bid

textos universitaris de biblioteconomia i documentació

ISSN 1575-5886
DL B-19.675-1998
DOI 10.1344/BiD2012.29.16

número 29
desembre de 2012

Facultat de Biblioteconomia i Documentació
Universitat de Barcelona

Seguiu-nos: estratègies de captura de tuits en català

Anita E. Locher
Facultat de Biblioteconomia i Documentació
Universitat de Barcelona
alocher@ub.edu

Aina Giones-Valls
VTLS Europe
gionesa@vtlseurope.com

Eli Ramírez
Bibliotecària *freelance*
eliramirez82@gmail.com

Grelda Ortiz
Real Academia de Ciencias Económicas y Financieras
biblioteca@racef.es

Resum [Abstract] [Resumen]

Twitter és una eina de *microblogging* utilitzada arreu del món per generar i compartir informació. La comunitat catalana la utilitza des del Govern fins als mitjans de comunicació, passant per persones amb diferent formació i estils de vida. No obstant això, a dia d'avui, a Catalunya, hi ha poques actuacions clares per preservar-la des d'una perspectiva de gestió bibliotecodocumental de Catalunya. Aquest article proposa una estratègia de captura de tuits per a la Biblioteca de Catalunya, que inclou criteris de selecció, metadades de preservació i previsions de creixement.

Metodologia: S'analitzen els aspectes legals i tecnològics del Twitter, de la informació de l'arxiu Twitter de la Library of Congress, del portal amb estadístiques Twiter'n'Català i del PADICAT (Patrimoni Digital de Catalunya), per descriure l'estat de la qüestió de les possibilitats de preservació dels tuits que té la Biblioteca de Catalunya.

Resultats: Els resultats recollits mostren que actualment fer servir una única eina no és suficient per filtrar, capturar i preservar els tuits. Els requeriments d'una institució que té com a objectiu la preservació a llarg termini fan semblar insuficients les eines existents a Internet. No obstant això, la necessitat d'actuar immediatament per no perdre el patrimoni digital fa que s'hagi d'actuar amb les eines i els recursos disponibles. Amb una combinació d'eines existents es poden millorar les estratègies de captura de tuits actuals.

1 Introducció

Twitter és una eina de *microblogging* que serveix per publicar breument, amb cent quaranta caràcters, un missatge d'estat ?també conegut com a *actualització*? per a persones i entitats. Aquesta eina permet seguir les actualitzacions que interessen els usuaris.

Com a eina social (Boyd; Ellison, 2007) compleix els requisits següents:

- És un servei basat en el web.
- Permet a l'usuari crear un perfil públic o semipúblic.
- Estableix una llista d'usuaris amb els quals es comparteix una relació.
- Visualitza les llistes de les connexions pròpies i de les d'altres usuaris.

Quan algú es vol donar d'alta a Twitter cal que proporcioni al sistema algunes dades (nom, correu electrònic i nom d'usuari). Aquestes dades formen part del perfil, que pot complementar amb una descripció, una imatge i un lloc que són públics, i també els seguidors que té l'usuari i els que segueix.[1]

Cada missatge publicat és un tuit o piulada, que pot contenir (o no) els elements següents:

- @usuari: serveix per mencionar un usuari (respondre, citar, etc.).
- # etiqueta (*hashtag*): identifica una paraula o una expressió significativa per a un tuit i serveix per recuperar-les posteriorment.
- RT: republicar continguts que ha publicat un altre usuari.

- També s'hi permet incloure text, URL, fotografies, vídeos, etc.

Twitter és un recurs nascut digitalment, publicat en obert i en temps real i que s'utilitza àmpliament a Catalunya.[2] És una font d'informació de primera mà sobre esdeveniments i notícies d'actualitat de caràcter no oficial, en què la societat reflecteix fets socials, concentracions, activitats, estats, d'una manera efímera (Bruns; Burgess, 2011). En aquest article volem investigar com i què s'està preservant. En general, les entitats que s'encarreguen de la preservació digital seleccionen, capturen i descriuen els arxius web i hi donen accés depenent de la legislació de cada país.[3] També cal destacar que els problemes trobats en la majoria de centres són comuns a totes les institucions; es tracta de problemes tecnològics i mancança de recursos humans (Castello; Priem, 2008).

