, així com entre pàgines i categories. Vegeu l'exemple descrit a [25].

— Reconeixement i classificació d'entitats en diversos idiomes. En [26] es proposa utilitzar les anotacions de la WP com les anotacions inicials necessàries en tot sistema d'aprenentatge automàtic. Per aconseguir aquest objectiu parteix de les anotacions en una llengua i, utilitzant l'estructura bigraf de la WP, es troben les anotacions en la llengua destí. L'anotació aconseguida és suficient per poder ampliar-la aplicant mecanismes d'aprenentatge automàtic.

6.3. *Terminologia*

La WP és un recurs que fa explícit un ampli ventall de conceptes en forma de títols de pàgines i nom de categories. Podem concebre dues maneres d'explotació de tota aquesta informació per al treball terminològic segons es faci una exploració de dalt a baix (*top-down*) o bé de baix a dalt (*bottom-up*) en el graf de categories. En aquest apartat posarem dos exemples reals d'explotació utilitzant ambdós mecanismes d'exploració.

24. *https://wiki.dbpedia.org.*
25. Aquest recurs és inactiu des de 2016, quan va ésser integrat a Viquidata.
26. *https://www.h-its.org/software/wikinet-2/.*
27. *https://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago/.*

La idea subjacent en els projectes que es descriuran és que es poden establir fronteres de domini en el graf de categories. És a dir, trobar categories per sota de les quals podem considerar, amb un cert grau de confiança, que totes les categories i les pàgines associades són pertinents en un cert domini.

Una vegada establertes aquestes fronteres, segons el tipus d'exploració que es faci podrem:

1. obtenir tots els termes d'un domini que estan registrats en aquest recurs (exploració *top-down*) o bé
2. determinar si un mot trobat a la Viquipèdia (com a pàgina o categoria) és o no pertinent en el domini d'interès (exploració *bottom-up*).

En una primera aproximació, l'establiment d'aquestes fronteres és molt simple: el nom del domini (o la categoria de la WP més propera) és la frontera cercada. Aquesta solució, però, no sempre és suficient; pot ésser necessari un treball d'exploració per afegir altres categories que la complementin. Un exemple d'aquesta situació es dona quan volem establir les fronteres per a la medicina. En una aproximació podríem escollir la categoria «medicina». Però una anàlisi del graf de categories ens mostra que les seves supercategories són: «ciències de la salut» i «assistència sanitària». Una anàlisi d'aquestes dues categories ens podria portar a acceptar la primera i a descartar la segona. En el primer cas s'inclouen subcategories com ara «infermeria», «nutrició» i «odontologia», entre d'altres. Mentre que la categoria «assistència sanitària» inclou subcategories com ara «metges asiàtics», «sanitat per país» i «organitzacions sanitàries», entre d'altres, la gran majoria de les quals no són d'interès terminològic. No obstant això, seria convenient una revisió acurada d'aquestes categories per si es consideren rellevants per a la tasca que ens proposem dur a terme. En qualsevol cas, l'exploració ha de prendre les precaucions oportunes per evitar considerar aquestes categories (i probablement les pàgines que en depenen) com a terminològiques.

Un problema comú a tot treball terminològic, ja sigui de compilació o d'extracció, és l'avaluació del material obtingut. No existeix una solució per a aquest tema, ja que, d'una banda, l'avaluació manual per experts és inviable pel seu cost i, de l'altra, si hi intervé més d'un expert, sorgeix el problema de la diferència de criteris entre ells. Per a una avaluació automàtica fora necessari tenir una llista de referència. En el cas de l'extracció, aquestes llistes, si existeixen, difícilment són completes. En el cas de la compilació és el producte que volem obtenir. Per tant, cal recórrer a avaluacions parcials o indirectes.

En els dos subapartats següents analitzarem amb més detall les possibilitats d'ambdós mecanismes i els resultats que s'han obtingut.

6.3.1. Recull de la terminologia d'un àmbit

La compilació de la terminologia d'un àmbit és una tasca necessària per a moltes aplicacions de PLN. L'adaptació al domini dels recursos existents per processos com ara l'etiquetatge semàntic, l'extracció de relacions, l'etiquetatge de rols semàntic, el resum automàtic, la traducció automàtica i altres depenen en gran mesura del fet de disposar d'aquestes terminologies.

L'obtenció de la terminologia d'un domini és una tasca complexa. Es presenten dos problemes: en primer lloc, s'ha de disposar d'un corpus de textos ben construït i, en segon lloc, cal extreure els termes d'aquest corpus.

La compilació d'un corpus d'un domini qualsevol és una tasca costosa en temps i recursos. L'obtenció dels termes del corpus ja compilat també és problemàtica. El processament manual és una tasca inassolible, per la qual cosa s'ha de recórrer a mètodes automàtics. Existeixen diverses solucions per a aquest últim problema (en aquest mateix subapartat se'n presenta una) amb resultats variables. Una altra solució seria disposar d'un recull ja compilat amb els termes del domini d'interès. D'aquesta manera, el problema es reduiria a cercar aquests termes en el text que s'està processant.[28]

En [27] i [28] es mostren dues aproximacions per recollir automàticament els termes d'un domini. Ambdues propostes utilitzen com a punt de partida la Viquipèdia, ja sigui directament (en el primer cas) o indirectament, a través de DBpedia (en el segon). L'objectiu en la primera aproximació és obtenir tots els termes d'alguns dominis en castellà i anglès, mentre que, en la segona, l'objectiu és semblant però estenent la tasca a una col·lecció de dominis. També s'afegeix la possibilitat de trobar la correspondència entre els termes de les dues llengües. En la resta d'aquest apartat descriurem breument el procediment proposat a [27].

L'objectiu d'aquesta proposta és recopilar de les fonts lèxiques de cada domini i per a les dues llengües tants termes com sigui possible i cercar la correspondència entre ells. Per aconseguir aquest objectiu es disposa de: *a*) un conjunt d'etiquetes de domini, *b*) un parell de llengües i *c*) recursos de lèxics que cobreixen tots els dominis en les dues llengües.[29]

El sistema proposat utilitza aquests recursos lèxics:

1. Multilingual Central Repository (MCR).[30] Aquest recurs segueix el model proposat pel projecte EWN [29] per crear una base de dades lèxica multilingüe

28. Òbviament, es podria dir que difícilment un recull de termes ja compilat tindria tots els termes del domini. De totes maneres, també es pot dir que segons l'aplicació de què es tracti aquest podria ésser suficient.

29. D'aquí es pot desprendre una limitació del sistema: només recollirà els termes de domini que estiguin presents a la WP.

30. *http://adimen.si.ehu.es/web/mcr/*.

amb els *wordnets* per a diverses llengües europees. Els *wordnets* de cadascuna de les llengües s'estructuren com el WordNet (WN) [30] de Princeton.

2. Extended WordNet Domains (XWND). Aquest recurs neix a partir de Wordnet Domains (WND) [31] i fou ampliat en [32]. L'objectiu és assignar informació de domini als *synsets* de WN. Aquesta informació té la forma d'una taxonomia amb 164 dominis. XWND és un recurs lèxic en què els *synsets* de WN incorporen informació de domini com a WND, però aquest recurs assigna a cada *synset* una probabilitat de pertànyer a cada domini del conjunt de dominis.

3. Wikipedia (WP). Projecte objecte d'aquest treball. Vegeu-ne els detalls a l'apartat 2.

4. DBpedia. Projecte que extreu informació estructurada i multilingüe de la Viquipèdia per fer-la accessible lliurement en la web utilitzant les tecnologies de la web semàntica i dades enllaçades.

El procés global està esquematitzat en la figura 4, s'aplica iterativament a cada parella domini-llengua la seqüència de processament descrita a continuació:

— Pas 1. Per utilitzar XWND com a etiquetari semàntic és necessari un procés de normalització per tal que les assignacions de probabilitats de domini a cada *synset* de l'MCR siguin comparables.

— Pas 2. Per a cada domini i llengua s'utilitza l'MCR i XWND per identificar tots els *synsets* que tinguin una alta probabilitat de pertànyer al domini. Es busquen les variants d'aquests *synsets* en la WP. El conjunt d'aquestes categories és avaluat utilitzant la mateixa WP (vegeu el subapartat 6.3.2) per eliminar aquelles categories que es troben per sota d'un cert llindar.

— Pas 3. S'obtenen totes les categories principals de cada domini.

— Pas 4. Es filtren les categories principals utilitzant XWND, les supercategories i la distància al *top*.

— Pas 5. Les categories principals seleccionades s'expandeixen utilitzant els enllaços del tipus «subcategoria».

— Pas 6. S'aplica un conjunt de mesures específiques que filtren les categories obtingudes en el pas anterior.

— Pas 7. S'obtenen el conjunt de pàgines enllaçades amb el conjunt de categories.

— Pas 8. Es posa en marxa un procés iteratiu en què, a cada cicle, el conjunt de pàgines i categories es reforça o restringeix mútuament. El procés continua fins que no hi ha cap variació d'un cicle al següent. El resultat és el conjunt de pàgines i categories definitiu.

— Pas 9. S'apliquen una sèrie de filtres per millorar la qualitat dels resultats. El primer consisteix a aplicar un algoritme *page-rank* sobre una representació en forma de graf dirigit del conjunt de categories i pàgines per a cada domini-llengua. Els termes menys fiables són eliminats. A continuació es resol el problema

dels candidats que pertanyen a més d'un domini. També s'eliminen els casos de pàgina i categoria que es refereixen al mateix terme. Quan un terme apareix en singular i en plural s'elimina l'última forma.

— Pas 10. S'utilitza la DBpedia per obtenir, on sigui possible, els termes equivalents en l'altre idioma.

— Pas 11. Es procedeix a l'avaluació del resultat per a cada domini i llengua i a la creació dels fitxers en format OLIF.

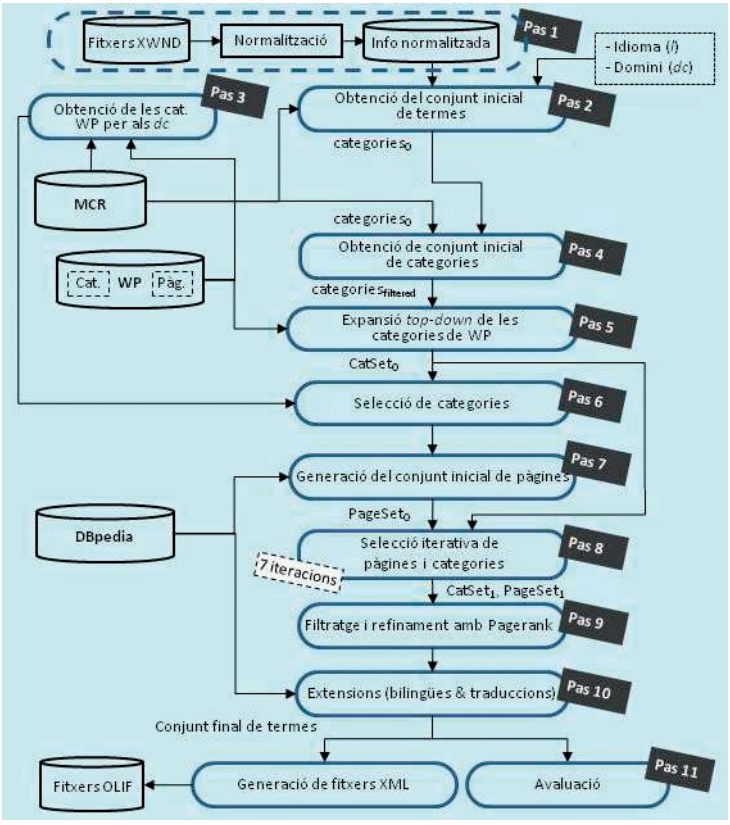


FIGURA 4. Esquema global per obtenir termes de tots els dominis inclosos en XWND.
FONT: Elaboració pròpia.

L'aplicació de la metodologia ja descrita ha permès obtenir 635.527 termes per als 164 dominis i les dues llengües. En la taula 1 es mostren els deu dominis en els quals s'han capturat més termes. L'última fila i l'última columna mostren les xifres globals. Per a cada domini es mostra el nombre de termes corresponent a

pàgines i categoria. S'inclou també el nombre de termes per als quals s'ha trobat una traducció.

Si es comparen aquests resultats amb el nombre de pàgines de la WP per a les llengües de treball,[31] els resultats quantitatius poden semblar escassos. S'ha de considerar, però, el caràcter enciclopèdic de la WP, fet que motiva la inclusió de moltes pàgines i categories que s'han de descartar perquè no es poden considerar terminològiques.

Taula 1. *Domini més freqüent per als termes obtinguts*

Domini	*Nombre de pàgines EN*	*Nombre de cat. EN*	*Nombre de pàgines ES*	*Nombre de cat. ES*	*Correspondència*	*Total*
social	41.710	2.230	6.206	808	6.583	50.954
free_time	26.819	461	1.223	136	716	28.639
animals	16.281	636	6.936	206	4.661	24.059
person	17.163	589	5.502	293	3.100	23.547
biology	13.847	754	4.318	339	3.036	19.258
medicine	13.353	852	4.227	423	3.473	18.855
plants	5.366	271	10.436	1.428	472	17.501
environment	14.124	901	2.105	235	1.996	17.365
sociology	13.715	1.315	1.874	452	1.738	17.356
industry	13.774	229	2.165	215	1.020	6.383
…						
Total	444.653	28.032	146.140	16.702	79.946	635.527

Font: Elaboració pròpia.

La col·lecció completa de tots els termes per a tots els dominis definits a WND es guarda en fitxers[32] amb el format definit per l'estàndard OLIF.[33] A tall d'exemple, la figura 5 mostra el contingut per al concepte «med_4434», que en

31. Segons les dades obtingudes a *https://en.wikipedia.org/wiki/List_of_Wikipedias* hi ha 5.845.347 pàgines en anglès i 1.516.327 pàgines en castellà.

32. La col·lecció completa serà disponible per descàrrega lliure.

33. *Open lexicon interchange format*. Es tracta d'un format lliure per a l'intercanvi d'informació lèxica i terminològica; vegeu-ne més informació a *http://www.olif.net/*.

anglès correspon a *assisted reproductive technology* i en castellà a *reproducción asistida*. Els dos termes es *mapegen* mútuament i, en conseqüència, comparteixen identificador. L'entrada inclou també altres informacions, com ara la categoria morfosintàctica, l'origen, el coeficient de fiabilitat, etc.

```
<entry ConceptUserId="med_4434">
 <mono>
  <keyDC>
   <canForm>assisted reproductive
     technology</canForm>
   <language>en</language>
   <ptOfSpeech>noun</ptOfSpeech>
  </keyDC>
 </mono>
 <monoDC>
  <monoAdmin>
   <confidence>0.801</confidence>
   <entrySource>WP page</entrySource>
  </monoAdmin>
 </monoDC>
</entry>
```

```
<entry ConceptUserId="med_4434">
 <mono>
  <keyDC>
   <canForm>reproducción asistida</canForm>
   <language>es</language>
   <ptOfSpeech>noun</ptOfSpeech>
  </keyDC>
 </mono>
 <monoDC>
  <monoAdmin>
   <confidence>0.800</confidence>
   <entrySource>WP page</entrySource>
  </monoAdmin>
 </monoDC>
</entry>
```

FIGURA 5. Exemple d'un terme en castellà i el seu equivalent en anglès en format OLIF.
FONT: Elaboració pròpia.

Pel que fa a l'avaluació dels termes obtinguts i per tal de minimitzar els problemes ja mencionats, s'han identificat escenaris diversos:

1. Avaluació parcial d'acord amb termes que apareixen tant a l'MCR com a la WP. Degut a la manca de fonts fiables de termes validats per la majoria dels dominis, el primer escenari consisteix a fer una avaluació restringida dels termes que apareixen tant a l'MCR com a la WP. En aquest cas, es donen per certes les assignacions fetes per XWND. Aquesta avaluació pot fer-se per a qualsevol parella domini-llengua.

En la figura 6 es mostren gràficament els termes trobats a l'MCR i la WP i es defineixen diferents grups de termes segons la seva pertinença. Aquesta avaluació es basa en el conjunt *C* perquè és el que agrupa els termes presents tant a l'MCR com a la WP i que pertanyen al domini. En la mateixa figura es mostren les fórmules de càlcul de precisió i cobertura.

2. Avaluació mitjançant fonts de domini externes. Es fa una avaluació completa per als dominis on existeixi una font fiable de referència. Aquest és el cas de medicina, que s'avalua mitjançant SNOMED-CT [33].

3. Avaluació mitjançant fonts externes. Per a l'idioma anglès i només per als termes inclosos en la Viquipèdia, es fa una comparació amb les assignacions fetes pel sistema de Niemann-Gurewych i descrit a [34] (NG).[34]

34. En [33] es proposa una alineació entre els sentits de WN anglès i les pàgines de la WP per a l'anglès.

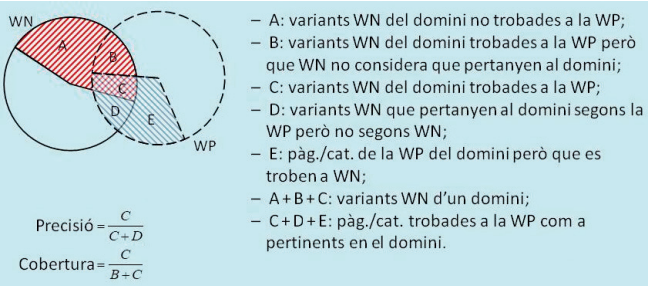


FIGURA 6. Definició dels conjunts de termes per a l'avaluació mitjançant l'MCR i la WP.
FONT: Elaboració pròpia.

4. Avaluació indirecta. L'últim escenari és indirecte i consisteix en l'ús del text dels articles de la WP associat als termes d'un domini per aprendre *word embeddings*[35] per a aquest domini i a continuació avaluar aquests *embeddings*.

L'avaluació es va limitar als dominis següents: agricultura, antropologia, arquitectura, medicina, música i turisme.

A continuació es mostren breument els resultats obtinguts amb els dos primers escenaris d'avaluació.

1. Avaluació parcial d'acord amb els termes que apareixen tant a WN com a la WP. Utilitzant el conjunt de termes segons s'ha definit en la figura 6, s'han calculat els valors de precisió i cobertura que es mostren en la taula 2. Per a cada idioma i domini es mostren el nombre inicial de termes a WN i els valors corresponents de precisió i cobertura.

2. Avaluació mitjançant fonts de domini externes. La utilització de SNOMED-CT permet, en principi, una avaluació millor dels termes de medicina, ja que es tracta d'una font reconeguda i molt utilitzada. De totes maneres, s'ha de tenir en compte que aquest recurs es defineix[36] com un «vocabulari de terminologia clínica utilitzat per professionals de la medicina per a l'intercanvi electrònic d'informació de salut». Per tant, no inclou tota la terminologia del domini sinó la que s'utilitza normalment en la pràctica diària. Aquest fet pot causar alguna indicació d'error falsa degut a:

a) Termes especialitzats. Algunes entrades es refereixen només a termes especialitzats. Vegeu, per exemple, el terme castellà *glándula*, que només existeix com a part d'un terme més específic com ara *glándula esofágica* o *glándula lagrimal*.

35. *Word embedding* és una tècnica molt utilitzada per a la representació del vocabulari d'un document. És capaç de capturar el context d'una paraula en un document, les similituds sintàctica i semàntica, la relació amb altres paraules, etc.

36. *https://searchhealthit.techtarget.com/definition/SNOMED-CT*.

TAULA 2. *Resultat de l'avaluació bàsica per a tots els dominis escollits*

Domini		*Turisme*		*Arquitectura*		*Música*	
Llengua		EN	ES	EN	ES	EN	ES
Termes a WN	Total (A+B+C)	556	180	219	30	1.121	234
	A la WP (C)	6	10	7	4	18	144
Precisió (%)		0,86	0,59	0,80	0,50	1,00	0,55
Cobertura (%)		1,00	1,00	0,89	1,00	1,00	1,00
Termes nous (D+E)		3.466	311	7.928	1.486	209	1.373
Total de termes nous al domini (C+D+E)		3.472	321	7.935	1.490	227	1.517

Domini		*Agricultura*		*Antropologia*		*Medicina*	
Llengua		EN	ES	EN	ES	EN	ES
Termes a WN	Total (A+B+C)	394	94	417	64	3.468	512
	A la WP (C)	4	6	11	15	499	196
Precisió (%)		0,67	0,55	0,79	0,54	0,78	0,55
Cobertura (%)		0,80	1,00	0,85	1,00	0,98	1,00
Termes nous (D+E)		394	510	2.294	1.040	13.282	2.523
Total de termes nous al domini (C+D+E)		398	516	2.395	1.055	13.881	2.719

NOTA: Els valors s'han d'interpretar d'acord amb la figura 6.
FONT: Elaboració pròpia.

b) Termes absents. Alguns termes relativament comuns estan presents a la WP però no en aquest recurs. Aquest fet és una de les causes de la baixa precisió mostrada en la taula 3.

c) Existència de termes complexos. S'inclouen com a termes simples alguns que en realitat són coordinats (p. ex., *enfermedades hereditarias y degenerativas del sistema nervioso central*).

En la taula 4 s'inclouen alguns exemples de termes validats i no validats, correctament i incorrectament.

TAULA 3. *Resultat de l'avaluació dels termes de l'àmbit de la medicina mitjançant SNOMED-CT*

	Llengua	
	EN	*ES*
Total de termes	13.382	2.523
Termes trobats	4.195	994
Precisió (%)	31,3	39,4

FONT: Elaboració pròpia.

TAULA 4. *Exemples de termes validats i no validats utilitzant SNOMED-CT*

	Llengua	
	Anglès	*Castellà*
Vàlid/*true*	*fibrosarcoma* *Kirschner wire* *pneumaturia*	*bronquiolitis* *duodeno* *placa dental*
Vàlid/*false*	*Alzheimer Research Forum* *Birmingham Accident Hospital*	*especialidades médicas*
No vàlid/*true*	*Rivalta test*	*otitis externa* *circulación portal hepática* *ortesis*
No vàlid/*false*	*high-throughput screening* *Kawasaki Medical School* *Goldwater rule*	*Condenado a vivir* *Asociación Pablo Ugarte* *huérfano del sida*

FONT: Elaboració pròpia.

Els resultats dels altres dos escenaris d'avaluació confirmen la validesa de la proposta que acabem de presentar i es poden consultar en [28].

6.3.2. Extracció de terminologia

La identificació dels termes presents en un text representa un coll d'ampolla per a la mineria de textos i, en conseqüència, és un tema de recerca important en l'àmbit del PLN. Podem veure l'extracció de termes com una tasca de marcatge semàntic per tal d'afegir al text informació sobre el significat. La manera d'abor-

dar aquesta tasca depèn dels recursos disponibles, principalment ontologies i llistes de termes. Si no es disposa d'aquesta informació cal recórrer a fonts d'informació indirecta de tipus lingüístic i/o estadístic. Els resultats que s'obtenen amb aquests mecanismes són limitats i per això aquestes eines tendeixen a afavorir la cobertura sobre la precisió. La conseqüència és que molts extractors obtenen llargues llistes de candidats que cal verificar manualment. Una de les raons d'aquest comportament és la manca d'informació semàntica. Les poques eines que utilitzen aquest tipus d'informació funcionen per a l'anglès. YATE [35] en constitueix una de les poques excepcions, ja que utilitza l'EWN com a font d'informació semàntica. Una altra possibilitat és utilitzar la WP com a font de coneixement. La WP representa una alternativa vàlida, i la utilització d'aquest recurs en aquest context és l'objectiu d'aquest apartat.

L'eina YATE està formada per diversos mòduls, un dels quals té com a funció determinar la terminologicitat d'un candidat utilitzant l'EWN. L'experiment que descriurem a continuació consisteix a construir un mòdul que tingui la mateixa funció utilitzant, però, la WP com a font de coneixement. L'àmbit escollit per a aquestes proves és la medicina, ja que disposa d'un cert nombre de fonts de coneixement que permeten superar les barreres comentades en el paràgraf anterior.[37] Per desenvolupar aquesta tasca s'haurà d'utilitzar l'estructura bigraf de la WP. A continuació veurem com utilitzar aquest recurs per calcular la terminologicitat d'un candidat.

Com ja s'ha comentat a l'inici d'aquest apartat, l'exploració dels grafs de la WP serà, en aquest cas, de baix cap a dalt; o sigui, a la inversa de l'exposada en el subapartat anterior. A partir del mot a valorar es procedeix a recórrer cap al *top* l'estructura bigraf fins a trobar una categoria que pugui considerar-se frontera de domini (FD) o bé arribar al *top*.

La figura 7 presenta de manera simplificada un fragment de l'estructura bigraf de la WP per tal de mostrar l'anàlisi del candidat a terme (CAT) *sang*. Aquesta figura mostra la situació que es presenta quan es vol analitzar el candidat *sang*. A la WP existeix la pàgina «sang», que té associada la categoria «sang». També es mostra la categoria «medicina», que és considerada com a frontera de domini. Podem imaginar una exploració del graf des de la pàgina amb el títol «sang» cap al *top*. És fàcil veure que existeixen múltiples camins possibles; alguns passen per la frontera de domini i altres no. Una possibilitat per valorar la terminologicitat o coeficient de domini (CD) del candidat *sang* seria considerar la relació entre aquest nombre de camins i que es concreta en l'equació, en què $NC_{frontera}(t)$ representa el nombre de camins al *top* que passen per la frontera i $NC_{total}(t)$, el total de camins al *top*.

37. Lamentablement, la majoria d'aquestes fonts són per a l'anglès, però n'hi ha d'altres (EWN) que es van poder ampliar en aquest àmbit en castellà i català.

$$CD_{nc}(t) = \frac{NC_{frontera}(t)}{NC_{total}(t)} \qquad \text{Equació 1}$$

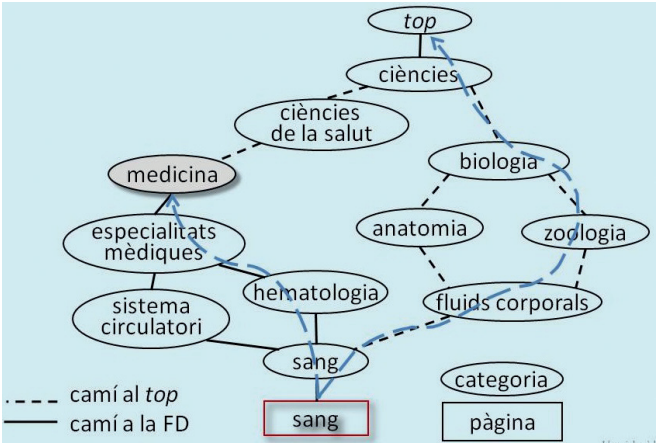


FIGURA 7. Exploració de l'estructura bigraf de la Viquipèdia per tal de valorar el candidat a terme *sang*.
FONT: Elaboració pròpia.

Utilitzant la mateixa figura 7, es poden identificar altres maneres de calcular la terminologicitat d'un candidat a terme. En les equacions 2 i 3 es mostren dues possibilitats més de càlcul. La primera es basa en la longitud dels camins (és a dir, el nombre de salts pàgina-categoria o categoria-categoria) i la segona, en la longitud mitjana dels camins.

$$CD_{lc}(t) = \frac{LC_{frontera}(t)}{LC_{total}(t)} \qquad \text{Equació 2}$$

en què $LC_{frontera}(t) =$ longitud dels camins a la frontera de domini
$LC_{total}(t) =$ longitud dels camins al *top*

$$CD_{lmc}(t) = \frac{LMC_{frontera}(t)}{LMC_{total}(t)} \qquad \text{Equació 3}$$

en què $LMC_{frontera}(t) =$ longitud mitjana dels camins a la frontera de domini
$LMC_{total}(t) =$ longitud mitjana dels camins al *top*

El resultat de l'aplicació d'aquests coeficients en l'avaluació d'un candidat a terme segons la informació de què disposa la WP pot ésser dividida en quatre grups:

1. CD..(t) = 1 → El candidat és un terme del domini;
2. 1 $>$ CD..(t) $>$ 0 → El candidat pot ésser utilitzat en diversos dominis. Normalment, com més gran és aquest valor més forta és la seva relació amb el domini;
3. CD..(t) = 0 → El candidat no pertany al domini;
4. CD..(t) = –1 → El candidat no està registrat a la WP i, en conseqüència, no se'n pot fer cap valoració.

A continuació es mostren els resultats que s'obtindrien si apliquéssim aquestes mesures al cas del candidat a terme *sang* utilitzant la figura 7:

$$CD_{nc}(t) = \frac{NC_{frontera}(t)}{NC_{total}(t)} = \frac{2}{2+2} = 0,5$$

$$CD_{lc}(t) = \frac{LC_{frontera}(t)}{LC_{total}(t)} = \frac{4+4}{6+6} = 0,66$$

$$CD_{lmc}(t) = \frac{LMC_{frontera}(t)}{LMC_{total}(t)} = \frac{4}{6} = 0,66$$

Aquest mecanisme té una limitació: només es poden valorar els termes que estan registrats com a pàgina o com a categoria. Una manera de superar aquesta barrera és utilitzar les pàgines de redirecció; en particular, les que fan referència a la redirecció dels adjectius relacionals.

Considerem, per exemple, el candidat a terme *maduració pulmonar*; a la WP no hi és i, per tant, no es pot reconèixer directament. Però si considerem que, aïlladament, ambdós mots es poden reconèixer com a termes en medicina podem també reconèixer el candidat sencer. En efecte, *maduració* té un coeficient de domini més gran que zero en medicina, mentre que *pulmonar* és un adjectiu relacional que la WP remet a *pulmó*, que també és reconegut com a terme. En conseqüència, assignem al terme complet un valor de terminologicitat que és una combinació del que obtenen cadascun dels mots separadament.

En [36] es presenta l'experiment que anticipàvem a l'inici d'aquest apartat i que permet comparar el comportament de YATE utilitzant l'EWN o bé la WP. A continuació, es presenta la metodologia emprada, els resultats obtinguts i la seva validació.

En la figura 8 es mostra l'esquema utilitzat per fer aquesta comparació. En aquest esquema es veu com el resultat del mòdul d'extracció de CAT de YATE s'utilitza per alimentar els dos mòduls d'anàlisi. A continuació, el resultat d'ambdós mòduls es compara amb el resultat de l'avaluació dels candidats del mateix text realitzada per especialistes del domini mèdic.

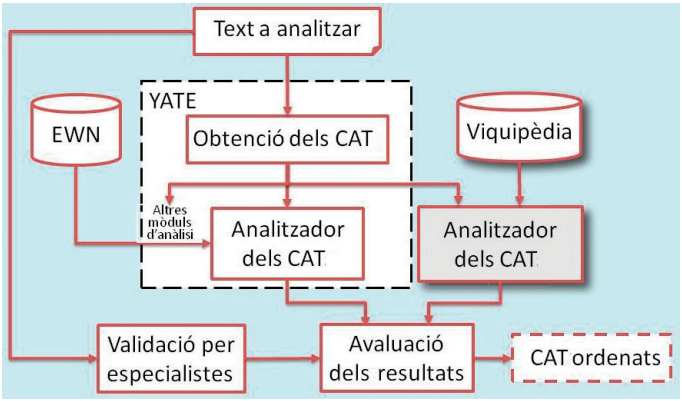


FIGURA 8. Esquema utilitzat per avaluar el mòdul de validació dels candidats a terme amb la Viquipèdia.
FONT: Elaboració pròpia.

Per realitzar l'estudi proposat s'han utilitzat textos del Corpus Tècnic de l'IULA [36] per a un total de 100 K paraules, aproximadament. Els textos van ser processats lingüísticament com és usual en PLN. L'avaluació s'ha efectuat amb les mesures de precisió i cobertura. Els candidats extrets per YATE s'han analitzat en primer lloc amb el mòdul de YATE (és a dir, utilitzant l'EWN com a font d'informació semàntica) i a continuació amb el mòdul que acabem de descriure (amb la WP i les tres mesures ja proposades). Cal mencionar que aquest extractor de termes només extreu els candidats que segueixen els patrons següents: *nom*, *nom-adjectiu* i *nom-preposició-nom*.[38] Els resultats obtinguts per a cada patró individualment, en termes de precisió i cobertura, es mostren en la figura 9, la figura 10 i la figura 11, mentre que en la figura 12 es mostren els resultats globals sense fer diferències entre patrons.

38. Es considera que aquests patrons cobreixen la gran majoria de termes, almenys en el domini de la medicina.

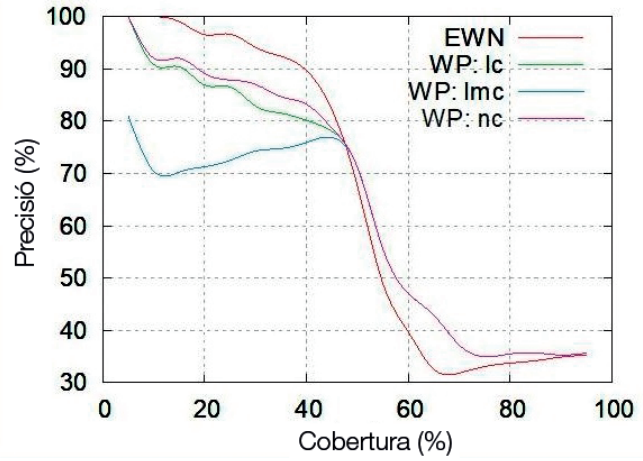


FIGURA 9. Avaluació de candidats a terme monolèxics nominals.
FONT: Elaboració pròpia.

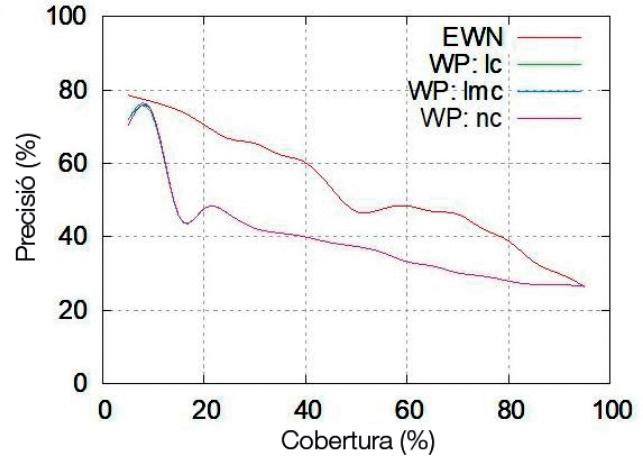


FIGURA 10. Avaluació de candidats a terme amb el patró *nom-adjectiu*.
FONT: Elaboració pròpia.

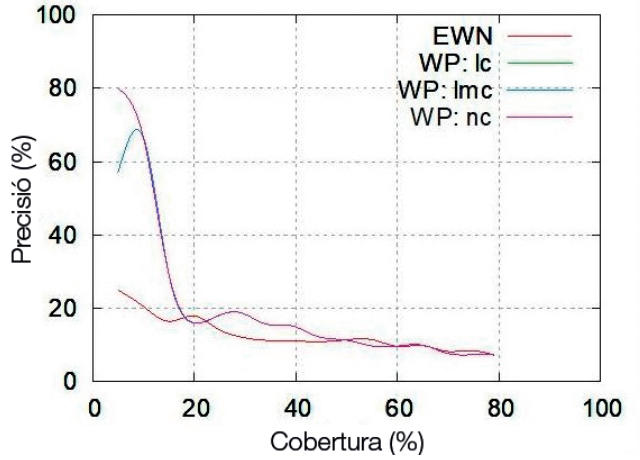


FIGURA 11. Avaluació de candidats a terme amb el patró *nom-preposició-nom*.
FONT: Elaboració pròpia.

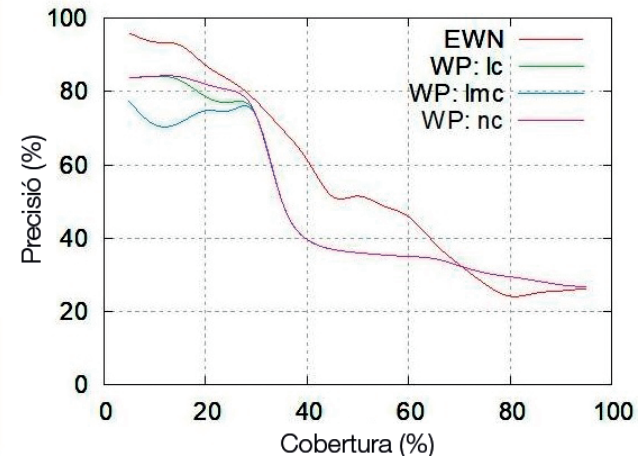


FIGURA 12. Avaluació de tots els candidats a terme conjuntament.
FONT: Elaboració pròpia.

Com es pot veure en la figura 9, la figura 10, la figura 11 i la figura 12, els resultats obtinguts utilitzant l'EWN són lleugerament superiors. Es considera que aquest comportament és, en part, degut al fet que la versió de l'EWN utilitzada està especialment adaptada per realitzar la tasca d'extracció de termes en medicina. De totes maneres, la diferència no és tan alta com podria esperar-se. Analitzem, a continuació, els resultats obtinguts per a cada patró individualment:

— Patró N: la diferència en els resultats obtinguts amb l'EWN i la WP varia entre un 10 % (CD_{nc}) i un 25 % (CD_{lmc}). Malgrat que aquesta diferència és important, cal mencionar que el coeficient CD_{nc} posiciona molt bé els CAT que estan inclosos a l'EWN. Cal mencionar també que hi ha CAT que existeixen a l'EWN però no a la WP.

— Patró NJ: en aquest cas el comportament de tots els CAT és molt similar i la diferència se situa al voltant del 25 %. Termes com ara *historia clínica* o *signo clínico* es classifiquen millor que amb l'EWN. Alguns termes són detectats amb la WP però no amb l'EWN (*poliarteritis nodosa*) i viceversa (*infección viral*).

— Patró NPN: en aquest cas tots els coeficients basats en la WP tenen un comportament millor que utilitzant l'EWN. La diferència és de prop del 10 % i la raó s'ha de buscar en el fet que l'EWN disposa de poques entrades per a aquest patró i l'estratègia de detecció no és gaire eficient. Al mateix temps, la WP incorpora moltes unitats amb aquest patró com, per exemple, *protocolo de tratamiento*, *grupo de riesgo* o *índice de mortalidad*, que reben la màxima qualificació amb la WP. En canvi, la qualificació d'aquests termes és molt reduïda amb l'EWN degut al fet que l'entrada completa no hi és i cada paraula per separat no té un significat especial en medicina. Cal mencionar també que, en el text analitzat, hi ha 910 candidats però només n'hi ha 39 a la WP i només 14 tenen un coeficient més gran que zero.

— Tots els patrons: en aquest comportament global la diferència es redueix un 5 % pel que fa a la precisió i un 30 % a la cobertura.

— Cal prestar una atenció especial al fet que, com ja s'ha comentat, la selecció de termes feta per especialistes és problemàtica. Com a exemple, podem dir que mots com ara *epitelio* o *medicina interna* s'han detectat correctament per ambdós sistemes però no s'han considerat com a termes pels especialistes; en conseqüència, el sistema d'avaluació els ha considerat errors.

En resum, podem considerar que l'extracció de termes en textos de biomedicina mitjançant la Viquipèdia pot ésser vàlida encara que la seva eficàcia pugui variar en funció del domini i del grau d'especialització del text a analitzar.

Aquesta estratègia per reconèixer termes ha sigut aplicada en diverses ocasions i àmbits:

— extracció de termes en medicina [36];
— estudi de la terminologia emprada en llibres de text a Mèxic: [38] i [39];
— projectes de màster en llengua italiana, portuguesa i francesa;
— projectes Alinea i APLE2 de l'IULA.[39]

39. A *http://eines.iula.upf.edu/WikiYATE/wikiYate.html* hi ha disponible una interfície gràfica que permet visualitzar els documents d'aquest projecte per a tots els àmbits del Corpus Tècnic de l'IULA i els termes escollits amb els contextos respectius.

7. Conclusions

En aquest treball s'han mostrat diferents aspectes de la Viquipèdia que van des d'una descripció detallada de l'estructura interna fins a la informació que conté. S'han analitzat algunes característiques tècniques que obliguen a prendre certes precaucions quan es consulta aquest recurs i, en particular, quan es vol utilitzar l'estructura bigraf. Aquestes qüestions no impedeixen que la Viquipèdia s'hagi utilitzat en multitud d'aplicacions per al PLN. Una qüestió sempre pendent quan es parla de la Viquipèdia és la credibilitat. El fet que no existeix un equip d'editors centralitzat fa que aquest tema surti freqüentment a la palestra. Al respecte, es mostren molts estudis i fins i tot propostes de millora i/o detecció de vandalisme en algunes pàgines. La Fundació Wikimedia és molt conscient d'aquest problema i la política de creació i edició de pàgines ha anat evolucionant a mesura que van apareixent manipulacions i altres problemes. S'han descrit algunes inconsistències quan analitzem la posició d'una mateixa categoria (cat.) en diferents llengües. Aquestes inconsistències no s'han estudiat com mereixen i poden representar un problema en funció de l'aplicació que s'estigui donant a la WP. A més a més, posa de relleu la importància del treball dels editors en la definició de quines categories s'assignen a una pàgina i de quin paper tenen aquestes categories en l'arbre de categories.

Malgrat el que acabem de mencionar, la Viquipèdia ha sigut molt utilitzada en diverses àrees del PLN. Al mateix temps, és part essencial d'altres fonts de coneixement de gran difusió que la prenen com a referència. En aquest treball es mencionen algunes d'aquestes aplicacions i, en particular, s'han descrit amb cert detall dues aplicacions específiques de l'àmbit de la terminologia, com són construir automàticament un recull de termes d'un domini i, donat un text, fer l'extracció de termes d'un domini. Aquests treballs demostren que és factible utilitzar la Viquipèdia en treballs terminològics. En ambdós casos queden alguns dubtes sobre el seu comportament quan l'àmbit de treball és molt especialitzat o quan som davant d'un àmbit transversal. En qualsevol cas, és un recurs que mereix ser tingut en compte en aplicacions que requereixen l'us de la terminologia.

Finalment, cal lamentar que la majoria dels treballs esmentats en l'apartat 6 facin referència fonamentalment a tasques fetes amb la llengua anglesa. El català, tot i ser una llengua molt present a l'univers Wikipedia, és clarament minoritari pel que fa als projectes que l'utilitzen.

Bibliografia

[1] Jullien, N. (2012). «What we know about Wikipedia: A review of the literature analyzing the project(s)». *HAL* [en línia]. 86 p. <https://hal.archives-ouvertes.fr/hal-00857208/document>.