Catalunya no n'és una excepció i des de fa temps preserva arxius web digitalment a través d'alguns projectes, com ara la Memòria Digital de Catalunya i ARCA, entre d'altres, desenvolupats per la Biblioteca de Catalunya (BC) (Serra; Pérez; Llueca, 2011; Llueca *et al.*, 2010). Amb la finalitat de garantir la perdurabilitat de les dades de diferents projectes digitals, s'ha creat el sistema COFRE (COnservem per al Futur Recursos Electrònics), un repositori de preservació digital d'alta seguretat (Serra; Pérez; Llueca, 2011). En el cas de les pàgines web catalanes, PADICAT (Llueca *et al.*, 2010) captura el domini .cat semestralment, mentre que altres pàgines web que tracten de temes seleccionats pels professionals de la informació es capturen periòdicament. En aquesta selecció temàtica s'inclouen els comptes públics de Twitter d'alguns polítics i partits (Llueca *et al.*, 2011). És a dir, que la BC forma part de les biblioteques a l'avantguarda perquè ja està capturant contingut de Twitter. No obstant això, en aquest article proposem fer un pas més en aquest procés de captura.

En aquests darrers dos anys ha augmentat notablement el nombre d'usuaris de Twitter. Només en un dia del mes de febrer del 2011 es van crear una mitjana de 460.000 comptes nous que van arribar a generar una mitjana de 140.000 tuits al dia.[4]

Des de juliol del 2012, el català és un idioma oficial a Twitter. Actualment té més de 52.000 usuaris.[5]

1.1 Per què és important preservar tuits catalans?

Algunes institucions públiques, com ara la Generalitat de Catalunya,[6] ajuntaments o entitats culturals, han triat crear un perfil a Twitter com un mitjà de comunicació més per interactuar amb la gran quantitat de catalans que cada vegada més fan servir aquesta eina. La informació que es genera a través d'aquests tuits pot ser interessant per a futurs estudis sociològics, històrics, lingüístics, estadístics o periodístics, entre d'altres. Investigadors com Banks (2009) consideren que els tuits són literatura grisa. Per aquest motiu, la seva preservació forma part ja del patrimoni documental català.

Aquest article vol cridar l'atenció sobre la rapidesa amb la qual desapareixen els tuits, si es té en compte que Twitter només permet recuperar els últims 3.200 tuits d'un usuari i 1.500 tuits per cada cerca, ja sigui etiqueta o text lliure. Per exemple, la Generalitat de Catalunya, en el seu perfil @gencat, va publicar el seu primer tuit el 29 de març del 2009. Ara, que té 5.181 tuits, ja no es poden recuperar els primers. Si la BC no actua, haurà de dependre de la Library of Congress (LC) per accedir-hi perquè és qui els conserva tots per ara.[7]

2 Metodologia

Per conèixer quines altres iniciatives duen a terme centres similars a la BC arreu del món, es va contactar amb les diferents institucions que formen part de l'International Internet Preservation Consortium (d'ara endavant, IIPC). L'IIPC fomenta l'intercanvi d'experiències en preservació web dutes a terme per les diferents entitats: biblioteques, arxius, universitats i entitats privades.

Ens vam posar en contacte amb vint-i-sis membres de l'IIPC,[8] dels quals van contestar catorze, i d'acord amb aquesta informació hem desenvolupat les propostes per a la BC. L'objectiu era esbrinar quines eren les tecnologies en ús i els criteris d'inclusió o exclusió de les xarxes socials i especialment de Twitter. En el primer correu electrònic es van plantejar les preguntes següents:

1. *Which are the exclusion or inclusion criteria or attitudes towards social media or web 2.0 for your archive?*
2. *Are there such documents regarding social media, specially Twitter, and your web archive, that you could share?*
3. *Can or will your web archive capture public tweets?*

Segons les respostes obtingudes es van elaborar preguntes noves sobre freqüència de captura, la tecnologia utilitzada, els problemes derivats, etc., que han servit per arribar als resultats que presentem a continuació.