[2] Nielsen, F. A. (2012). «Wikipedia research and tools: review and comments». *SSRN Electronic Journal* [en línia]. 66 p. <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2129874>.
[3] Zesch, T.; Müller, C.; Gurevych, I. (2008). «Extracting lexical semantic knowledge from Wikipedia and Wiktionary». A: *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. París: ELRA, p. 1646-1652.
[4] Kaptein, R.; Kamps, J. (2013). «Exploiting the category structure of Wikipedia for entity ranking». *Artificial Intelligence,* vol. 194, p. 111-129.
[5] Azer, S. [*et al.*] (2015). «Accuracy and readability of cardiovascular entries on Wikipedia: are they reliable learning resources for medical students?». *BMJ Open*, vol. 5 (10), p. 1-14. DOI: 10.1136/bmjopen-2015-008187.
[6] Tomaszewski, R.; MacDonald, K. I. (2016). «A study of citations to Wikipedia in scholarly publications». *Science & Technology Libraries*, vol. 35 (3), p. 246-261.
[7] Milne, D.; Witten, I. H. (2013). «An open-source toolkit for mining Wikipedia». *Artificial Intelligence*, vol. 194, p. 222-239.
[8] Lehmann, J. [*et al.*] (2015). «DBpedia - A large-scale, multilingual knowledge base extracted from Wikipedia». *Semantic Web*, vol. 6 (2), p. 167-195.
[9] Lewoniewski, W.; Wecel, K.; Abramowicz, W. (2018). «Determining quality of articles in polish Wikipedia based on linguistic features». A: *International Conference on Information and Software Technologies*.
[10] Harpalani, M. [*et al.*] (2011). «Language of vandalism: improving Wikipedia vandalism detection via stylometric analysis». A: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Portland: Association for Computational Linguistics, p. 83-88.
[11] Sarabadani, A.; Halfaker, A.; Taraborelli, D. (2017). «Building automated vandalism detection tools for Wikidata». A: *Proceedings of the 26th International Conference on World Wide Web Companion*. Cantó de Ginebra (Suïssa): International World Wide Web Conferences Steering Committee, p. 1647-1654.
[12] Jemielniak, D.; Masukume, G.; Wilamowski, M. (2019). «The most influential medical journals according to Wikipedia: quantitative analysis». *Journal of Medical Internet Research*, vol. 21 (1). També disponible en línia a: <https://www.jmir.org/2019/1/e11429/>.
[13] Heilman, J. M. [*et al.*] (2011). «Wikipedia: a key tool for global public health promotion». *Journal of Medical Internet Research*, vol. 13 (1). També disponible en línia a: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3221335/>.
[14] Martin-Carreras, T.; Kahn, C. E. (2019) «Integrating Wikipedia articles and images into an information resource for radiology patients». *Journal of Digital Imaging* (1 juny), vol. 32 (3), p. 349-353.
[15] Tufiş, D. [*et al.*] (2013). «Wikipedia as an SMT training corpus». A: *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*. Hissar: INCOMA, p. 702-709.

[16] Plamada, M.; Volk, M. (2013). «Mining for domain-specific parallel text from Wikipedia». A: *Proceedings of the 6th Workshop on Building and Using Comparable Corpora*. Sofia: ACL, p. 112-120.

[17] Nastase, V.; Filippova, K.; Milne, D. (2009). «Summarizing with encyclopedic knowledge». A: *Proceedings of the 2nd Text Analysis Conference*. Gaithersburg: National Institute of Standards and Technology.

[18] Melo, G.; Weikum, G. (2010). «MENTA: inducing multilingual taxonomies from Wikipedia». A: *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*. Nova York: ACM, p. 1099-1108.

[19] Sedigheh, K.; Abolghasem, M. S. (2015). «Automatic construction of domain ontology using Wikipedia and enhancing it by Google search engine». *Journal of Information Systems and Telecommunication* [ACM], vol. 3 (4), p. 248-258.

[20] Mihalcea, R.; Csomai, A. (2007). «Wikify!: linking documents to encyclopedic knowledge». A: *CIKM'07: Proceedings of the sixteenth ACM Conference on Information and Knowledge Management*. Nova York: ACM, p. 233-242. També disponible en línia a: <https://dl.acm.org/doi/proceedings/10.1145/1321440>.

[21] Ferragina, P.; Scaiella, U. (2010). «TAGME: on-the-fly annotation of short text fragments (by Wikipedia entities)». A: *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*. Nova York: ACM, p. 1625-1628.

[22] Wang, Z. [*et al.*] (2017). «Entity linking in queries using word, mention and entity joint embedding». A: Wang, Z.; Turhan, A. Y.; Wang, K.; Zhang, X. (ed.). *Semantic Technology*, vol. 10675. Cham: Springer. (Lecture Notes in Computer Science)

[23] Gabrilovich, E.; Markovitch, S. (2009). «Wikipedia-based semantic interpretation for natural language processing». *Journal of Artificial Intelligence Research*, vol. 34, p. 443-498.

[24] Gabrilovich, E.; Markovitch, S. (2007). «Computing semantic relatedness using Wikipedia-based explicit semantic analysis». A: *Proceedings of the 20th International Joint Conference on Artificial Intelligence* (Índia). San Francisco: Morgan Kaufmann, p. 1606-1611.

[25] Yazdani, M.; Popescu-Belis, A. (2013). «Computing text semantic relatedness using the contents and links of a hypertext encyclopedia». *Artificial Intelligence*, vol. 194, p. 176-202.

[26] Fernandes, E. R. [*et al.*] (2016). «Using Wikipedia for cross-language named entity recognition». A: *Big Data Analytics in the Social and Ubiquitous Context*. Cham: Springer.

[27] Vivaldi, J.; Rodríguez, H. (2012). «Using Wikipedia for domain terms extraction». A: *Proceedings of CHAT 2012: The 2nd Workshop on the Creation; Harmonization and Application of Terminology Resources*. Linköping: Linköping University Electronic Press, p. 3-10.

[28] Vivaldi, J.; Rodríguez, H. (en premsa). «Automatically producing semantically tagged bilingual terminologies».

[29] Vossen, P. (1998). «The EuroWordNet Annual Report 1998». Amsterdam: Vrije Universiteit. [Document de treball]
[30] Fellbaum, C. (1999). «Wordnet: an electronic lexical database». *The Library Quarterly* [Chicago: The University of Chicago Press], vol. 69, p. 406-408.
[31] Magnini, B.; Cavaglià, G. (2000). «Integrating subject field codes into WordNet». A: *Proceedings of the Language Resources and Evaluation Conference (LREC 2000)*. Atenes: ELRA, p. 1413-1418.
[32] Bentivogli, L. [*et al.*] (2004) «Revising the Wordnet domains hierarchy: semantics, coverage and balancing». A: *Proceedings of the Workshop on Multilingual Linguistic Ressources*. Stroudsburg: Coling, p. 94-101.
[33] Spackman, K. A.; Campbell, K. E.; Côté, R. A. (1997). «SNOMED RT: a reference terminology for health care». A: *Proceedings of the AMIA Annual Fall Symposium*. Nashville: Hanley & Belfus, vol. 4, p. 640-644.
[34] Niemann, E.; Gurevych, I. (2011). «The people's web meets linguistic knowledge: automatic sense alignment of Wikipedia and WordNet». A: *Proceedings of the 9th International Conference on Computational Semantics*. Oxford: ACL, p. 205-214.
[35] Vivaldi, J. (2001). *Extracción de candidatos a término mediante combinación de estrategias heterogéneas*. Tesi doctoral. Barcelona: Universitat Politècnica de Catalunya.
[36] Vivaldi, J.; Rodríguez, H. (2011). «Using Wikipedia for term extraction in the biomedical domain: first experience». *Procesamiento del Lenguaje Natural*, vol. 45, p. 251-254.
[37] Vivaldi, J. (2009). «Corpus and exploitation tool: IULACT and bwanaNet». A: *A Survey on Corpus-based Research. Proceedings of the I International Conference on Corpus Linguistics (CICL 2009)*. Múrcia: Universidad de Murcia, p. 224-239.
[38] Cabrera-Diego, L. A. [*et al.*] (2011). «Using Wikipedia to validate term candidates for the Mexican basic scientific vocabulary». A: *Proceedings of the First International Conference on Terminology, Languages, and Content Resources (LaRC)*. Seül, p. 76-85.
[39] Cabrera-Diego, L. A. [*et al.*] (2012). «Using Wikipedia to validate the terminology found in a corpus of basic textbooks». A: *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. Istanbul: ELRA, p. 3820-3827.
[40] Vivaldi, J.; Rodríguez, H.; Rigau, G. (2013). «Combining Wikipedia and WordNet for improving domain terms compilation». A: *Proceedings of the 14th International Conference, CICLing 2013* (Samos). Berlín: Springer.
[41] Vivaldi, J.; Rodríguez, H. (2014). «Arabic medical terms compilation from Wikipedia». A: *2014 Third IEEE International Colloquium in Information Science and Technology (CIST)*. Tetuan: IEEE, p. 248-253.

Judit Feliu i Mireia Trias (cur.)
Viquipèdia i terminologia
Barcelona: Institut d'Estudis Catalans, 2021, p. 49-54 DOI: 10.2436/15.2503.02.65

És fiable la Viquipèdia? Manteniment, estandardització i control a la Viquipèdia en català

Pau Cabot i Bonnín
Amical Wikimedia

Un dels debats que existeixen des del naixement de la Viquipèdia, ara fa divuit anys, és el de la fiabilitat d'aquesta eina. És fiable la Viquipèdia en català? Com tantes d'altres obres, té els seus errors, però almanco és tremendament útil. Les dades de consulta d'aquesta eina són aclaparadores, en aquest sentit.

Per tal de millorar la fiabilitat de l'obra, desenes de viquipedistes treballen per mantenir i millorar la qualitat de l'enciclopèdia lliure. Quines són les eines de control que tenen aquests viquipedistes? Aquí, breument, intentarem explicar quines són aquestes eines que permeten millorar la qualitat d'aquesta enciclopèdia en línia.

1. Wikipedia

Wikipedia és el cinquè lloc web més visitat d'Internet. Rep mensualment 8.000 milions de visites i es publica en 303 llengües. La versió en anglès tota sola té actualment uns 6 milions d'articles i totes les edicions lingüístiques sumen un total de 40 milions d'articles. Estam parlant d'un fenomen que ha vengut per quedar-se i que, a causa de la llicència lliure que usa, podria ser la font de moltes derivades en el futur.

2. La Viquipèdia en català

La Viquipèdia en català es va crear el 16 de març de 2001 i, avui dia, té uns 600.000 articles i rep uns 20 milions de visites mensuals. És la vintena edició lingüística en nombre d'articles i la trenta-sisena en nombre de consultes.

Pel que fa al nombre d'articles, i en l'edat més madura de la Viquipèdia en català, podem constatar dues fases: una primera (que va entre 2003 i 2011) en què s'hi crearen una mitjana de 5.200 articles mensuals (170 de diaris) i una fase més recent (del 2011 fins a l'actualitat) en què la mitjana baixa a 3.100 articles per mes (100 de diaris). Això indica que els temps en què havíem de reivindicar la quantitat d'articles ja han passat i que la comunitat és conscient que ens hem de centrar més en la qualitat.

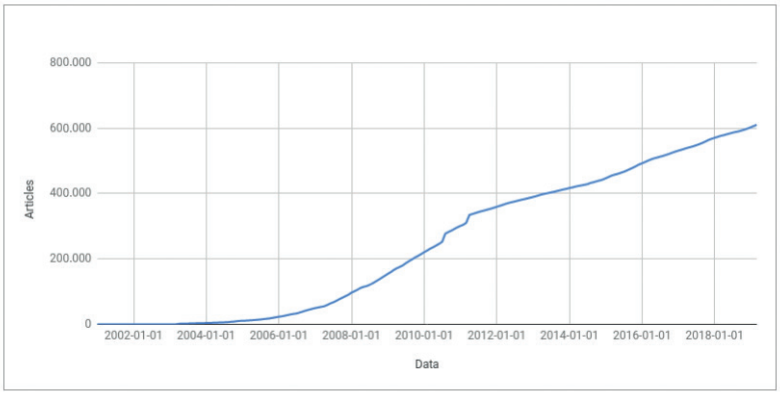


FIGURA 1. Evolució del nombre d'articles a la Viquipèdia en català.
FONT: Elaboració pròpia.

3. RECURSOS HUMANS

Totes aquestes feines (i d'altres que no hem apuntat), les fa la comunitat viquipedista, que cada dia es connecta per tenir cura de l'enciclopèdia. En xifres mensuals, podríem descriure aquesta comunitat de la manera següent:

— 1.700 usuaris que han fet almanco una edició;
— 800 usuaris actius, que han fet almanco cinc edicions;
— 70 usuaris molt actius, que han fet més de cent edicions;
— 21 administradors, que són escollits per la comunitat i que tenen la seva confiança per esborrar articles que no són enciclopèdics o que infringeixen drets d'autor i que, a més, poden bloquejar usuaris que fan edicions perjudicials per a l'obra.

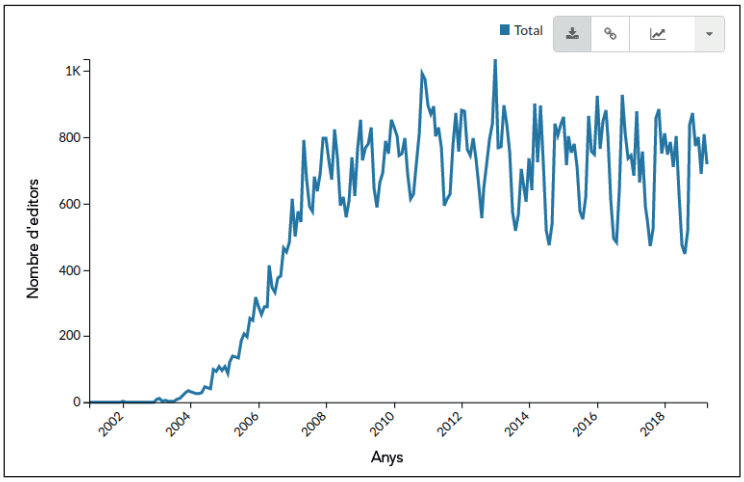


Figura 2. Evolució del nombre d'editors actius.
Font: *stats.wikimedia.org*.

4. Tasques de manteniment

Els viquipedistes, a part d'anar afegint articles nous, fan una tasca molt valuosa de corregir, mantenir i actualitzar el que ja és present a l'enciclopèdia.

Les tasques de manteniment que du a terme la comunitat són, per exemple: ajudar els companys, corregir articles i actualitzar-los, estandarditzar els articles, posar avisos quan es detecta una mancança i traduir la interfície del programari que usa la Viquipèdia, el Mediawiki.

Entre les tasques més lleus que sí que han de fer diàriament els viquipedistes hi ha:

— Corregir enllaços incorrectes a pàgines de desambiguació.
— Ampliar articles massa curts.
— Fusionar articles que parlen del mateix concepte.
— Arreglar enllaços externs no actius.

I entre les greus, podríem destacar:

— Revertir vandalismes.
— Eliminar publicitat.
— Corregir articles amb faltes d'ortografia.
— Traduir articles amb traducció per millorar.
— Eliminar textos o imatges amb drets d'autor.
— Esborrar articles no admissibles.
— Esmenar articles no neutrals.

5. Volum de feina

Algunes dades ens poden ajudar a entendre quin és el volum de feina diari que ha de gestionar la comunitat viquipedista. A part de crear les ja mencionades 100 pàgines noves, podríem destacar també:

— Eliminar 3.000 edicions que no són creacions de pàgines.
— Esborrar 25 pàgines.
— Bloquejar 8 usuaris.
— Detectar 2 persones amb conflicte d'interessos que venen a fer publicitat.

6. Canvis recents, pàgines noves i llista de seguiment

Les tres eines principals que ajuden la comunitat viquipedista a revisar els nous continguts que s'afegeixen diàriament a la Viquipèdia són tres pàgines especials que descriurem a continuació:

— Canvis recents:[1] llista els canvis que es produeixen a la Viquipèdia en temps real. Els viquipedistes poden revisar quines edicions s'han fet, en quin moment, quin usuari les ha fet, si s'ha afegit o s'ha eliminat informació, etc.

— Pàgines noves:[2] llista els articles creats recentment, indicant-ne el moment de creació, l'usuari, etc.

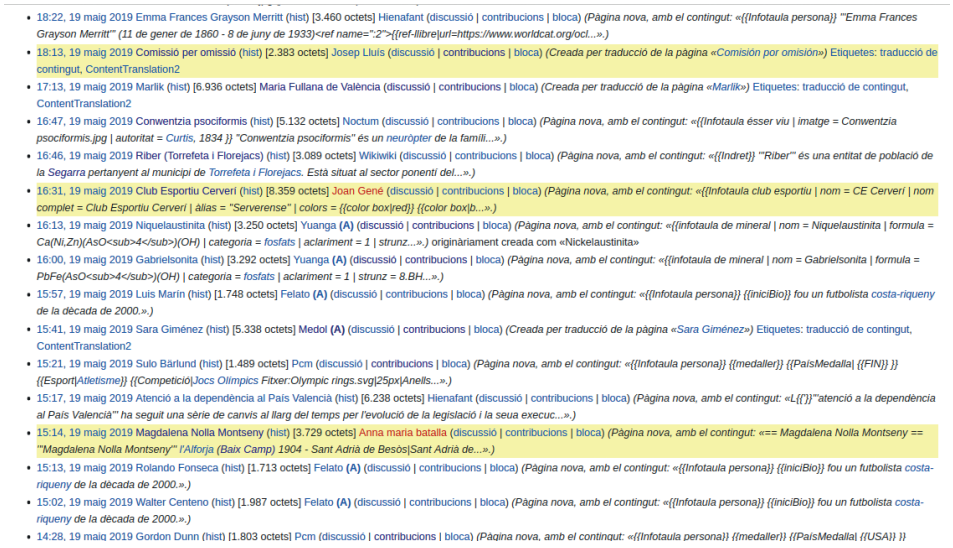
— Llista de seguiment:[3] llista les modificacions recents que s'han efectuat als articles que estan vigilats. Aquesta és una llista privada i específica de cada viquipedista. Només està accessible si ens hem donat d'alta com a usuaris i hem iniciat la sessió.

7. Uns ajudants molt especials: els bots

Gestionar centenars de milers d'articles és complex. Fer un petit canvi en l'estructura dels articles, corregir una falta d'ortografia comuna o adaptar el text dels articles a la nova ortografia del català pot implicar des de centenars fins a milions de canvis en els articles. Per sort, aquestes tasques repetitives les fan, molt més ràpidament i de manera molt més fiable, robots. Aquests robots (que a la Viquipèdia anomenam, simplement, *bots*) són programats per viquipedistes i fan tasques repetitives a un conjunt d'articles que podem delimitar.

Sense aquesta eina d'edició massiva, la Viquipèdia no podria haver evolucionat tan ràpid com ho ha fet: ha permès, a part de fer correccions de faltes d'ortografia, actualitzar els articles a les tecnologies d'avui en dia, que són molt diferents

1. *ca.wikipedia.org/wiki/Especial:Canvis_recents*.
2. *https://ca.wikipedia.org/wiki/Especial:P%C3%A0gines_noves*.
3. *https://ca.wikipedia.org/wiki/Especial:Llista_de_seguiment*.



- 18:22, 19 maig 2019 Emma Frances Grayson Merritt (hist) [3.460 octets] Hienafant (discussió | contribucions | bloca) *(Pàgina nova, amb el contingut: «{{Infotaula persona}} '"Emma Frances Grayson Merritt"' (11 de gener de 1860 - 8 de juny de 1933)<ref name=":2">{{ref-llibre|url=https://www.worldcat.org/ocl...».)*
- 18:13, 19 maig 2019 Comissió per omissió (hist) [2.383 octets] Josep Lluís (discussió | contribucions | bloca) *(Creada per traducció de la pàgina «Comisión por omisión»)* Etiquetes: traducció de contingut, ContentTranslation2
- 17:13, 19 maig 2019 Marlik (hist) [6.936 octets] Maria Fullana de València (discussió | contribucions | bloca) *(Creada per traducció de la pàgina «Marlik»)* Etiquetes: traducció de contingut, ContentTranslation2
- 16:47, 19 maig 2019 Conwentzia psociformis (hist) [5.132 octets] Noctum (discussió | contribucions | bloca) *(Pàgina nova, amb el contingut: «{{Infotaula ésser viu | imatge = Conwentzia psociformis.jpg | autoritat = Curtis, 1834 }} "Conwentzia psociformis" és un neuròpter de la famili...».)*
- 16:46, 19 maig 2019 Riber (Torrefeta i Florejacs) (hist) [3.089 octets] Wikiwiki (discussió | contribucions | bloca) *(Pàgina nova, amb el contingut: «{{Indret}} '"Riber"' és una entitat de població de la Segarra pertanyent al municipi de Torrefeta i Florejacs. Està situat al sector ponenti del...».)*
- 16:31, 19 maig 2019 Club Esportiu Cerverí (hist) [8.359 octets] Joan Gené (discussió | contribucions | bloca) *(Pàgina nova, amb el contingut: «{{Infotaula club esportiu | nom = CE Cerverí | nom complet = Club Esportiu Cerverí | àlias = "Serverense" | colors = {{color box|red}} {{color box|b...».)*
- 16:13, 19 maig 2019 Niquelaustinita (hist) [3.250 octets] Yuanga **(A)** (discussió | contribucions | bloca) *(Pàgina nova, amb el contingut: «{{infotaula de mineral | nom = Niquelaustinita | formula = Ca(Ni,Zn)(AsO₄)(OH) | categoria = fosfats | aclariment = 1 | strunz...».)* originàriament creada com «Nickelaustinita»
- 16:00, 19 maig 2019 Gabrielsonita (hist) [3.292 octets] Yuanga **(A)** (discussió | contribucions | bloca) *(Pàgina nova, amb el contingut: «{{infotaula de mineral | nom = Gabrielsonita | formula = PbFe(AsO₄)(OH) | categoria = fosfats | aclariment = 1 | strunz = 8.BH...».)*
- 15:57, 19 maig 2019 Luis Marín (hist) [1.748 octets] Felato **(A)** (discussió | contribucions | bloca) *(Pàgina nova, amb el contingut: «{{Infotaula persona}} {{iniciBio}} fou un futbolista costa-riqueny de la dècada de 2000.».)*
- 15:41, 19 maig 2019 Sara Giménez (hist) [5.338 octets] Medol **(A)** (discussió | contribucions | bloca) *(Creada per traducció de la pàgina «Sara Giménez»)* Etiquetes: traducció de contingut, ContentTranslation2
- 15:21, 19 maig 2019 Sulo Bärlund (hist) [1.489 octets] Pcm (discussió | contribucions | bloca) *(Pàgina nova, amb el contingut: «{{Infotaula persona}} {{medaller}} {{PaísMedalla| {{FIN}} }} {{Esport|Atletisme}} {{Competició|Jocs Olímpics Fitxer:Olympic rings.svg|25px|Anells...».)*
- 15:17, 19 maig 2019 Atenció a la dependència al País Valencià (hist) [6.238 octets] Hienafant (discussió | contribucions | bloca) *(Pàgina nova, amb el contingut: «L{{'}}'atenció a la dependència al País Valencià'' ha seguit una sèrie de canvis al llarg del temps per l'evolució de la legislació i la seua execuc...».)*
- 15:14, 19 maig 2019 Magdalena Nolla Montseny (hist) [3.729 octets] Anna maria batalla (discussió | contribucions | bloca) *(Pàgina nova, amb el contingut: «== Magdalena Nolla Montseny == '"Magdalena Nolla Montseny"' l'Alforja (Baix Camp) 1904 - Sant Adrià de Besòs|Sant Adrià de...».)*
- 15:13, 19 maig 2019 Rolando Fonseca (hist) [1.713 octets] Felato **(A)** (discussió | contribucions | bloca) *(Pàgina nova, amb el contingut: «{{Infotaula persona}} {{iniciBio}} fou un futbolista costa-riqueny de la dècada de 2000.».)*
- 15:02, 19 maig 2019 Walter Centeno (hist) [1.987 octets] Felato **(A)** (discussió | contribucions | bloca) *(Pàgina nova, amb el contingut: «{{Infotaula persona}} {{iniciBio}} fou un futbolista costa-riqueny de la dècada de 2000.».)*
- 14:28, 19 maig 2019 Gordon Dunn (hist) [1.803 octets] Pcm (discussió | contribucions | bloca) *(Pàgina nova, amb el contingut: «{{Infotaula persona}} {{medaller}} {{PaísMedalla| {{USA}} }}*

FIGURA 3. Registre de pàgines noves de la Viquipèdia.
FONT: *https://ca.wikipedia.org/wiki/Special:Newpages.*

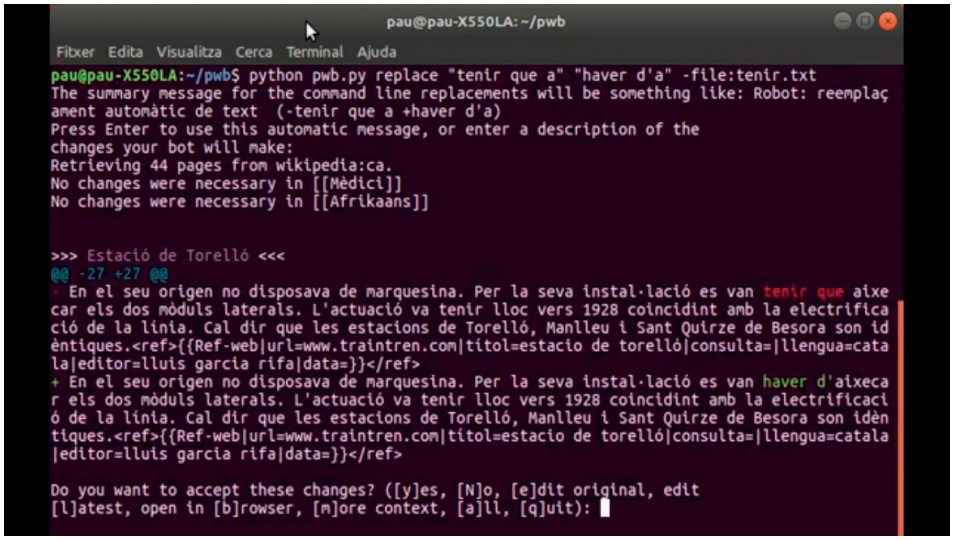


FIGURA 4. Funcionament d'un bot corrector a la Viquipèdia.
FONT: Captura de pantalla del programa Pywikibot feta per l'autor.

de les que teníem fa deu anys, quan la Viquipèdia tenia una quarta part dels articles que té ara.

Fonts

Les estadístiques provenen de la mateixa Viquipèdia i de la pàgina d'estadístiques *https://stats.wikimedia.org*.

Judit Feliu i Mireia Trias (cur.)
Viquipèdia i terminologia
Barcelona: Institut d'Estudis Catalans, 2021, p. 55-58 DOI: 10.2436/15.2503.02.66

Projecte Viquiterm

Ramon Garriga
Fundació Torrens-Ibern

Toni Hermoso Pulido
Serveis Cientificotècnics, Centre for Genomic Regulation (CRG)
Amical Wikimedia[1]
Societat Catalana de Biologia

1. La Fundació Torrens-Ibern

La iniciativa de la Fundació Torrens-Ibern, orientada a normalitzar l'ús del català en els àmbits científic i tècnic, s'inscriu en el procés democratitzador de la universitat que implica l'impuls de l'ús del català en aquesta institució. Van ser els anomenats professors no numeraris (PNN) els que en una assemblea que tingué lloc el 7 d'abril de 1975 decidiren proposar a la Junta de l'Escola d'Enginyers de Barcelona dues coses:

a) Oferir classes de català a l'Escola a dos nivells; classes de comprensió per a persones que desconeguessin la llengua i classes d'escriptura i gramàtica per a persones que coneguessin la llengua parlada.

b) Promoure una fundació per a l'estudi de la terminologia tècnica i científica catalana a la qual posarien el nom de Joaquim Torrens Ibern, catedràtic de l'Escola, gran persona, que havia mort recentment i que des de la tornada de l'exili havia promogut la llengua i la cultura catalanes. Ell fou qui presentà la proposta de col·laboració de l'Escola amb el Congrés de Cultura Catalana.

Els anys 1975-1976 va ser quan es va formar el grup de llançament de la Fundació, el període 1976-1977 va ser ja el d'actuació d'una comissió promotora i el 13 de maig de 1977 es va establir l'escriptura de constitució de la Fundació i se'n nomenà el primer Patronat. L'aprovació oficial arribà el 20 de juny de 1978, és a

1. En el moment de la conferència, Toni Hermoso Pulido era membre d'Amical Wikimedia, però actualment ja no ho és.

dir, fa més de quaranta anys. Els objectius fundacionals eren l'estudi, la difusió i la promoció de l'aplicació de la llengua i la cultura catalanes en els camps tècnic i científic, principalment a Catalunya, mitjançant la realització de treballs de la mateixa entitat; la concessió de beques, ajuts i subvencions a persones o entitats; l'organització de cursos, seminaris, concursos i certàmens; la distribució de premis a les persones que es distingissin en el camp de la finalitat fundacional i l'edició o distribució de llibres i revistes que contribuïssin a la finalitat expressada.

Amb el temps, i amb la creació d'entitats oficials (TERMCAT, serveis de llengües i terminologia a les universitats…) responsables d'alguns dels objectius que s'havia fixat la Fundació en un moment en què no existien, la Fundació ha anat adaptant-se a una funció d'impuls i de complementarietat. Ha editat vocabularis bàsics, ha participat en el Projecte Scriptorium amb l'IEC i altres entitats, ha creat el Projecte Ubertas, ha col·laborat en la traducció de normes ISO que ha derivat en un conveni de col·laboració del TERMCAT amb AENOR, ha donat suport a l'edició de diccionaris com el de física i el de química del TERMCAT, ha organitzat sessions de diàleg sobre terminologia, ha creat els premis Joaquim Torrens Ibern i Enric Freixa, aquest en record de qui va ser el primer president de la Fundació, etc.

I ara tot fa pensar que és el moment de posar en marxa el fòrum Viquiterm.

2. El Projecte Viquiterm

2.1. *Introducció*

El predomini de l'anglès com a idioma generalitzat en l'àmbit científic i tècnic i l'alt ritme de desenvolupament tecnològic actual representen tot un repte perquè la terminologia emergent que hi va associada pugui ser assumida satisfactòriament pel conjunt de la comunitat catalanoparlant, tant pel que fa als experts en l'àmbit científic o tècnic en concret com per la resta de la societat. Tot i l'excellent feina i el rigor tant de les societats cientificotècniques com dels centres de terminologia (sense ànim de ser exhaustius podem pensar en diferents filials de l'Institut d'Estudis Catalans, en el TERMCAT, principalment amb el Cercaterm…), els ritmes i procediments que han de seguir a vegades no s'adeqüen a les necessitats de comunicació actuals, en què la immediatesa de les comunicacions (per exemple, amb les xarxes socials) requereix l'adopció de solucions ràpides, encara que siguin temporals o de compromís.

Per això pensem que cal desenvolupar un fòrum obert, àgil i rigorós en el qual es puguin suggerir traduccions amb les característiques acabades d'esmentar a l'espera que les autoritats en la matèria, TERMCAT i Institut d'Estudis Catalans (IEC), donin la traducció definitiva.

2.2. *Entitats participants*

El projecte està obert a la participació de tots aquells que hi tinguin alguna cosa a dir. És evident que sense el vistiplau del TERMCAT i de l'IEC aquest projecte no seria possible, ja que a ells caldrà reportar els resultats dels debats que s'hagin produït al fòrum, així com els mateixos debats. Amb l'IEC, a més, s'està estudiant la possibilitat d'allotjar el fòrum en els seus equipaments informàtics.

L'impuls inicial és de la Fundació Torrens-Ibern (FT-I), que ha trobat en Amical Wikimedia i en la seva magnífica experiència en actuacions equivalents un soci de primer nivell.

La FT-I liderarà i es farà càrrec del cost de la primera fase del projecte i buscarà altres entitats que puguin ajudar per a una segona fase, més potent, com es veurà més endavant, una vegada el prototip hagi estat validat. Es preocuparà de buscar els possibles usuaris en l'àmbit tècnic, universitari i de recerca, així com d'ampliar el nombre d'entitats participants per fer créixer el projecte tant des del punt de vista d'usuaris com del de suports lingüístics que facilitin la moderació dels debats del fòrum.

Amical Wikimedia i Softcatalà ajudaran en el desenvolupament tècnic de la idea a partir de la seva experiència en altres fòrums. Amical, a més, també ajudarà en el desenvolupament dels rols dels diferents actors a mesura que avancin amb la solució tècnica.

2.3. *Planificació del projecte*

Una vegada en funcionament normal, el fòrum consistirà en diferents fils de debat en què els participants plantejaran un dubte, una qüestió o una proposta terminològica com a conseqüència d'haver descobert un mot o una expressió la traducció dels quals no ha trobat en cap de les eines a l'abast. Cada fil s'assignarà a una categoria temàtica concreta, amb la possibilitat que alguns fils puguin tenir-ne més d'una, i rebrà comentaris dels diferents participants. Aquestes propostes inicials i comentaris en algun moment esdevindran propostes definitives (per votació, consens o decisió del moderador responsable), que seran relacionades com a tals en el diccionari repositori del fòrum i trameses amb tota la deliberació al TERMCAT per tal d'iniciar el procés oficial. Per això la deliberació ha de ser pública i plenament referenciable per terceres parts.

Ara bé, abans d'entrar en aquest funcionament normal i amb la plataforma tecnològica totalment desenvolupada per donar-hi servei, caldrà definir un prototip que haurà de validar a petita escala la resposta per part dels usuaris potencials i que ha de ser escalable per a desenvolupaments posteriors.

Una primera tasca a fer, i que ja s'ha fet parcialment, és detectar l'interès dels possibles usuaris en el fòrum. S'ha entrat en contacte amb els serveis lingüístics de diferents universitats i la resposta ha estat totalment positiva. Una vegada es disposi del prototip, la definició i realització del qual està també en curs, s'iniciarà una campanya de contactes amb professors i recercadors que puguin estar més en línia amb els objectius del fòrum per acabar de polir la metodologia i també per adquirir un cert compromís de participació per part seva.

En la primera fase de posada en marxa, totes les propostes s'inclouran en un únic fil i el fòrum tindrà un moderador que a la vegada assumirà les funcions d'animador. Es tracta de veure la resposta que rep la iniciativa. En funció d'aquesta resposta es decidirà com i quan es faran els passos següents.

2.4. *Plataforma tecnològica i consideracions tècniques*

La plataforma base triada és Discourse, possiblement el sistema de gestió de fòrums més popular a Internet avui en dia. El sistema està desenvolupat majoritàriament en els llenguatges de programació Javascript i Ruby, i per darrere funciona amb una base de dades PostgreSQL. Tot i la diversitat de programes i llenguatges utilitzats, la seva instal·lació i gestió està simplificada per l'ús de contenidors de programari, en concret de Docker, i moltes de les tasques d'administració es poden dur a terme des de la mateixa interfície web.

Allò que per defecte proporciona la plataforma, juntament amb unes quantes modificacions i personalitzacions, per exemple, de fils de debat, configura el que seria el sistema prototip que es compartirà amb el públic objectiu. Segons la resposta i comentaris rebuts, s'implementaran modificacions, sigui mitjançant connectors de Discourse ja existents, o bé amb desenvolupaments específics per aconseguir el fòrum final que es vulgui obtenir. Paral·lelament també es confeccionaria i es traslladaria al web un disseny d'acord amb les preferències i comentaris que s'hagin rebut fins llavors.

Als estadis inicials d'aquest procés de personalització de la plataforma, també es procediria a la localització del programari, que es troba en el portal de traducció col·laborativa Transifex. Es traduirien totes les cadenes pendents que hi hagi i es revisarien i corregirien la resta d'acord amb un criteri terminològic coherent amb el d'altres projectes de programari lliure sota la supervisió de Softcatalà. Com a resultat, el Discourse passaria a estar disponible plenament en català, no només per al Viquiterm, sinó també per a iniciatives de tercers que optin per utilitzar aquesta plataforma a partir d'aquell moment.

Judit Feliu i Mireia Trias (cur.)
Viquipèdia i terminologia
Barcelona: Institut d'Estudis Catalans, 2021, p. 59-88 DOI: 10.2436/15.2503.02.67

Taula rodona: «Experiències de les societats filials de l'IEC en la Viquipèdia»

1. Viquipèdia, matemàtiques, terminologia[1]

1.1. *Viquipèdia i les matemàtiques en català*

Dins de la comunitat matemàtica, tant en l'ensenyament primari i secundari com en la universitat i la recerca, la Wikipedia, o la seva versió catalana, la Viquipèdia, tenen un paper molt important, i d'importància creixent.

En la meva manera de veure, allò més important que ens ofereix és l'accés enciclopèdic ràpid al coneixement matemàtic. Això vol dir que la Wikipedia no és el lloc on aprendre matemàtiques, però sí que és el lloc on recordar fórmules o conceptes per a aquells que ja n'han tingut nocions prèviament. I això és ben útil, i substitueix el paper de molts llibres, i a més amb un accés molt fàcil i còmode.

Moltes vegades, els professors recomanem l'accés a la Wikipedia als estudiants que han oblidat o no dominen prou un tema concret. En la meva experiència com a professor de primer curs d'universitat, així ho he fet amb temes com el teorema del binomi, els nombres complexos o d'altres, que més o menys podria dir-se que podrien saber-se des de l'ensenyament secundari, però que de vegades els estudiants de primer curs no dominen suficientment.

També els investigadors recorrem a la Wikipedia per buscar fórmules o gràfiques que recordem una mica, però no amb prou precisió com pugui ser requerida en el treball que estiguem fent.

Algú podria preguntar-nos si ens fiem completament del que diu la Wikipe-

1. Aquest article correspon a la participació de Joan de Solà-Morales en la XVI Jornada SCATERM: «La Viquipèdia i la terminologia» (30 de maig de 2019).

dia. Aquesta és una pregunta pertinent. La resposta és, d'una banda, que el propi esperit crític i els coneixements generals del lector són una bona arma per a detectar possibles errors, i, de l'altra, que forma part de l'estadística personal de tots els usuaris, si més no de matemàtiques, que a la Wikipedia hi ha molt pocs errors. La mateixa facilitat per a editar i canviar un article és una garantia d'aquest fet.

Però sí que és veritat que molts articles contenen coses que no són falses, però són opinables. Això afecta els conceptes i les precisions terminològiques, més que no pas la correcció de les afirmacions matemàtiques. També afecta la manera com s'expliquen les coses. Avui en dia (escric això a la primavera del 2019) preferim els exemples entenedors i el foment de la intuïció que no pas el rigor en les definicions. El rigor també hi ha de ser, no ho nego pas, però s'ha de començar per la intuïció, i no quedar-se mai paralitzat per la manca de precisió. La precisió sempre arriba, però després. Aquestes reflexions m'han fet discrepar de la redacció d'alguns articles de la Viquipèdia, però haig de dir que he procurat de no deixar-me portar massa per les meves opinions subjectives, i admetre que al coneixement s'hi pot accedir sempre per camins diferents.

Cap a l'any 2015 vam iniciar un treball sobre la Viquipèdia en l'àrea que vam anomenar «Matemàtica intermèdia». El treball es va iniciar a la Secció de Ciències i Tecnologia de l'IEC i va consistir en una recopilació de temes de matemàtiques als nivells de batxillerat i primer curs d'universitat, centrant-nos principalment en coneixements no tan especialitzats com els que formen part de carreres universitàries com matemàtiques, estadística o física. El nostre objectiu era fixar-nos en el tipus de matemàtiques que una majoria important d'estudiants universitaris de primer curs han assolit durant el batxillerat, o assoleixen durant el primer curs de carrera.

Aquesta recopilació, que va dur-se a terme durant el curs acadèmic 2016-2017, va donar lloc a una llista de 319 entrades ja existents de la Viquipèdia, divididas en les matèries d'aritmètica i àlgebra (76), geometria i àlgebra lineal (89), càlcul (96) i probabilitat i estadística (58). Amb aquesta llista vam centrar una mica els nostres objectius, que es van concretar a revisar i mantenir-ne la qualitat, així com a fer un seguiment del nombre d'accessos a aquestes pàgines.

Pensem que la franja d'edat, diguem, dels setze als dinou anys, és la que pot ser la usuària més important de la Viquipèdia en català. Creiem que els estudis més avançats fàcilment tendiran a utilitzar la Wikipedia en anglès. També creiem que el català sempre serà un idioma, fins i tot en aquest nivell intermedi, en competència amb el castellà, ja que tots els usuaris seran, en termes generals, bilingües.

De cadascuna d'aquestes entrades en vam fer una anàlisi general, principalment destinada a detectar errades o mancances greus, i una anàlisi bibliomètrica, basada en comparacions dels accessos als mateixos termes en català, castellà i anglès.

Cal dir que no van aparèixer ni errades ni mancances greus. Però la Societat Catalana de Matemàtiques va acollir la celebració de cinc viquimaratons entre el

març del 2017 i el novembre del 2018 destinades a reflexionar sobre aquests articles més importants i fer-hi modificacions i adaptacions.

En la taula següent hi ha les dades de longitud de l'entrada i de nombre de visites diàries en català, espanyol i anglès de les vint primeres entrades, ordenades pel nombre de visites diàries en anglès.

TAULA 1. *Longitud i visites diàries de les entrades principals*

Entrades	*Longitud*			*Visites diàries*		
	Català	*Espanyol*	*Anglès*	*Català*	*Espanyol*	*Anglès*
Desviació tipus	32.538	7.545	49.971	11	1.983	10.163
Distribució normal	9.635	58.506	131.800	6	1.631	7.518
Nombre primer	62.711	80.433	76.022	59	4.808	5.074
Variància	2.227	9.947	49.732	10	1.549	4.250
Distribució de Poisson	7.979	9.878	60.854	3	839	4.187
Sèrie de Taylor	35.434	13.929	39.051	6	603	3.652
R (llenguatge)	14.830	32.865	48.649	3	307	3.388
Distribució binomial	5.406	5.541	36.087	3	855	3.327
Interval de confiança	1.464	15.516	57.461	4	693	3.218
Estadística	29.219	44.092	62.145	18	3.925	3.171
Nombre *e*	4.781	47.896	37.491	12	1.220	3.159
Logaritme	47.359	43.710	91.062	14	1.976	3.127
Teorema de Bayes	8.507	5.773	29.785	4	478	3.075
Equació de segon grau	20.102	13.299	51.099	27	2.131	2.712
Producte vectorial	6.126	14.059	69.425	4	623	2.694
Matriu (matemàtiques)	21.798	38.528	104.549	8	1.203	2.670
Producte escalar	6.404	16.841	20.349	6	648	2.579
Regressió lineal	14.488	18.224	63.016	6	999	2.562
Distribució exponencial	2.501	4.787	32.150	1	358	2.547
Funció trigonomètrica	59.656	19.788	63.850	17	2.743	2.538

FONT: «Matemàtica intermèdia», IEC, Secció de Ciències i Tecnologia, 2017.

Per a interpretar aquestes dades i fer comparacions entre els idiomes, podem dir, en nombres rodons, que les proporcions de parlants entre els tres idiomes, en tot el món són d'1 a 50 i a 100 (català, espanyol i anglès).

També per a interpretar-les convé adonar-se de la gran importància de l'estadística per als angloparlants. Això ens podria fer reflexionar sobre, potser, algunes mancances del nostre sistema educatiu.

1.2. *Stet the author's choice*

Stet és una conjugació llatina del verb *stare*, que en alguna de les seves accepcions vol dir 'deixa-ho com està' o 'permet'. S'utilitza molt en el que en anglès es diu *obelism*, que desconec si té traducció catalana, i que són les abreviatures habituals en anglès dels correctors de textos, ja sigui per a correccions gramaticals, terminològiques o tipogràfiques, i usualment forma part d'un diàleg entre autor, editor i correctors per a dir 'deixa-ho com estava'.