3 Resultats

Segons les respostes rebudes, s'ha pogut observar que són pocs els centres que capturen tuits amb regularitat. En els centres on s'han fet proves, la captura no ha estat satisfactòria o ha presentat errors.

Els centres que capturen tuits ho fan perquè aquests missatges estan relacionats amb altres pàgines i no per fer una

mostra representativa de la xarxa social i d'allò de què s'hi parla. Les captures, segons aquests criteris, no són representatives per a la xarxa, ja que són fruit d'una captura d'un moment concret, i a Twitter l'actualització és constant i la freqüència de captura no està a l'altura de les circumstàncies.

Aquestes institucions capturen els tuits o bé perquè formen part de pàgines d'un domini seleccionat, o bé perquè són d'un usuari rellevant.

La intenció dels arxius no és, en cap cas, crear un arxiu de Twitter separat, desconnectat de l'arxiu web; més aviat volen capturar totes les formes d'expressió relacionades amb un tema concret i fer-les accessibles en conjunt. En les polítiques de col·lecció no es mencionen explícitament les xarxes socials. Aquest fet es pot explicar de dues maneres:

1. Les polítiques s'han fet en moments en què les xarxes socials no existien o no tenien un paper gaire important.
2. Moltes de les polítiques de col·lecció són prou generals com per incloure xarxes socials.

De fet, si s'apliquen els criteris de "nacionalitat" que moltes biblioteques nacionals apliquen al seu fons imprès, els tuits inclosos serien:

- Produïts al país.
- Creats per un ciutadà del país.
- En l'idioma del país.
- Sobre una temàtica del país.

Per tant, no cal que les xarxes socials estiguin incloses explícitament en les polítiques de col·lecció, encara que alguns centres tenen previst incloure-les dins els seus arxius de preservació web. Només l'arxiu web de la República Txeca exclou intencionadament les xarxes socials per dues raons: en primer lloc, perquè les consideren converses privades i, en segon lloc, pels problemes tècnics. Els arxius que només capturen per domini necessitaran noves estratègies per tenir representades les xarxes socials de les seves comunitats. En el cas de captura selectiva (en tant que no captura tot el domini ja que se segueixen criteris temàtics, per exemple), l'arxiu pot triar la part que cal preservar. El problema rau a saber què és rellevant per a la institució.

Davant d'aquesta problemàtica es pot optar per, o bé seleccionar perfils coneguts, o bé capturar exhaustivament totes les publicacions fetes en l'idioma del país amb mètodes automàtics.

Per exemple, l'arxiu de la República Txeca ha desenvolupat un connector per reconèixer automàticament l'idioma i per filtrar pàgines d'una temàtica concreta. D'altres, com ara les biblioteques nacionals d'Àustria i Dinamarca, capturen les pàgines de perfil de Facebook i Twitter, malgrat que queden fora de context a causa de la ràpida freqüència d'actualització dels usuaris.

La majoria de les captures es fan una o dues vegades l'any, tot i que en períodes electorals la freqüència sol augmentar. Malauradament, aquesta freqüència resulta del tot insuficient per esdevenir un resultat representatiu del moviment social. En cap cas no ens hem trobat amb institucions que detectin i capturin esdeveniments espontanis com ara el moviment dels indignats amb l'etiqueta #15m[9] o la lluita contra els preus de les autopistes catalanes amb l'etiqueta #novullpagar.[10]

A continuació, abans d'exposar els problemes trobats, expliquem algunes eines que han mencionat les biblioteques consultades o que apareixen en la bibliografia; hem descartat les que no són de codi obert o que no permeten l'exportació dels resultats. Entre aquestes eines, podem diferenciar les de selecció i les de captura.