O sigui, «respecta el redactat o la tipografia originals», sense els canvis. *Stet the author's choice*, diu el manual d'estil de la prestigiosa American Mathematical Society [1], *as long as it is consistent inside the given text*.

Com que avui en dia l'edició de textos matemàtics és gairebé sempre autoedició, i moltes vegades usa el llenguatge LaTex, que no és gaire popular fora dels matemàtics i els físics, és a dir, fora dels que utilitzen moltes fórmules, *stet* acaba volent dir que sigui respectada l'elecció inicial de l'autor.

Potser perquè jo sempre m'he situat, excepte en poquíssimes ocasions, a la banda de l'autor, aquesta actitud de respectar l'opinió inicial de l'autor és la que jo prefereixo. Mentre sigui coherent, és clar, i l'autor no faci canvis no justificables al llarg del text.

Aprofito per a fer primer una defensa de l'American Mathematical Society (AMS), com a principal societat matemàtica que hi ha al món i, en general, dels seus criteris. I en particular del seu llibre d'estil [1]. Això no vol dir que no hi hagi altres societats matemàtiques importants en el món, com la European Mathematical Society (EMS) o la nord-americana Society for Industrial and Applied Mathematics (SIAM). L'EMS és una importantíssima societat de societats, i en formen part les societats europees nacionals, com ara la Societat Catalana de Matemàtiques i la Real Sociedad Matemática Española, entre moltes d'altres. Però l'EMS no té un llibre d'estil, almenys que jo conegui.

La SIAM, en canvi, sí que en té [4]. Però en molts punts s'acaba referint al manual de l'AMS, i això em reafirma a acceptar el llibre d'estil de l'AMS com el llibre d'estil en matemàtiques més rellevant del món. Recomano als correctors de textos matemàtics, ja siguin en l'àmbit terminològic, gramatical, ortogràfic o tipogràfic, que hi donin una mirada de tant en tant, quan tinguin dubtes. I demano, si

és possible, que es mantinguin, com diu aquest manual, en l'estil dels autors. *Stet the author's choice*.

Darrerament he tingut algunes converses amb un expert en tipografia (a més d'expert lingüista) sobre si l'operador derivada s'ha d'escriure en lletra rodona —que els anglesos anomenen *romana*— o en lletra cursiva —que anomenen *itàlica*. Les converses han estat vertaderament interessants. Hi ha una tradició en matemàtiques que diu que les variables s'escriuen en cursiva, mentre que els operadors s'escriuen en rodona. Però jo he escrit sempre la derivada en cursiva. I com jo, la immensa majoria, diguem el 95 %, dels textos que llegeixo. A més, l'operador derivada no és a la llista que hi ha a l'esmentat llibre d'estil de l'AMS on es recullen els operadors que normalment s'escriuen amb lletra rodona. I en tots els exemples que he trobat en aquest llibre, exemples d'altres coses, però que hi surt la derivada, hi apareix en cursiva. Amb una excepció, que honradament haig de dir: apareix en una ocasió una forma diferencial simplèctica, amb motiu del *teorema de Darboux*, i hi surt amb la *d* rodona.

També hem parlat sobre la tipografia dels nombres *e*, *pi* i la unitat imaginària *i*. Com he dit, les variables en matemàtiques s'acostumen a escriure en cursiva. Potser per això les constants s'escriguin en rodona, principalment en textos de física i de química. Pròpiament, en textos matemàtics és difícil que apareguin constants universals, llevat de les tres esmentades i potser la gamma minúscula d'Euler, enteses amb el mateix sentit que en física o química poden tenir la constant de Boltzmann, la constant d'Avogadro, la velocitat de la llum o tantes d'altres. Jo sempre he escrit *e*, *pi* i *i* en cursiva, i així ho he vist escrit gairebé sempre. Amb això no vull pas dir ni molt menys que escriure-les en rodona s'hagi de desaconsellar. Però sí que vull dir que escriure-les en cursiva no hauria de ser motiu de correcció. *Stet the author's choice*.

Aprofito per a fer una referència genèrica a la correcció de textos en català. El català és la meva llengua materna, i en ella he llegit, escoltat, parlat i escrit sempre que ha tingut sentit de fer-ho, que ha estat molt sovint, però no sempre. La recerca matemàtica, i cada cop més la docència especialitzada en l'àmbit universitari, es fan sempre en anglès, i això no veig que hagi de canviar en els propers anys. Però degut a les meves mancances en llengua catalana soc tot sovint un usuari agraït dels serveis dels correctors lingüístics. Reitero aquest agraïment davant de la SCATERM, que alguna cosa hi té a veure, amb la correcció lingüística en català, sens dubte.

Però demano que se'm deixi fer una crida a la moderació. *Stet the author's choice*, per favor. He publicat al llarg de la meva vida professional força textos matemàtics en revistes prestigioses, quasi sempre en anglès. Com que el nivell del meu anglès és força deficient, no puc ni imaginar que el meu nivell de català sigui pitjor. I haig de dir que mai he tingut en els meus textos tantes correccions de tota mena, vull

dir ortogràfiques, gramaticals i tipogràfiques, com quan he escrit textos en català. Això no hauria de ser, o acabarem amb les matemàtiques en català. Per sempre.

Permeteu-me una anècdota. En matemàtiques és molt freqüent donar noms de persones a les fórmules o als resultats: el mètode d'Euler, el problema de Cauchy, la fórmula de Gauss-Bonnet, el teorema de Noether. L'esment d'aquests noms és ben sabut que es refereix a objectes matemàtics precisos, però no a escrits precisos d'aquests autors. És ben conegut l'acudit entre matemàtics que afirma que en matemàtiques quasi cap afirmació que porti el nom d'una persona pot atribuir-se directament a aquesta persona, i que això val també per a aquesta mateixa afirmació i qui la fa.

Doncs bé, en certa ocasió algú va escriure un text de matemàtiques en català i el corrector devia tenir instruccions per part de l'editor que la primera vegada que un autor és citat en un text ha d'aparèixer amb el nom complet i l'any de naixement i mort. Això fa que *el problema de Cauchy*, que és una cosa molt concreta i freqüent en equacions diferencials, es converteixi en *el problema d'Agustin-Louis Cauchy (1789-1857)*, una cosa que, com que cap matemàtic del món pot reconèixer, finalment sembla que posi de manifest que l'autor és un ignorant o un pedant. I, com tots sabem, la petita (o gran) vanitat de l'autor és la darrera cosa que un corrector s'ha d'atrevir a reptar, si no és que està disposat a suportar les respostes irritades del corregit.

Fem, per favor, i especialment en català, que el diàleg entre autors, editors i correctors sigui sempre amistós. I, si pot ser, *stet the author's choice*.

1.3. *Un exemple: el binomi de Newton*

M'agradaria analitzar des del punt de vista terminològic l'entrada «binomi de Newton» de la Viquipèdia. Per començar, voldria dir en paraules el que diu la fórmula

$$\left(x+a\right)^n = \sum_{k=0}^{n} \binom{n}{k} x^k a^{n-k}.$$

Diu així: «*x* més *a* <u>*elevat*</u> a *n* és igual a la <u>*suma*</u> per a *k* des de zero fins a *n* del coeficient <u>*n sobre k*</u> multiplicat per *x* elevat a *k* i per *a* elevat a *n* menys *k*.» I ara, comentem un per un aquests termes:

— <u>*elevat*</u>: fa referència a la potenciació. Respecte a la potenciació, podem llegir a la Wikipedia les expressions següents, en quatre idiomes diferents:

Català: b^n es pot llegir de les següents maneres: «*b* elevat a *n*», «*b* elevat a la *n*-èsima potència» o, de manera curta, «*b* a la *n*».

Espanyol: a^n se lee normalmente como «*a* elevado a la *n*».

Anglès: b^n is called "*b* raised to the *n*-th power", "*b* raised to the power of *n*", "the *n*-th power of *b*", "*b* to the *n*th", or most briefly as "*b* to the *n*".

Francès: a^n se lit «*a* puissance *n*» ou «*a* exposant *n*».

Com veiem, sempre hi ha diverses maneres de dir-ho. La meva opinió és que no és necessàriament bo tenir una única manera de dir les coses, i que és preferible que l'elecció del qui escriu el text introdueixi el matís que cregui convenient. Faig notar que la Wikipedia m'ha estat molt útil per a trobar aquestes expressions escrites en diversos idiomes.

També faig notar que la traducció automàtica de textos matemàtics, a la qual em referiré més avall, presenta problemes verdaderament difícils de resoldre.

I finalment observo l'ús de cometes baixes o *guillemets* (« ») en certes llengües en contraposició a l'ús de cometes altes (" "), molt típic de l'anglès. Potser per aquest motiu els matemàtics usem les cometes altes, preferentment, perquè el llenguatge LaTex ens hi porta de manera més natural.

— *Suma* o *sumatori*: la lletra sigma majúscula es refereix a una suma indexada, i el conjunt en el qual varien els índexs es posa al voltant d'aquesta lletra, de vegades, com aquí, en què les indicacions apareixen com a subíndexs i superíndexs de la sigma majúscula, de vegades, totes les indicacions completament a sota de la sigma o, de vegades, la indicació també en dues parts, però una part completament a sota i l'altra part completament a sobre. En LaTex no és difícil d'escriure-ho d'una o manera o l'altra.

Quan aquesta sigma majúscula s'usa indicant una suma, moltes vegades el signe s'anomena *sumatori*. Fa la sensació que aquesta distinció és la que està implícita en la definició de *sumatori* que hi ha a la segona edició del *Diccionari de la llengua catalana* de l'IEC (DIEC2) [2]. Però és difícil distingir en aquest cas entre el signe i el concepte. El signe seria la sigma majúscula, i el concepte una suma múltiple, indexada. Per aquesta raó molts matemàtics haurien llegit la fórmula de dalt dient «el sumatori per a *k* des de zero fins a *n*», a diferència del que he fet jo, que he dit «la suma per a *k* des de zero fins a *n*».

— *n sobre k*: aquesta és la manera habitual d'anomenar aquest nombre combinatori. Recordem que

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}.$$

Aquesta accepció de la paraula *sobre* no es troba al DIEC2, i potser tampoc és tan necessari que hi sigui. Però té un ús freqüent en matemàtiques. En castellà també es diu *n sobre k*, però, a causa del seu significat combinatori, en anglès es llegeix *n choose k* i en francès, *k parmi n*.

Però el meu comentari terminològic principal sobre l'entrada «binomi de Newton» de la Viquipèdia és precisament sobre el títol de l'entrada. És ben sabut que aquesta fórmula no és deguda a Newton, sinó que és molt més antiga. La contribució de Newton va ser precisament la de considerar els casos en què n no és un nombre enter, sinó fraccionari. En aquest cas la suma finita es converteix en una suma infinita, per a k des de zero fins a infinit, amb la interpretació apropiada del coeficient combinatori corresponent. Tot i això, tant en català com en castellà aquest nom està ben establert, encara que sigui inapropiat. Pel que fa al català, així ho veiem a la Viquipèdia, però també al *Diccionari de matemàtiques i estadística* (DME) [3] i al TERMCAT [5]. En castellà, però, la Wikipedia en dona el nom molt més correcte de *teorema del binomio*.

La meva opinió particular és que hauria d'anomenar-se *fórmula del binomi*, perquè posar èmfasi en la paraula *teorema* em sembla poc apropiat per a una cosa tan pràctica. Però la terminologia matemàtica no serà mai la que digui un especialista o l'altre, sinó aquella que s'usi entre la gent que la utilitza. En tot cas, per aquesta raó em vaig permetre de canviar les primeres línies de l'article «binomi de Newton» a la Viquipèdia, i ara hi posa la solució que en aquell moment em va semblar millor: «El binomi de Newton o teorema del binomi és una fórmula que…».

Per cert, dues observacions sobre això. La primera és que si a la Wikipedia en espanyol es busca «binomio de Newton» no diu pas que aquesta pàgina no existeix, sinó que redirigeix automàticament a «teorema del binomio». L'altra observació és remarcar el títol de la pàgina en francès: ni més ni menys que «formule du binôme de Newton».

Com a conclusió: no sé si hem d'aspirar a una terminologia molt fixa en matemàtiques, ni tan sols en les coses elementals. La meva opinió ja l'he dit abans: *stet the author's choice*.

1.4. *Altres temes de «matemàtiques en català»*

Seria molt útil poder disposar d'un bon corrector ortogràfic de català adaptat al LaTex. Com ja he dit, el LaTex ha esdevingut el llenguatge en el qual els matemàtics i físics escriuen principalment els seus textos. Això és especialment cert en l'àmbit universitari i en l'àmbit de la recerca. Però també és cert que en altres àmbits s'usen altres processadors de textos, però amb resultats que jo, personalment, considero menys satisfactoris que els que s'obtenen amb el LaTex.

Igualment, haig de dir que la recerca matemàtica, i cada cop una part més gran de la docència, s'escriu en anglès. El LaTex està molt ben adaptat a l'anglès, perquè va crear-se en aquesta llengua. Això no vol dir que no es pugui escriure perfectament amb grafia catalana, però d'una manera que els correctors ortogràfics en català tenen dificultats per a processar.

Poso com a exemple un petit text matemàtic en català, extret de la solució d'un problema en una assignatura del grau de matemàtiques:

```
{\bf Soluci\'{o}}: {\sl La definici\'{o} apropiada de $A$ \'{e}s
\centerline{$A((u^1,v^1),(u^2,v^2))=
\int_0^1 (u^1_xu^2_x+v^1_xv^2_x+ru^1u^2+su^1v^2+rv^1v^2+su^2v^1)\ dx$}
perqu\`{e} prenent $(\phi,\psi)$ amb $\psi=0$ queda la definici\'{o} habitual de
soluci\'{o} feble de la primera de les dues equacions, i prenent $\phi=0$ queda la
definici\'{o} habitual per a la segona equaci\'{o}. Est\`{a} clar que \'{e}s bilineal cont\'{\i}
nua a $H$.
A fi que sigui coerciva es necessita que la forma quadr\`{a}tica de $\R^2$ dona-
da per $rx_1^2+2sx_1x_2+rx_2^2$ sigui definida positiva. Analitzant els valors pro-
pis, es veu que aix\`{o} passa si $r>0$ i $|s|<r$. El teorema de Lax-Milgram pot apli-
car-se perqu\`{e} $(\phi,\psi)\mapsto \int_0^1f\phi+g\psi$ tamb\'{e} \'{e}s cont\'{\i}nua
sobre $H$ quan $f,g\in L^2(0,1)$.}
```

Aquí es veu que el LaTex té formes pròpies per a escriure accents, i convindria tenir un corrector ortogràfic que sabés llegir-les automàticament i proposar les correccions adients. El corrector hauria de saber prescindir de les indicacions tipogràfiques, com ara les «\bf», «\sl» o «\centerline» o d'altres que apareixen en el text i de totes les expressions matemàtiques, que en LaTex apareixen entre dos signes $.

Naturalment, seria encara millor si aquest corrector ortogràfic fes també suggeriments gramaticals, com fan els millors correctors automàtics. Per a saber fer-ho hauria de ser capaç d'interpretar les fórmules matemàtiques entre signes $ com a equivalents a expressions gramaticals. Si això arribés a ser possible, podríem també encarar el problema de la traducció automàtica de textos matemàtics, que es fa necessària principalment entre el català i l'anglès, però també entre el català i el castellà.

Cal esmentar el traductor Apertium, accessible a *https://apertium.org*, el qual accepta documents en LaTex. És una eina molt útil, però que pel que fa a textos matemàtics necessitaria encara algunes millores.

Un altre tema ben diferent que convé que sigui conegut per la seva importància terminològica és que l'IEC posarà, com a diccionari en línia dins de la seva col·lecció de diccionaris de ciència i tecnologia, una nova edició del *Diccionari de matemàtiques i estadística* [3]. Per aquest motiu, ha signat recentment un conveni amb els propietaris dels drets d'edició, que són la Universitat Politècnica de Catalunya i Enciclopèdia Catalana.

Aquesta edició en línia del DME serà una eina terminològica molt útil, que augmentarà els recursos terminològics accessibles pel web. Els diccionaris en línia

permeten consultes creuades i associacions de termes, que tenen molta importància en la discussió terminològica.

A més, el conveni autoritza que s'actualitzin, es millorin i es completin els seus continguts, i això permetrà d'anar-lo perfeccionant de mica en mica. És d'esperar, doncs, que les discussions i estudis que generi l'actualització d'aquest diccionari tinguin una transcendència útil per a tota l'expressió matemàtica en català.

El DME és pròpiament un diccionari amb força contingut enciclopèdic, especialment pel que fa a biografies de matemàtics. La referència a matemàtics illustres és, sens dubte, també d'interès terminològic perquè, com ja hem dit abans, molts objectes i resultats matemàtics tenen els noms dels matemàtics que els van crear, o a qui s'atribueixen.

En relació amb això darrer, convido a qui hi estigui interessat que consulti les grafies dels noms de matemàtics importants que apareixen en una llista en un apèndix del llibre d'estil de l'AMS [1].

Finalment, voldria deixar entre les feines pendents la continuació de les viquimaratons temàtiques al voltant de conceptes i termes matemàtics. Els projectes endegats des de la Societat Catalana de Matemàtiques han tingut prou ressò, però haurien de continuar. Això és també una invitació a totes les persones interessades perquè continuïn col·laborant amb la Viquipèdia i mantenint i millorant-ne el contingut.

Bibliografia

[1] Letourneau, Mary; Wright Sharp, Jennifer (2017). *AMS Style Guide* [en línia]. American Mathematical Society. <http://www.ams.org/arc/styleguide/index.html>.

[2] Institut d'Estudis Catalans (2007). *Diccionari de la llengua catalana*. 2a ed. Barcelona: IEC. També disponible en línia a: <http://dlc.iec.cat/>.

[3] Mateu, Rosa; Torras, Montserrat (coord.) (2002). *Diccionari de matemàtiques i estadística*. Barcelona: Universitat Politècnica de Catalunya: Enciclopèdia Catalana.

[4] Society for Industrial and Applied Mathematics (2013). *SIAM Style Manual* [en línia]. <https://www.siam.org/journals/pdf/stylemanual.pdf>.

[5] TERMCAT. *Cercaterm* [en línia]. Barcelona: TERMCAT. <https://www.termcat.cat/ca>.

Joan de Solà-Morales Rubió
Secció de Ciències i Tecnologia de l'Institut d'Estudis Catalans
Societat Catalana de Matemàtiques
Universitat Politècnica de Catalunya

2. Experiències de les societats filials de l'IEC en la Viquipèdia: aportacions des de la Societat Catalana de Química[2]

En el marc de la visualització de la relació entre la Viquipèdia i la terminologia, aquesta aportació se centra en la situació d'aquesta relació pel que fa a la terminologia química.

Amb aquesta visió, començarem situant els referents de la terminologia química i revisant les traduccions disponibles en català. En un segon moment, explorarem el rol de la Viquipèdia en relació amb la química i la seva terminologia. Tancarem l'exposició visualitzant quin rol han tingut els projectes promoguts des de l'IEC i Amical Wikimedia en els darrers anys pel que fa al contingut en química a la Viquipèdia catalana.

2.1. *Terminologia i química*

Fa set anys, a la setena edició del Seminari de Terminologia, el doctor Salvador Alegret (Alegret, 2013) presentava amb força detall el rol que la Unió Internacional de Química Pura i Aplicada (IUPAC) i els seus llibres de colors tenen pel que fa a la terminologia química. No repetiré el seu treball, que podeu llegir en línia, però deixeu-me començar situant l'estat actual d'aquests referents (a maig de 2019). Els vuit llibres de colors de la IUPAC, juntament amb les seves diferents edicions, es mostren recollits en la taula 1. Des del 2012, diversos dels llibres s'han actualitzat, la majoria són accessibles en línia i molts d'ells tenen projectes actius treballant en les edicions següents. Resumint el que podem observar en aquesta taula, la terminologia química és molt extensa, té diversos milers de pàgines i diverses desenes de milers de termes, i la IUPAC i la comunitat científica fan esforços significatius per mantenir-la actualitzada i accessible.

Com feia el doctor Alegret (2013) llavors, mirem ara les traduccions d'aquests documents a la nostra llengua, que es mostren en la taula 2. Com podem observar, i malgrat l'esforç fet per a disposar de traduccions en català d'aquests documents, només quatre dels vuit llibres de colors tenen traduccions al català. A més, la darrera edició anglesa només està traduïda en el cas del llibre taronja, i només hi ha edicions electròniques del llibre blau i del llibre verd.

Abans d'entrar en el rol de la Viquipèdia, deixeu-me esmentar altres recursos terminològics en l'àmbit de la química, com els que han estat promoguts per l'IEC i estan recollits al portal CiT (Ciències i Tecnologia) (*https://cit.iec.cat/*) o els recollits a la biblioteca terminològica del TERMCAT (*https://www.termcat.cat/ca/*

2. Aquest article correspon a la participació de Jordi Cuadros en la XVI Jornada SCATERM: «La Viquipèdia i la terminologia» (30 de maig de 2019). Les anàlisis i les conclusions que s'hi reflecteixen són de l'autor i no són atribuïbles a cap de les institucions en què participa.

biblioteca-en-linia/biblioteca-terminologica/arees-tematiques/Qu%C3%ADmica). En l'àmbit de la química, el CiT recull set obres terminològiques i el TERMCAT, vint-i-sis obres, entre les quals el *Diccionari de química* (Universitat Politècnica de Catalunya, TERMCAT i Enciclopèdia Catalana, 2019), *http://www.termcat.cat/ca/diccionaris-en-linia/212*, reeditat recentment.

Un darrer aspecte a comentar pel que fa a les dimensions de la terminologia química correspon al nombre de productes químics identificats. Aquest mateix 2019, el registre del Chemical Abstracts Service anunciava els cent cinquanta milions de productes registrats (Wang, 2019). Encara que no tots són d'ús comú, aquesta xifra ens dona un altre referent del nombre de termes que es poden arribar a considerar.

Taula 1. *Llibres de colors de la IUPAC (versions originals en anglès)*

Llibre	*Edicions*	*Nombre de pàgines*	*En revisió?*
Compendium of Chemical Terminology (*Gold Book*) Disponible a: <https://goldbook.iupac.org/>	1987 (1a), 1997 (2a) En línia: 2006 (1.0.0), 2009 (2.1.0), 2014 (2.3.3)	1.670 (en el PDF de la versió 2.3.3)	Sí, projecte 2016-046-1-024
Quantities, Units and Symbols in Physical Chemistry (*Green Book*) 3a edició en PDF disponible a: <https://iupac.org/wp-content/uploads/2019/05/IUPAC-GB3-2012-2ndPrinting-PDFsearchable.pdf>	1981 (1a), 1993 (2a), 2007 (3a)	250 (en el PDF de la 3a edició, 2a reimpressió)	No hi consta
Nomenclature of Inorganic Chemistry (*Red Book*) Disponible a: <https://iupac.org/wp-content/uploads/2016/07/Red_Book_2005.pdf >	1959 (1a), 1971 (2a), 1990 (*Red Book I*), 2000 (*Red Book II*), 2005	377 (en el PDF de l'edició del 2005)	Sí, en projectes relacionats[3]

3. Per exemple, els projectes 2006-038-1-800, *Preferred IUPAC Names (PINs) for Inorganic Compounds*, i 2017-033-1-800, *Alignment of Principles For Specifying Ligands And Substituent Groups Across Various Areas of Nomenclature*.

TAULA 1. *Llibres de colors de la IUPAC (versions originals en anglès) (Continuació)*

Llibre	*Edicions*	*Nombre de pàgines*	*En revisió?*
Nomenclature of Organic Chemistry (*Blue Book*) Edicions de 1979 i de 1993 disponibles a: <https://www.acdlabs.com/iupac/nomenclature/> Un esborrany de la tercera edició està disponible a: <http://old.iupac.org/reports/provisional/abstract04/favre_310305.html>	1979 (1a), 1993 (2a), 2013 (3a)	1.612	Sí, en projectes relacionats
Compendium of Polymer Terminology and Nomenclature (*Purple Book*) Disponible a: <https://iupac.org/cms/wp-content/uploads/2016/07/ONLINE-IUPAC-PB2-Online-June2014.pdf>	1991 (1a), 2009 (2a)	465 (en el PDF de la segona edició)	No hi consta
Compendium of Analytical Nomenclature (*Orange Book*) Disponible a: <https://media.iupac.org/publications/analytical_compendium/>	1977 (1a), 1987 (2a), 1998 (3a)	964	Sí, projecte 2012-005-1-500
Compendium of Terminology and Nomenclature of Properties in Clinical Laboratory Sciences (*Silver Book*) Disponible a: <https://pubs.rsc.org/en/content/ebook/978-1-78262-107-2>	1995 (1a), 2017 (2a)	182	No hi consta
Biochemical Nomenclature and Related Documents (*White Book*)	1978 (1a), 1992 (2a)	347	No hi consta

FONT: Elaboració pròpia.

TAULA 2. *Llibres de colors de la IUPAC en català*

Llibre	*Correspondència amb l'edició original*
Magnituds, unitats i símbols en química física. Unió Internacional de Química Pura i Aplicada. Versió catalana de la segona edició anglesa a cura de Josep M. Costa. Barcelona: Institut d'Estudis Catalans, 2004. *Magnituds, unitats i símbols en química física* [recurs electrònic]. Unió Internacional de Química Pura i Aplicada. Versió catalana de la segona edició anglesa a cura de Josep M. Costa. 2a ed., corregida. Barcelona: Institut d'Estudis Catalans, 2008. <http://publicacions.iec.cat/repository/pdf/00000049%5C00000040.PDF>. *Magnituds, unitats i símbols en química física* [en línia]. Unió Internacional de Química Pura i Aplicada. Versió catalana de la segona edició anglesa a cura de Josep M. Costa. 2a ed., corregida. Barcelona: Institut d'Estudis Catalans, febrer 2009. <http://cit.iec.cat/QUIMFIS>.	~~1981 (1a),~~ 1993 (2a), ~~2007 (3a)~~
Nomenclatura de química inorgànica: *Recomanacions de 1990.* Unió Internacional de Química Pura i Aplicada. Versió catalana per Enric Casassas i Simó i Joaquim Sales i Cabré. Barcelona: Institut d'Estudis Catalans, 1997.	~~1959 (1a), 1971 (2a),~~ 1990 (Red Book I), ~~2000 (Red Book II), 2005~~
Nomenclatura de química orgànica: *Seccions A, B i C*: *Regles definitives de 1979.* Unió Internacional de Química Pura i Aplicada. Divisió de Química Orgànica. Edició a cura d'Àngel Messeguer i Peypoch i Miquel A. Pericàs i Brondo. Barcelona: Consell Superior d'Investigacions Científiques: Institut d'Estudis Catalans, 1989. *Nomenclatura de química orgànica [recurs electrònic]*: *Seccions A, B, C i H*: *Regles definitives de 1979.* Unió Internacional de Química Pura i Aplicada. 2a ed., corr. i ampl. Barcelona: Institut d'Estudis Catalans: Consell Superior d'Investigacions Científiques, 2013. <http://publicacions.iec.cat/repository/pdf/00000195/00000013.pdf>. *Guia de la IUPAC per a la nomenclatura de compostos orgànics*: *Recomanacions de 1993 (incloent-hi les revisions, tant publicades com no publicades fins ara, de l'edició del 1979 de la Nomenclature of organic chemistry).* Unió Internacional de Química Pura i Aplicada. Divisió de Química Orgànica. Comissió de Nomenclatura de Química Orgànica (2017). Versió catalana a cura d'Àngel Messeguer. Barcelona: Institut d'Estudis Catalans, 2017. <https://publicacions.iec.cat/repository/pdf/00000241/00000059.pdf>.	1979 (1a), 1993 (2a), ~~2013 (3a)~~

TAULA 2. *Llibres de colors de la IUPAC en català (Continuació)*

Llibre	*Correspondència amb l'edició original*
Compendi de nomenclatura de química analítica: *Regles definitives de 1977*. Unió Internacional de Química Pura i Aplicada. Divisió de Química Orgànica. Edició a cura d'Enric Casassas i Salvador Alegret. Barcelona: Institut d'Estudis Catalans, 1987. (Monografies de la Secció de Ciències; 4) *Compendi de nomenclatura de química analítica*: *Regles definitives de 1997*. Unió Internacional de Química Pura i Aplicada. Edició a cura d'Enric Casassas, Elisabeth Bosch i Salvador Alegret. Barcelona: Institut d'Estudis Catalans, 2007. 3 v.	1977 (1a), ~~1987 (2a)~~, 1998 (3a)

NOTA: En la columna «Correspondència amb l'edició original», les edicions ratllades indiquen que no han estat traduïdes al català.
FONT: Elaboració pròpia.

2.2. *Química a la Viquipèdia*

Vist el panorama del que són els pilars de la terminologia química, quin és el rol que pot tenir la Viquipèdia en aquest àmbit?

Podem pensar en dues funcions que la Viquipèdia pot assumir respecte a la terminologia química; com a font o com a forma de difusió i normalització terminològica.

En qualsevol dels dos casos, la Viquipèdia serà un recurs vàlid, és a dir, a potenciar i a promoure si compleix com a mínim els quatre criteris següents:

1. Si els continguts són correctes (tant en els termes com en les definicions).
2. Si el contingut és fàcilment accessible.
3. Si el nombre de termes recollits és extens, com a mínim equiparable al de les obres terminològiques més accessibles.
4. Si els termes s'actualitzen de forma ràpida quan se'n proposen de normalitzats o quan aquests es normativitzen.

A través d'algunes dades i alguns exemples, intentarem acostar-nos a aquests criteris a fi d'avaluar quin valor podem donar a aquest recurs. Proposaré aquestes anàlisis tant per a la Viquipèdia catalana, la realitat actual per a la nostra llengua, com per a la Wikipedia anglesa, com a referent d'un horitzó al qual podríem aspirar d'aquí a uns quants anys.

2.2.1. Correcció: fixem-nos en alguns termes químics

Hi ha una certa diversitat d'estudis que discuteixen, de forma general, la fiabilitat de la Viquipèdia. Un resum d'alguns d'aquests estudis es pot trobar a Cuadros, Dengra i Marginet (2017). Centrem-nos, però, en la terminologia química i fem-ho amb alguns exemples; comparem les definicions dels termes *element químic*, *reacció química* i *pH* de la Wikipedia (en anglès i en català) amb les que apareixen a les fonts normatives catalanes, a partir de la consulta a l'Optimot (*https://aplicacions.llengua.gencat.cat/llc/AppJava/index.html*), i amb les definicions del *Gold Book* de la IUPAC (International Union of Pure and Applied Chemistry, 2017).

a) *Element*

Comencem pel terme *element*, *element químic* o *chemical element*. Les figures 1, 2, 3 i 4 mostren, respectivament, els resultats de la cerca en els quatre recursos esmentats: l'Optimot, que mostra resultats obtinguts dels diccionaris del TERMCAT; la Viquipèdia catalana; la Wikipedia anglesa, i el *Gold Book* de la IUPAC.

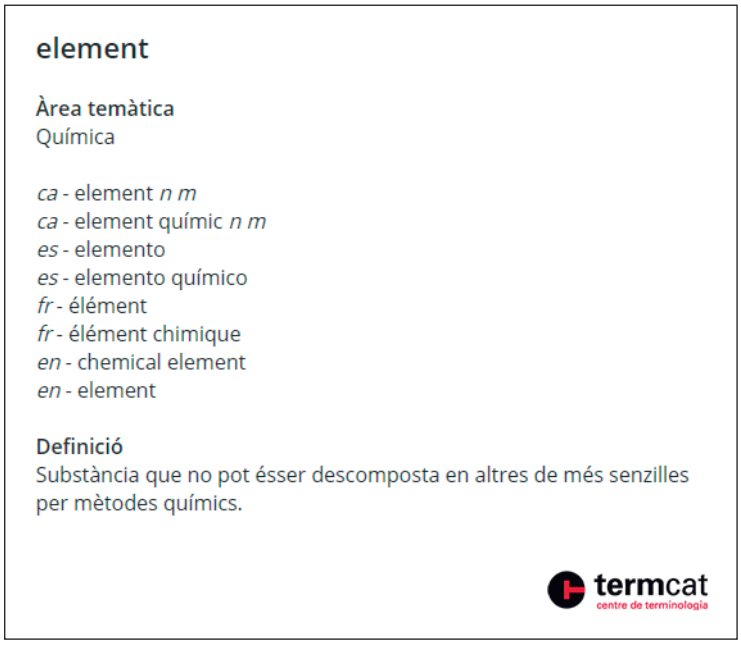
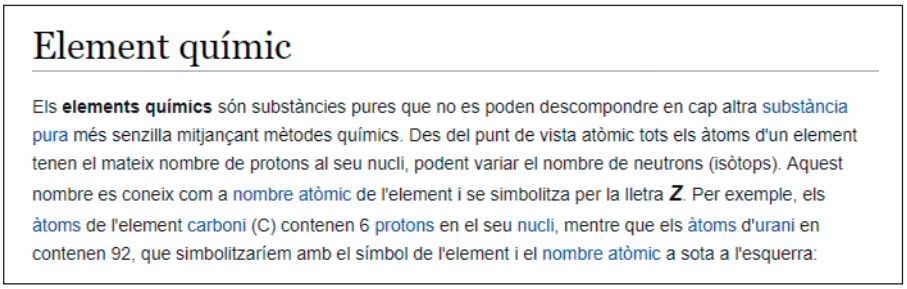


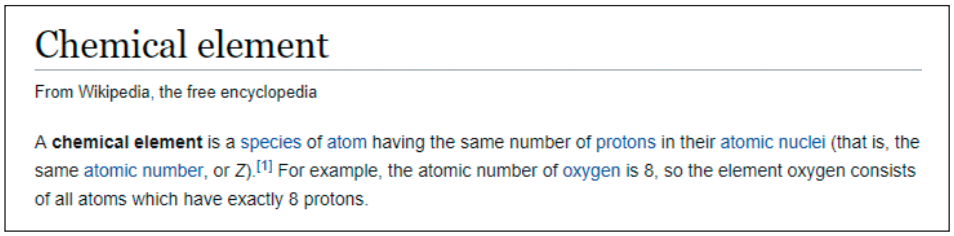
Figura 1. Consulta del terme *element* a l'Optimot.



Element químic

Els **elements químics** són substàncies pures que no es poden descompondre en cap altra substància pura més senzilla mitjançant mètodes químics. Des del punt de vista atòmic tots els àtoms d'un element tenen el mateix nombre de protons al seu nucli, podent variar el nombre de neutrons (isòtops). Aquest nombre es coneix com a nombre atòmic de l'element i se simbolitza per la lletra ***Z***. Per exemple, els àtoms de l'element carboni (C) contenen 6 protons en el seu nucli, mentre que els àtoms d'urani en contenen 92, que simbolitzaríem amb el símbol de l'element i el nombre atòmic a sota a l'esquerra:

FIGURA 2. Consulta del terme *element químic* a la Viquipèdia catalana.

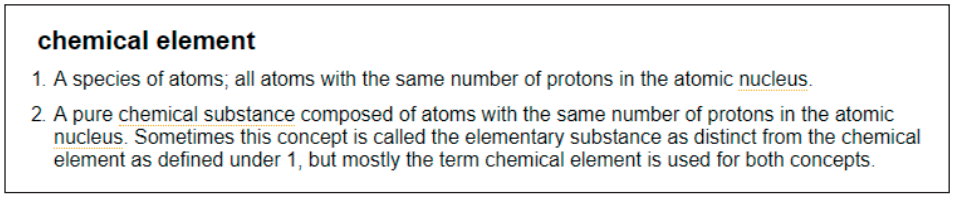


Chemical element

From Wikipedia, the free encyclopedia

A **chemical element** is a species of atom having the same number of protons in their atomic nuclei (that is, the same atomic number, or *Z*).[1] For example, the atomic number of oxygen is 8, so the element oxygen consists of all atoms which have exactly 8 protons.

FIGURA 3. Consulta del terme *chemical element* a la Wikipedia anglesa.



chemical element

1. A species of atoms; all atoms with the same number of protons in the atomic nucleus.
2. A pure chemical substance composed of atoms with the same number of protons in the atomic nucleus. Sometimes this concept is called the elementary substance as distinct from the chemical element as defined under 1, but mostly the term chemical element is used for both concepts.

FIGURA 4. Consulta del terme *chemical element* al *Gold Book*.

La lectura amb cert detall d'aquestes definicions mostra que la més completa és la definició del *Gold Book*, que inclou tant l'accepció submicroscòpica com l'accepció macroscòpica del terme. Curiosament, la Wikipedia anglesa se centra en l'accepció submicroscòpica, mentre que les fonts en català, molt semblants les dues, mostren només la dimensió macroscòpica.

b) *Reacció química*

El mateix treball amb els termes *reacció química* i *chemical reaction* es mostra en les figures 5, 6, 7 i 8. Les observacions que poden fer-se són semblants. Les tres fonts de tipus general, l'Optimot i la Wikipedia (en català i en anglès), presenten definicions de qualitat semblant; totes elles de menys precisió que la que s'obté llegint el *Gold Book*.

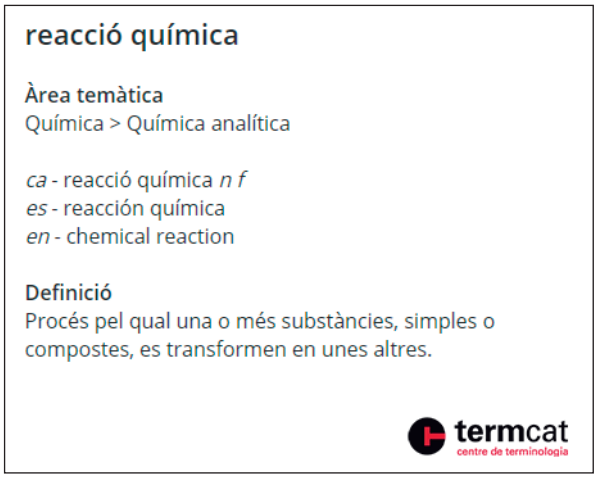


FIGURA 5. Consulta del terme *reacció química* a l'Optimot.

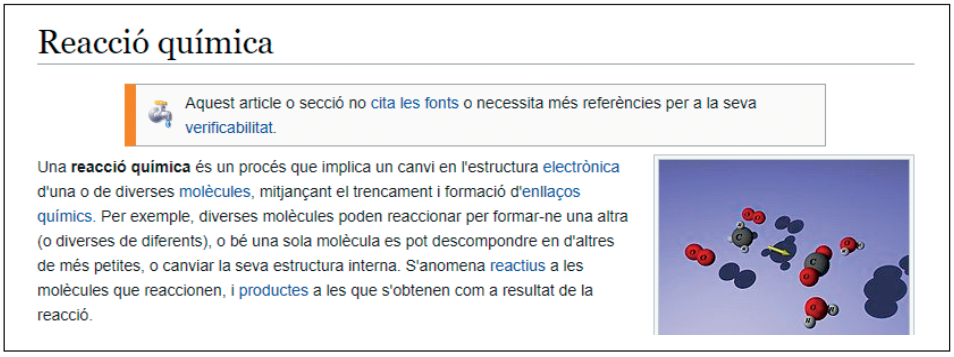


FIGURA 6. Consulta del terme *reacció química* a la Viquipèdia catalana.

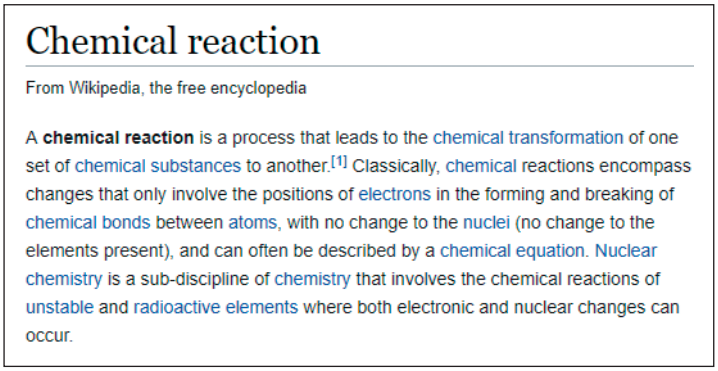


FIGURA 7. Consulta del terme *chemical reaction* a la Wikipedia anglesa.

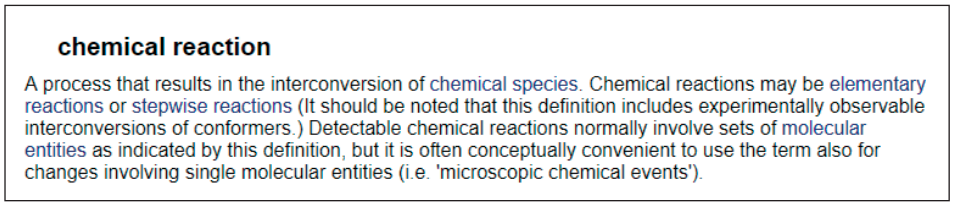


FIGURA 8. Consulta del terme *chemical reaction* al *Gold Book*.

c) *pH*

També en el cas del *pH*, el tercer cas estudiat (figures 9, 10, 11 i 12), les definicions són semblants en les quatre fonts. Dit d'una altra manera, les tres fonts generals consultades mostren definicions prou properes a la del *Gold Book*.

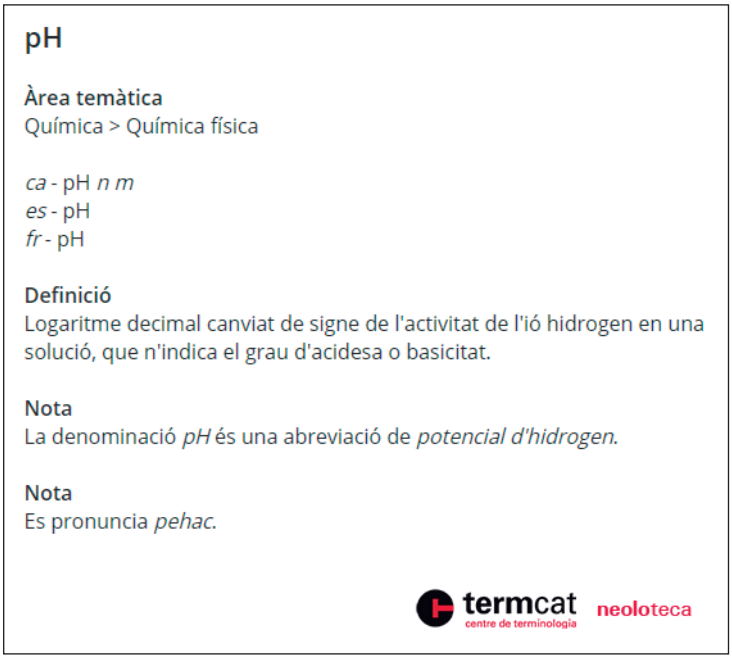
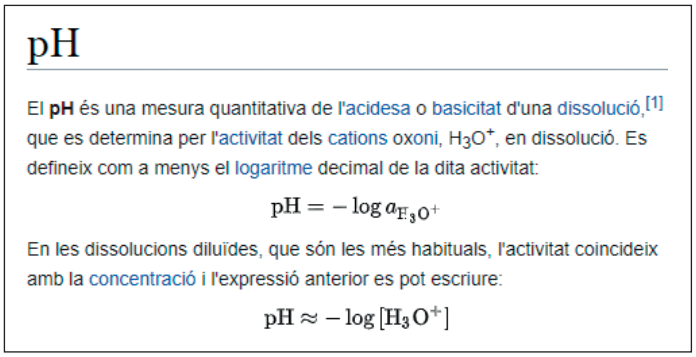


FIGURA 9. Consulta del terme *pH* a l'Optimot.



pH

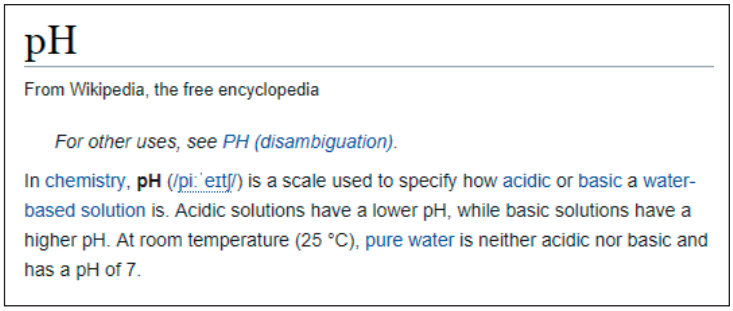
El **pH** és una mesura quantitativa de l'acidesa o basicitat d'una dissolució,[1] que es determina per l'activitat dels cations oxoni, H_3O^+, en dissolució. Es defineix com a menys el logaritme decimal de la dita activitat:

$$\mathrm{pH} = -\log a_{\mathrm{H_3O^+}}$$

En les dissolucions diluïdes, que són les més habituals, l'activitat coincideix amb la concentració i l'expressió anterior es pot escriure:

$$\mathrm{pH} \approx -\log [\mathrm{H_3O^+}]$$

FIGURA 10. Consulta del terme *pH* a la Viquipèdia catalana.



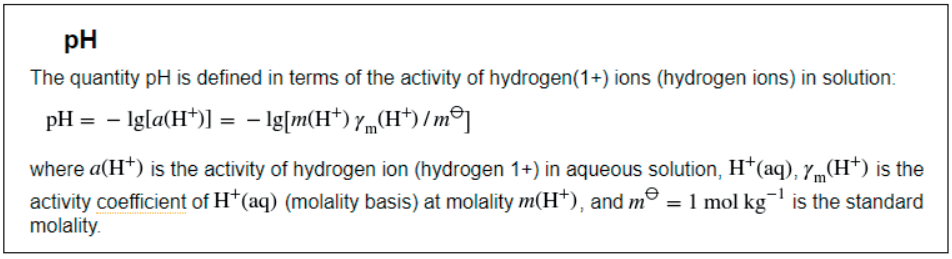
pH

From Wikipedia, the free encyclopedia

For other uses, see PH (disambiguation).

In chemistry, **pH** (/ˌpiːˈeɪtʃ/) is a scale used to specify how acidic or basic a water-based solution is. Acidic solutions have a lower pH, while basic solutions have a higher pH. At room temperature (25 °C), pure water is neither acidic nor basic and has a pH of 7.

FIGURA 11. Consulta del terme *pH* a la Wikipedia anglesa.



pH

The quantity pH is defined in terms of the activity of hydrogen(1+) ions (hydrogen ions) in solution:

$$\mathrm{pH} = -\lg[a(\mathrm{H^+})] = -\lg[m(\mathrm{H^+})\,\gamma_{\mathrm{m}}(\mathrm{H^+})\,/\,m^{\ominus}]$$

where $a(\mathrm{H^+})$ is the activity of hydrogen ion (hydrogen 1+) in aqueous solution, $\mathrm{H^+(aq)}$, $\gamma_{\mathrm{m}}(\mathrm{H^+})$ is the activity coefficient of $\mathrm{H^+(aq)}$ (molality basis) at molality $m(\mathrm{H^+})$, and $m^{\ominus} = 1\ \mathrm{mol\ kg^{-1}}$ is the standard molality.

FIGURA 12. Consulta del terme *pH* al *Gold Book*.

Els tres exemples analitzats il·lustren prou bé, en opinió de l'autor, la situació de la Viquipèdia respecte a la qualitat dels continguts terminològics en l'àmbit de la química. Aquests són sovint prou correctes i d'un nivell de qualitat similar al que es troba en les referències terminològiques generals que usem en la nostra llengua.

2.2.2. Accessibilitat: termes de química a Google

Segona idea, comparem l'accessibilitat de la terminologia recollida a la Viquipèdia amb la d'altres fonts terminològiques de referència.

És prou conegut que la Viquipèdia és un recurs obert, accessible a tothom que disposi dels recursos tecnològics necessaris.[4] Però, són igualment accessibles els altres recursos terminològics dels quals disposem?

Tot i que alguns usem el Cercaterm (*https://www.termcat.cat/ca/cercaterm*), l'Optimot (*https://aplicacions.llengua.gencat.cat/llc/AppJava/index.html*) o el CercaCiT (*https://cit.iec.cat/entrada.asp?pagina=10*), el més habitual és que acudim a Google per resoldre els nostres dubtes terminològics. Però, on ens porta Google?

TAULA 3. *Posició en què apareixen els primers resultats de la Viquipèdia catalana, de documents terminològics de l'IEC, dels diccionaris del TERMCAT o de les obres d'Enciclopèdia Catalana, quan se cerquen a Google diferents termes químics. Resultats obtinguts el 29 de maig de 2019, des d'un ordinador amb IP localitzable a la província de Barcelona i usant com a navegador Mozilla Firefox 67.0, sense galetes emmagatzemades*

Terme	*Viquipèdia catalana*	*IEC*	*TERMCAT*	*Enciclopèdia Catalana*
Arsènic	2	> 10	> 10	3
Element químic	1	> 10	> 10	3
Espectroscòpia	2	> 10	> 10	> 10
Naftalè	2	> 10	> 10	4
pH-metre	2	> 10	> 10	8

NOTA: «> 10» indica que no hi ha resultats en la primera pàgina de resultats.
FONT: Elaboració pròpia.

D'acord amb els resultats que s'observen en la taula 3, és clar que Google ens porta a la Viquipèdia catalana quan cerquem terminologia química, amb uns resultats més que meritoris en comparació amb els de les obres d'Enciclopèdia Catalana. Una mínima reflexió sobre aquests resultats ens condueix a pensar a tenir cura de la Viquipèdia catalana i a revisar la visibilitat i l'accessibilitat que tenen els treballs de l'IEC i del TERMCAT.

4. Deixant de banda els problemes d'accés que la censura digital provoca en alguns països.

2.2.3. Extensió: articles de química a la Wikipedia

Tercer aspecte important: té la Wikipedia prou contingut de química? Les anàlisis que segueixen s'han dut a terme la setmana del 27 de maig de 2019, usant l'eina Petscan (Manske, 2019), que es troba a *https://petscan.wmflabs.org/*.

L'anàlisi del nombre d'articles de química a la Viquipèdia catalana, que es mostra en la taula 4, situa el contingut terminològic de química en aquesta Viquipèdia al nivell d'un glossari de mida mitjana, amb unes dimensions que poden equiparar-se amb les de les obres terminològiques especialitzades de les quals disposem en català.

Taula 4. *Resultats de l'anàlisi per a la Viquipèdia catalana*

Cerca	*Nombre de pàgines*	*URL*
Articles dins la categoria «química» (profunditat: 2)	2.085	*https://petscan.wmflabs.org/?language=ca&project=wikipedia&depth=2&categories=Qu%C3%ADmica&ns%5B0%5D=1&search_max_results=500*
i excloent-hi persones, institucions, esdeveniments i publicacions	1.724	*https://petscan.wmflabs.org/?language=ca&project=wikipedia&depth=2&categories=Qu%C3%ADmica&ns%5B0%5D=1&templates_no=Infotaula%20persona%0D%0AIPP%0D%0AInfotaula%20d%27organitzaci%C3%B3%0D%0AInfotaula%20esdeveniment&search_max_results=500*
Articles dins la categoria «química» (profunditat: 3)	3.972	*https://petscan.wmflabs.org/?language=ca&project=wikipedia&depth=3&categories=Qu%C3%ADmica&ns%5B0%5D=1&search_max_results=500*
i excloent-hi persones, institucions, esdeveniments i publicacions	3.371	*https://petscan.wmflabs.org/?language=ca&project=wikipedia&depth=3&categories=Qu%C3%ADmica&ns%5B0%5D=1&templates_no=Infotaula%20persona%0D%0AIPP%0D%0AInfotaula%20d%27organitzaci%C3%B3%0D%0AInfotaula%20esdeveniment&search_max_results=500*
Articles sobre productes químics	1.166	*https://petscan.wmflabs.org/?language=ca&project=wikipedia&ns%5B0%5D=1&templates_any=Infotaula%20d%27element%20qu%C3%ADmic%0D%0Ainfotaula%20de%20compost%20qu%C3%ADmic%0D%0AICQ%0D%0Ainfotaula%20de%20f%C3%A0rmac&search_max_results=500*

Font: Elaboració pròpia.

Si fem una mirada a l'horitzó i analitzem el nombre d'articles de química en la Wikipedia anglesa (taula 5), aquest se situa ja en les desenes de milers d'articles i, per tant, en magnituds properes al que són els llibres de colors de la IUPAC.

Crec que és important en aquest punt fer menció als articles sobre productes químics, més d'un miler en català i vora els divuit mil en anglès. Difícilment trobarem, en els documents terminològics de referència, reculls de productes químics (noms, explicacions i propietats) d'una extensió similar. És més, és especialment rellevant, en aquest cas, la possibilitat que ofereix la Wikipedia de traduir els noms dels productes químics a través dels enllaços entre les diferents versions de la Wikipedia. Usant aquest recurs no és difícil aconseguir la traducció d'un producte químic a l'àrab, al rus o al xinès, per posar-ne tres exemples.

TAULA 5. *Resultats de l'anàlisi per a la Wikipedia anglesa*

Cerca	*Nombre de pàgines*	*URL*
Articles dins la categoria «chemistry» (profunditat: 2)	26.813	*https://petscan.wmflabs.org/?language=en&project=wikipedia&depth=2&categories=Chemistry&ns%5B0%5D=1&search_max_results=500*
i excloent-hi persones, institucions, esdeveniments i publicacions	25.448	*https://petscan.wmflabs.org/?language=en&project=wikipedia&depth=2&categories=Chemistry&ns%5B0%5D=1&templates_no=Infobox%20scientist%0D%0AInfobox%20organization%0D%0AInfobox%20journal%0D%0AInfobox%20news%20event&search_max_results=500*
Articles dins la categoria «chemistry» (profunditat: 3)	64.371	*https://petscan.wmflabs.org/?language=en&project=wikipedia&depth=3&categories=Chemistry&ns%5B0%5D=1&search_max_results=500*
i excloent-hi persones, institucions, esdeveniments i publicacions	60.411	*https://petscan.wmflabs.org/?language=en&project=wikipedia&depth=3&categories=Chemistry&ns%5B0%5D=1&templates_no=Infobox%20scientist%0D%0AInfobox%20organization%0D%0AInfobox%20journal%0D%0AInfobox%20news%20event&search_max_results=500*
Articles sobre productes químics	18.013	*https://petscan.wmflabs.org/?language=en&project=wikipedia&depth=2&ns%5B0%5D=1&templates_any=chembox%0D%0Adrugbox&search_max_results=500*

FONT: Elaboració pròpia.

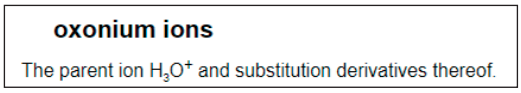
2.2.4. *Actualització: de l'hidroni a l'oxolà*

La darrera característica que proposo avaluar en aquesta comunicació és l'actualització del contingut terminològic de la Wikipedia, posant-ho de nou en el context dels referents terminològics que constitueixen els llibres de colors de la IUPAC, a més d'algunes de les seves darreres publicacions. Novament, em centraré en un grapat de casos que ens permetin traure alguna conclusió; fixem-nos en l'oxoni, l'arsà, el tennes i l'oxolà.

a) De l'hidroni a l'oxoni i l'oxidani

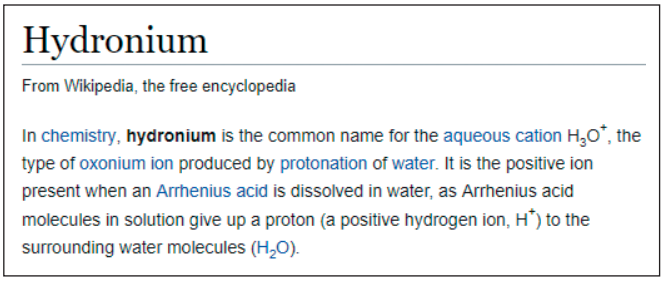
Tot i que el nom *hidroni* segueix sent el terme més usat per a referir-se al catió H_3O^+, és considerat com a incorrecte des del 2005 (darrera edició del *Red Book*). L'hauríem d'anomenar *oxoni* o *oxidani*. Vol dir això que hauríem de fer desaparèixer el terme *hidroni* dels nostres vocabularis?

Les figures 13, 14, 15 i 16 mostren els resultats que s'obtenen en cercar *hidroni*, *oxidani* i *oxoni* al *Gold Book*, la Wikipedia i l'Optimot. Ni el terme *oxidani* ni el terme *hidroni* figuren a l'Optimot ni al *Gold Book*. Només la Viquipèdia catalana fa referència als tres termes i la Wikipedia anglesa tracta el terme *oxoni* com a terme col·lectiu diferenciant-lo del terme *hidroni*.


oxonium ions

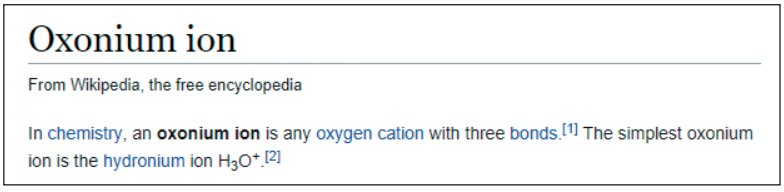
The parent ion H_3O^+ and substitution derivatives thereof.

FIGURA 13. Consulta del terme *oxonium* al *Gold Book*.


Hydronium

From Wikipedia, the free encyclopedia

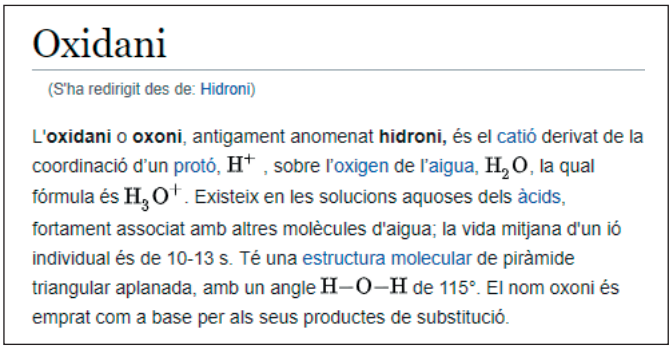
In chemistry, **hydronium** is the common name for the aqueous cation H_3O^+, the type of oxonium ion produced by protonation of water. It is the positive ion present when an Arrhenius acid is dissolved in water, as Arrhenius acid molecules in solution give up a proton (a positive hydrogen ion, H^+) to the surrounding water molecules (H_2O).


Oxonium ion

From Wikipedia, the free encyclopedia

In chemistry, an **oxonium ion** is any oxygen cation with three bonds.[1] The simplest oxonium ion is the hydronium ion H_3O^+.[2]

FIGURA 14. Consulta dels termes *hydronium* i *oxonium* a la Wikipedia anglesa.

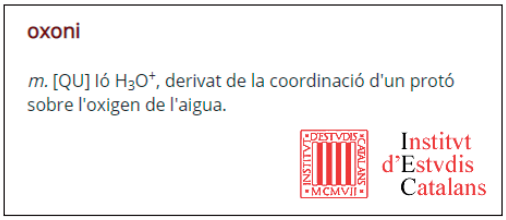


Oxidani

(S'ha redirigit des de: Hidroni)

L'**oxidani** o **oxoni**, antigament anomenat **hidroni,** és el catió derivat de la coordinació d'un protó, H^+ , sobre l'oxigen de l'aigua, H_2O, la qual fórmula és H_3O^+. Existeix en les solucions aquoses dels àcids, fortament associat amb altres molècules d'aigua; la vida mitjana d'un ió individual és de 10-13 s. Té una estructura molecular de piràmide triangular aplanada, amb un angle $H-O-H$ de 115°. El nom oxoni és emprat com a base per als seus productes de substitució.

FIGURA 15. Consulta del terme *hidroni*, que redirigeix al terme *oxidani*, a la Viquipèdia catalana.



oxoni

m. [QU] ió H_3O^+, derivat de la coordinació d'un protó sobre l'oxigen de l'aigua.

Institvt d'Estvdis Catalans

FIGURA 16. Consulta del terme *oxoni* a l'Optimot.

Val a dir que, de les solucions actuals, la més completa i més clara sembla la de la Viquipèdia catalana, en la qual no només es recullen els tres termes, sinó que s'indica amb claredat l'estat desactualitzat del terme *hidroni*.

b) De l'arsina a l'arsà

Les recomanacions del *Blue Book* de 1993 (i la corresponent traducció catalana de 2017) indiquen que el compost AsH_3, anomenat tradicionalment *arsina*, s'ha de denominar *arsà*.

Actualment, el *Gold Book* mostra els termes *arsines* i *arsanes* sense establir cap preferència terminològica, i l'Optimot només recull *arsina*. La Wikipedia recull ambdós termes, que redirigeixen a la mateixa pàgina. Només la Viquipèdia catalana prefereix actualment el terme *arsà*.

c) De l'ununsepti al tennes

Anem per un tercer cas, fixant-nos ara, de forma més concreta, en l'actualització d'un article concret, i especialment en el seu títol. Ens centrarem en el terme *tennes*, el nom de l'element químic de nombre atòmic 117.

Fem primer una mica d'història.

Els experiments que van portar al reconeixement dels elements 113, 115, 117 i 118 es van dur a terme entre 2004 i 2013. Aquests experiments van provocar que, a finals de desembre de 2015, la quarta IUPAC/IUPAP JWP (International Union of Pure and Applied Chemistry / International Union of Pure and Applied Physics Joint Working Party) reconegués el descobriment de núclids dels elements 113, 115, 117 i 118 i assignés prioritats als diferents equips de recerca per a la proposta de noms. Fins aquell moment, l'element 117 s'havia anomenat *ununsepti*, d'acord amb els noms sistemàtics provisionals de la IUPAC.

El 8 de juny de 2016, la IUPAC anuncià els noms proposats i l'inici d'un període de cinc mesos d'exposició pública. Aquest període acabà amb la proclamació dels noms següents, el mes de novembre de 2016: *nihonium* (Nh), *moscovium* (Mc), *tennessine* (Ts) i *oganesson* (Og). Dels quatre, el que ha presentat més dificultats en l'adaptació a les diferents llengües és l'element 117, el *tennessine*, en anglès.

En llengua catalana, i després de propostes diverses generades i/o difoses a través dels mitjans de comunicació, el mes de febrer de 2017, el TERMCAT publicà *tennessi* com a forma normalitzada del nom de l'element 117. El juny de 2018, la Comissió Terminològica de l'IEC revisà la decisió i establí la forma actual, *tennes*.

FIGURA 17. Historial de modificacions de l'article *tennes*.
FONT: *https://ca.wikipedia.org/w/index.php?title=Tennes&action=history*.

Com es mostra en l'historial de l'article *tennes* (figura 17), la comunitat d'editors anà reaccionant de forma ràpida als canvis de nom de l'element 117. El desembre de 2016, l'article canvià el seu nom a *tennès,* el maig de 2017 es canvià a *tennessi* i el setembre de 2018 a *tennes*. Tots tres canvis passaren en un màxim d'uns tres mesos.

d) Del tetrahidrofurà a l'oxolà

L'últim exemple en què em vull fixar és una mica més difícil. La darrera edició del *Blue Book* (2014) indica que el compost denominat fins ara *tetrahidrofurà* s'ha d'anomenar *oxolà* (*oxolane*, en anglès).

A hores d'ara, el terme només apareix en la Wikipedia anglesa, en la infotaula. Cal tenir en compte que la novetat del terme, i el seu ús limitat en la mateixa comunitat científica, farà que probablement calgui esperar que la recomanació prosperi.

Resumint, respecte a l'actualització dels continguts de química en la Viquipèdia catalana, dels casos analitzats se'n desprèn que la Wikipedia s'actualitza adequadament amb els canvis que es van produint en els referents terminològics. En alguns casos, aquesta actualització és més ràpida, o més adequada, que la que es produeix a través d'altres vies de difusió terminològica.

2.3. *Viquiprojectes amb química*

Per acabar aquesta revisió de la presència de la química a la Viquipèdia catalana, farem menció de dos viquiprojectes dels diversos que hi ha actius per a promoure la coordinació de les edicions de la Viquipèdia catalana. Aquests projectes són el Viquiprojecte de Química i el Viquiprojecte IEC, en els quals la presència de la química és significativa.

El Viquiprojecte de Química, que es troba a *https://ca.wikipedia.org/wiki/Viquiprojecte:Qu%C3%ADmica*, té vuit membres, i està essencialment inactiu. L'historial de la pàgina no mostra cap edició des de 2016.

Una situació semblant és la que s'observa amb el Viquiprojecte IEC, *https://ca.wikipedia.org/wiki/Viquiprojecte:IEC*. Aquest es va posar en marxa l'any 2016 amb l'objectiu, entre d'altres, de millorar els continguts de ciència i tecnologia a la Viquipèdia catalana (Cuadros, Dengra i Marginet, 2017). En l'àmbit de la química, es van seleccionar vint-i-set temes. A hores d'ara només consta la participació de dues persones en l'actualització i millora dels continguts de química, que han treballat en vuit dels temes proposats.

Com es pot observar, ambdós projectes han aconseguit un desenvolupament bastant limitat i sembla que actualment estan força aturats.

2.4. *Algunes conclusions*

Com a conclusió del treball presentat aquí i en resposta a la pregunta que hi donava peu, quin és el paper que té o pot tenir la Viquipèdia respecte a la terminologia química en català, es presenten les conclusions següents, que, per la metodologia seguida, són necessàriament provisionals:

La terminologia química és extensa i està constantment en revisió.

La seva adaptació al català, tot i significativa, és limitada i correspon sovint a edicions desactualitzades de les referències de l'àmbit.

El contingut de la Viquipèdia, pel que fa a terminologia química en català, té una extensió significativa, tot i que insuficient; presenta definicions similars a les d'altres obres de referència; té una actualització semblant o millor que aquestes, i és clarament un punt d'accés molt rellevant per a la terminologia química en català. Per aquests motius, és important tenir-ne cura i mantenir els esforços per a ampliar-la i millorar-ne la qualitat pel que fa a la terminologia.

Els projectes en marxa per a dotar la Viquipèdia catalana de més continguts en l'àmbit de la química són limitats i estan poc desenvolupats. Qualsevol esforç que pugui fer-se per donar-los més volada serà més que benvingut.

Bibliografia

Alegret i Sanromà, S. (2013). «L'adaptació dels llibres de la IUPAC al català». A: Sànchez Fèrriz, Miquel-Àngel (cur.). *La terminologia en les ciències de la vida, en la química i en el món educatiu*. Barcelona: Institut d'Estudis Catalans, p. 39-49. (Memòries de la Societat Catalana de Terminologia; 4)

Colomer i Artigas, R. (2005). «La terminologia química en català». *Panace@*: *Revista de Medicina, Lenguaje y Traducción*, 6 (20), p. 124-131.

Cuadros, J.; Dengra, X.; Marginet, R. (2016). «Usem la Viquipèdia per ensenyar química?». *Educació Química*, 22, p. 38-47.

International Union of Pure and Applied Chemistry (2017). *IUPAC Compendium of Chemical Terminology* [en línia]. <https://goldbook.iupac.org/> [Consulta: 30 maig 2019]. [Conegut com a *Gold Book*]

Manske, M. (2019). *Petscan* [en línia]. <https://petscan.wmflabs.org/> [Consulta: 30 maig 2019].

Unió Internacional de Química Pura i Aplicada (IUPAC). Divisió de Química Orgànica. Comissió de Nomenclatura de Química Orgànica (2017). *Guia de la IUPAC per a la nomenclatura de compostos orgànics*: *Recomanacions de 1993 (incloent-hi les revisions, tant publicades com no publicades fins ara, de l'edició del 1979 de la 'Nomenclature of organic chemistry')*. Versió catalana a cura d'Àngel Messeguer. Barcelona: Institut d'Estudis Catalans.

Universitat Politècnica de Catalunya; TERMCAT, Centre de Terminologia; Enciclopèdia Catalana (2019). *Diccionari de química* [en línia]. 2a ed. Barcelona: TERMCAT. <http://www.termcat.cat/ca/diccionaris-en-linia/212>. [Consulta: 30 maig 2019].

Wang, L. D. (2019). «CAS reaches 150 millionth substance». *Chemical & Engineering News*, 97 (22), p. 43.

Jordi Cuadros
IQS (Universitat Ramon Llull)
Societat Catalana de Química

3. Experiències viquipedistes a la Societat Catalana de Biologia[5]

Quelcom potser no prou conegut és que la Societat Catalana de Biologia (SCB) va ser la primera filial de l'Institut d'Estudis Catalans. El seu naixement es remunta al 1912, poc més de cinc anys després que el de la mateixa institució mare. En la seva llarga trajectòria n'han format part un gran nombre de personalitats i la seva tasca ha fet possible la publicació de moltes obres rellevants i un gran nombre d'activitats, evidenciant així el seu suport a la recerca feta al país. Però, no només això, cal destacar-ne també, sobretot, el seu compromís amb la llengua catalana.

Potser l'exponent més icònic de la seva implicació a favor de la normalització del català entre la comunitat de biòlegs va ser la publicació del *Què Cal Saber?* Iniciat el 1984, amb l'assessorament del Servei de Correcció Lingüística de l'IEC i posteriorment també del TERMCAT, consistia en unes fitxes, que es distribuïen entre els socis, sobre terminologia o altres aspectes clau de la llengua per a les ciències biològiques o, simplement, per a un ús professional (extensible a altres disciplines).

Anys després, tot i que ja feia temps que s'havien deixat de publicar més números d'aquesta iniciativa, continuava vigent l'interès perquè els avenços de la biologia poguessin seguir anomenant-se i descrivint-se en català. Això va portar la SCB a explorar com una plataforma llavors emergent centrada en el coneixement lliure i en el treball compartit hi podria ajudar: ens referim a la Viquipèdia.

Arran de l'experiència prèvia d'alguns professors que acollien pràctiques d'universitat en què s'animava els estudiants a crear-hi o millorar-hi articles, es va decidir fer un pas més enllà i fer una trobada oberta a tot el públic interessat a fer precisament això mateix. Aquest tipus d'esdeveniments s'anomenen *viquimaratons* i s'acostumen a aprofitar per a apropar el públic més neòfit a l'edició en la Viquipèdia, a les seves regles i a la seva governança.

Amb la col·laboració d'Amical Wikimedia, organització d'àmbit catalanoparlant que promou la Viquipèdia i els seus principis, es van organitzar fins a tres d'aquests esdeveniments.

La primera d'aquestes viquimaratons, impulsada per la Secció de Biologia Computacional i Bioinformàtica, es va fer el 2015 i va ser sobre bioinformàtica, precisament. Va tenir lloc a la seu de l'Institut d'Estudis Catalans, a la Casa de la Convalescència a Barcelona, i s'hi van aplegar investigadors, estudiants i també editors habituals de l'enciclopèdia lliure.

Un any després, el 2016, en col·laboració amb la secció regional valenciana i la Càtedra de Divulgació de la Ciència de la Universitat de València (UV), es va

5. Aquest article correspon a la participació de Toni Hermoso Pulido en la XVI Jornada SCATERM: «La Viquipèdia i la terminologia» (30 de maig de 2019). Podeu trobar les diapositives de la presentació en l'enllaç següent: *https://slides.com/similis/experiencies-viquipedia-scb-2019#/*.

organitzar un esdeveniment de matí i tarda a l'edifici Octubre, de caire més generalista, al voltant de la biologia. El públic va aprendre sobre els principis que regeixen la Viquipèdia i unes nocions pel que fa al seu ús. Tot seguit, van animar-se a editar-hi entrades. Cal destacar que entre els presents hi havia un professor que havia mantingut glossaris de termes biològics en català.

L'experiència, aquest cop de mig dia, es va repetir el 2017 de nou a Barcelona, aquesta vegada per a guiar membres locals de la SCB interessats a endinsar-se per primera vegada a la Viquipèdia.

Per tal de preparar aquestes viquimaratons s'han generat sovint glossaris de termes d'una temàtica (p. ex., la bioinformàtica). Aquests reculls partien de termes en la Wikipedia en anglès (la més poblada d'articles) i les versions corresponents en altres llengües (sempre que existissin), juntament amb el nombre de visites que rebien aquestes pàgines. Això permetia tenir una visió general de les solucions terminològiques en diferents llengües, com, també, a la comunitat viquipedista, veure quines pàgines calia crear o millorar en català d'acord amb la demanda que podien tenir en una altra llengua.

Finalment, cal recordar que la Viquipèdia, tot i que és l'exponent més conegut, no està sola dins del que és el món Wikimedia, una iniciativa més àmplia on conviuen altres projectes que treballen també pel coneixement lliure per a tothom des de diferents angles, no només en el marc d'una enciclopèdia. És el cas, per exemple, de Wikidata, un repositori de dades del qual avui en dia s'alimenta la mateixa Viquipèdia, omplint les taules laterals (o infotaules) a la part superior de les pàgines o generant llistes diverses en els cossos dels articles. En aquest sentit, s'ha fet un gran esforç perquè moltes dades considerades rellevants sobre la SCB, des de quins han estat els seus presidents fins a quins han estat els articles científics meritoris dels premis que s'atorguen cada any, puguin ser recollides en aquesta plataforma.

Tot això és el que s'ha pogut fer fins ara des de la SCB aprofitant les oportunitats que ofereix la Viquipèdia i els seus projectes germans a favor del coneixement lliure en català. Però, sens dubte, som conscients que encara hi ha molt per fer i esperem poder relatar més experiències en un futur.

Toni Hermoso Pulido
Serveis Cientificotècnics, Centre for Genomic Regulation (CRG)
Amical Wikimedia[6]
Societat Catalana de Biologia

6. En el moment de la conferència era membre d'Amical Wikimedia, però ja no ho és actualment.

Crònica del curs 2018-2019

XVI Jornada de la SCATERM: «La Viquipèdia i la terminologia»

Les intervencions de la XVI Jornada, enregistrades en vídeo, són accessibles a la videoteca de l'Institut, a l'enllaç *https://www.youtube.com/watch?v=do4KWYZc8Xo*

Vista de la mesa inaugural de la XVI Jornada de la SCATERM, celebrada el 30 de maig de 2019 a la Sala Prat de la Riba de l'Institut d'Estudis Catalans. D'esquerra a dreta, Miquel-Àngel Sànchez Fèrriz, president de la SCATERM; M. Teresa Cabré, presidenta de la Secció Filològica, i Ester Bonet (moderadora), vocal de la Junta Directiva de la SCATERM i d'Amical Wikimedia. FONT: SCATERM.

Vista dels ponents de la XVI Jornada. D'esquerra a dreta: Ester Bonet (moderadora), vocal de la Junta Directiva de la SCATERM; Ramon Garriga, de la Fundació Torrens-Ibern; Toni Hermoso, de la Societat Catalana de Biologia; Jorge Vivaldi, de l'Institut de Lingüística Aplicada; Pau Cabot, d'Amical Wikimedia; Joan de Solà-Morales, de la Societat Catalana de Matemàtiques, i Jordi Cuadros, de la Societat Catalana de Química. FONT: SCATERM.

Programa de la XVI Jornada

XVI Jornada de la SCATERM
«La Viquipèdia i la terminologia»

Sala Prat de la Riba de l'Institut d'Estudis Catalans, 30 de maig de 2019

9.00 h *Paraules de benvinguda*
Miquel-Àngel Sànchez Fèrriz
President de la Societat Catalana de Terminologia

Inauguració
M. Teresa Cabré
Presidenta de la Secció Filològica de l'Institut d'Estudis Catalans

9.30 h *Conferència*
Viquipèdia: un recurs útil per a la terminologia?
Jorge Vivaldi
Universitat Pompeu Fabra

10.30 h *Pausa cafè*

11.00 h *Ponències*
És fiable la Viquipèdia en català? Manteniment, estandardització i control a la Viquipèdia en català
Pau Cabot
Amical Wikimedia

Projecte Viquiterm
Ramon Garriga
Fundació Torrens-Ibern
Toni Hermoso
Amical Wikimedia

12.30 h *Taula rodona*
Experiències de les societats filials de l'IEC en la Viquipèdia
Joan de Solà-Morales (Societat Catalana de Matemàtiques)
Jordi Cuadros (Societat Catalana de Química)
Toni Hermoso (Societat Catalana de Biologia)

13.00 h *Debat i cloenda de la jornada*
Moderadora: Ester Bonet
Societat Catalana de Terminologia

Presentació de la XVI Jornada

Junta Directiva de la SCATERM

Quan ens assabentem que la Viquipèdia en català rep una mitjana de divuit milions de visites al mes, ens adonem que potser no aprofitem prou aquesta enciclopèdia. Tanmateix, les consideracions que tots ens fem sobre la seva validesa i la seva verificabilitat ens frenen a l'hora d'imaginar com la podríem utilitzar millor i, si ens centrem en l'àmbit de la terminologia, com la podríem convertir en una eina de divulgació terminològica i en un referent de qualitat divulgativa.

En aquesta jornada us convidem a escoltar diferents experiències (per exemple, sobre els mecanismes de control de la fiabilitat, sobre la capacitat d'autogestió de la verificabilitat del contingut i sobre els usos que se li poden donar en el context acadèmic) per poder entendre si la Viquipèdia és un camí que ens porta a la confusió i al desordre, i per tant cal abandonar-lo com una via en desús, o si, en canvi, pot ser una via òptima per a l'ús de la llengua catalana i una font d'informació multilingüe, sobretot en relació amb la terminologia.

Balanç i conclusions de la XVI Jornada

Ester Bonet
Societat Catalana de Terminologia
Amical Wikimedia

Quan la SCATERM us ha convidat a aquesta jornada, era perquè escoltéssiu diferents experiències, per poder entendre si la Viquipèdia és un camí que ens porta a la confusió i al desordre —i que, per tant, cal abandonar-lo com una via en desús— o si és una via òptima per a la normalització de la llengua catalana i una font d'informació multilingüe apta per a la recerca.

D'entrada, la validesa i la verificabilitat d'aquesta enciclopèdia ens frenen a l'hora d'imaginar com la podríem utilitzar millor, però quan ens assabentem que la Viquipèdia en català rep divuit milions de visites al mes, se'ns disparen les alarmes i ens preguntem si no ens estem perdent alguna oportunitat. Si ens centrem en el camp de la terminologia, que és el que ens ha reunit aquí, com la podríem convertir en una eina de divulgació terminològica?

Ens ha agradat, sobretot, portar aquesta reflexió aquí, a l'Institut d'Estudis Catalans, perquè tots els que estem darrere de la Viquipèdia catalana ens estimem la llengua catalana —això explica que tingui aquest volum d'entrades amb tan pocs parlants com tenim, comparat amb la Wikipedia anglesa o amb la castellana. És una militància voluntària que fem dia a dia. Hem tingut la sort de poder reunir aquí la gent que, dins de la Viquipèdia, representa pesos molt específics, que fa molts anys que hi treballen i que volen que això continuï endavant. Volem que aquest acte sigui una encaixada de mans per dir-vos que Amical Wikimedia és aquí i que Amical vol treballar amb vosaltres. No podem obviar aquests divuit milions de visites que té la Viquipèdia catalana. És impossible.

Cal tenir present també que hi ha alumnes de batxillerat i de primer de carrera que la consulten molt, i els hem de deixar una Viquipèdia que els sigui útil i que els permeti aprendre. Això és un treball que com a Viquipèdia catalana podem fer. No ens cal aspirar a tenir els sis milions d'entrades que té la Wikipedia anglesa, però sí que podem aspirar a proporcionar una eina de consulta a aquests estu-

diants, una eina en la qual basar-se i, sobretot, una eina en la qual més endavant podran col·laborar, perquè a Amical Wikimedia, a més de ser militants de la llengua, també som militants del coneixement obert. És cap aquí cap on hem d'anar: no és voler les coses gratuïtes, sinó voler eines compartides.

Per tant, crec que aquesta jornada ha estat un primer pas —que n'ha de tenir un de segon— i que ha aconseguit presentar-vos la Viquipèdia no només com a recurs de coneixement obert, sinó també com a mitjà de difusió terminològica.

Crònica de la XVI Jornada

Junta Directiva de la SCATERM

La XVI Jornada de la SCATERM, titulada «La Viquipèdia i la terminologia», va tenir lloc el dia 30 de maig de 2019 a l'Institut d'Estudis Catalans (IEC). Tenia com a objectiu analitzar l'enciclopèdia en línia Wikipedia, la versió catalana de la qual rep una mitjana de divuit milions de visites al mes, per poder entendre si és un camí que ens porta a la confusió i al desordre o si, en canvi, pot ser una via òptima per a l'ús de la llengua catalana i una font d'informació multilingüe, sobretot en relació amb la terminologia.

Miquel-Àngel Sànchez Fèrriz, president de la Societat Catalana de Terminologia (SCATERM), i M. Teresa Cabré, presidenta de la Secció Filològica de l'IEC, van inaugurar l'acte. Cabré va fer èmfasi en el fet que la jornada podria respondre a algunes preguntes que se solen fer sobre la Viquipèdia, per exemple en relació amb la qualitat lingüística o el control de la variació. Si cada cop més estudis científics —sobretot sobre terminologia— es basen en aquest recurs enciclopèdic, és important conèixer si existeix un biaix informatiu.

La conferència inaugural, a càrrec de Jorge Vivaldi, de la Universitat Pompeu Fabra, va presentar les característiques principals de la Viquipèdia i va mostrar com es pot aplicar al processament del llenguatge natural (PLN): des de la creació de corpus textuals fins a la traducció automàtica i la creació de bases de coneixement, entre altres aplicacions. A més, també va demostrar que, a partir de la informació lingüística que conté aquesta enciclopèdia, és possible l'extracció de terminologia bilingüe i l'extracció dels termes d'un text especialitzat. Tanmateix, Vivaldi també va destacar que ara per ara l'ús de la Viquipèdia en català per a aquest tipus d'aplicacions de PLN és escàs i, per tant, cal estar pendents de com evoluciona.

A continuació, van tenir lloc dues ponències. Pau Cabot, d'Amical Wikimedia, va sorprendre l'audiència afirmant que la Viquipèdia no és fiable. Va afe-

gir-hi, però, alguns matisos esperançadors. Malgrat que representa un entorn hostil per mantenir-ne la qualitat precisament a causa de la seva essència d'estar oberta a tothom, darrerament s'ha aconseguit millorar la qualitat i la quantitat d'articles en català, gràcies a un esforç pel que fa a les tasques de manteniment que duen a terme tant els usuaris com la Fundació Wikimedia. Ramon Garriga, de la Fundació Torrens-Ibern, acompanyat de Toni Hermoso, va presentar el Viquiterm, un fòrum basat en el funcionament essencial de la Viquipèdia. Pretén ser una plataforma oberta perquè tant científics i tècnics com professionals de la llengua i la terminologia puguin plantejar-hi dubtes terminològics, discutir-los i arribar a un consens. L'objectiu del Viquiterm és completar les mancances terminològiques dels recursos existents en llengua catalana.

L'últim acte de la jornada va consistir en una taula rodona en la qual diverses societats filials de l'IEC van presentar llurs experiències en relació amb la Viquipèdia. Joan de Solà-Morales, de la Societat Catalana de Matemàtiques, va explicar com es van seleccionar els termes que es van treballar en una viquimarató destinada especialment a cobrir la terminologia bàsica per a estudiants de batxillerat. També va fer èmfasi en els problemes que sorgeixen en les entrades de la Viquipèdia de l'àmbit de les matemàtiques, en referència a l'adequació dels gràfics i a la tipografia dels símbols i de les fórmules, ja que considera que, a causa de l'abast d'aquesta enciclopèdia en línia, podria ser un referent per a la fixació de la tipografia. Jordi Cuadros, de la Societat Catalana de Química, va destacar la utilitat de la Viquipèdia per a la difusió dels conceptes fonamentals de la química que es recullen en els llibres de colors de la IUPAC (Unió Internacional de Química Pura i Aplicada), ja que comparada amb la traducció tradicional comporta un gran estalvi de temps i contribueix a la immediatesa en la difusió del coneixement. Ara bé, perquè això sigui possible, la Viquipèdia ha de complir uns requisits mínims de correcció, accessibilitat, extensió i actualització. En aquest sentit, segons Cuadros, pel que fa a les entrades de termes químics en llengua catalana, caldria millorar-ne la qualitat lingüística, la precisió i l'abast. Finalment, Toni Hermoso, de la Societat Catalana de Biologia, va detallar les accions que duu a terme la societat filial en relació amb la Viquipèdia i es va centrar a explicar com s'hi detecten les mancances terminològiques de l'àmbit de la biologia i què es fa per poder cobrir-les. S'utilitzen estratègies diverses basades en els recursos que ofereix la mateixa eina en línia.

La sessió es va cloure amb un prolífic debat, moderat per Ester Bonet, membre de la Junta Directiva de la SCATERM i d'Amical Wikimedia, en el qual es va fer evident l'interès que suscita la Viquipèdia en relació amb la llengua i la terminologia. En definitiva, al llarg de tota la jornada es van respondre algunes de les preguntes que augurava M. Teresa Cabré en la inauguració i estem convençuts que se'n van generar d'altres. Igualment, va resultar útil com a font d'idees per a noves investigacions, sobretot des del punt de vista de la terminologia.

Memòries de la Societat Catalana de Terminologia

Títols publicats

1 Jaume Martí i Marina Salse (coord.), *La terminologia i la documentació: relacions i sinergies* (2010)
2 Eusebi Coromina i Josep M. Mestres (cur.), *Aspectes de terminologia, neologia i traducció* (2010)
3 Lluc Potrony i Joan Maria Romaní (cur.), *Indexació, terminologia i llenguatge jurídic* (2011)
4 Miquel-Àngel Sànchez Fèrriz (cur.), *La terminologia en les ciències de la vida, en la química i en el món educatiu* (2013)
5 Miquel Strubell i Trueta (cur.), *La terminologia instrumentalitzada* (2015)
6 Miquel-Àngel Sànchez Fèrriz i Rosa Mateu (cur.), *La ciència en català: des del segle xiii fins avui* (2018)
7 Judit Feliu i Mireia Trias (cur.), *Gramàtica, esport i terminologia* (2019)
8 Judit Feliu i Mireia Trias (cur.), *Viquipèdia i terminologia* (2021)

scatterm

SOCIETAT CATALANA DE TERMINOLOGIA
Filial de l'Institut d'Estudis Catalans

9 788499 656052