Eines de selecció

En el cas d'arxius selectius es poden utilitzar eines de selecció que permetin detectar perfils, tuits o converses amb un cert grau de repercussió i, d'aquesta manera, poder automatitzar el procés de selecció. Aquestes eines utilitzen diferents criteris per generar la llista dels més populars. Per això, cada centre pot utilitzar l'eina amb els criteris que millor s'adaptin a les seves necessitats, per exemple, nombre de seguidors d'un usuari, popularitat, freqüència de les actualitzacions, data de l'última actualització o nombre de retuits rebuts, entre d'altres. En destaquem dues:

- TwitterGrader i Twitleve: eines que serveixen per crear rànquings d'influència d'un perfil o d'un tuit dins de la xarxa social Twitter.

- Twitter 'n' Català: projecte de Data'n'press[11] que pren el relleu del portal Twit.cat,[12] i que dóna les estadístiques d'ús de Twitter utilitzant l'operador "language" de l'eina i altres factors com els tuitaires que segueixen el seu compte i un estudi de la xarxa en què detecta possibles usuaris catalans, i es crea així una primerenca base de dades.

Eines de captura

Són les que permeten cercar, copiar i exportar els tuits dels comptes dels usuaris. En destaquem les següents:

- Twitter Api (*application programming interface*, interfície de programació d'aplicacions): eines creades per Twitter que permeten als programadors desenvolupar les pròpies aplicacions. La REST API permet publicar

microentrades en les aplicacions, seguir algú o crear llistes. La Search API serveix per cercar tuits en un índex de tuits recents; no es poden recuperar tuits més enllà d'una setmana i només es recuperen els tuits considerats rellevants dins la cerca. Les Streaming API les utilitzen els desenvolupadors que volen tota la seqüència de tuits en temps real en el mateix moment en què es publiquen a Twitter. Trobem l'API per a la seqüència de tuits públics (*public streams*), per a la seqüència de tuits d'un compte d'un usuari (*user streams*) i per a la seqüència de molts comptes de diferents usuaris (*site streams*). Aquesta última és una eina molt recent i, per això, no totes les aplicacions hi tenen accés. Les dues eines següents utilitzen el Twitter Api com a base:

    - TwapperKeeper: programa finançat pel JISC (Joint Information Systems Committee) del Regne Unit. Permet crear un arxiu de tuits en un servidor propi en quatre formats diferents: HTML, RSS, XLS i JSON.[13]

    - Backupify: permet exportar els continguts dels tuits en arxius PDF indexats o fitxers JSON. Arxiva fins a 1 GB de tuits en el núvol de manera gratuïta utilitzant la infraestructura Amazon S3 (*simple storage service*). En cas que se sobrepassi el gigabyte, el servei passa a ser de pagament. Seguint els límits que imposa Twitter Api, només permet capturar els 3.200 tuits més recents.

- Archive-it: servei ofert per Internet Archive, entitat que treballa en la preservació web des de 1996. Archive-it és una aplicació web de pagament que permet crear, descarregar i gestionar col·leccions digitals amb diferents tipus de continguts i accedir-hi, com ara: HTML, vídeos o àudio.[14] Cal destacar que pot capturar xarxes socials, entre les quals hi ha Twitter. Aquesta eina permet exportar la col·lecció i rebre'n una còpia dins un disc dur amb les dades capturades.

- Heritrix: rastrejador web (*crawler*) gratuït i de codi obert que va crear Internet Archive amb la col·laboració del Nordic Web Archive el 2003 (Mohr *et al.*, 2004) implementat en llenguatge JAVA. El format de captura és HTML, i el d'emmagatzemament de la informació és ARC. És l'eina que utilitza PADICAT.

- Web Analyzer: aplicació que es pot integrar en un dels mòduls de Heritrix que permet identificar l'idioma de la pàgina web o filtrar-la per tema. Està desenvolupada per la Biblioteca de la República Txeca (Vlcek, 2008).

3.1 Problemes trobats

S'han identificat problemes comuns a totes les biblioteques o a tots els centres de documentació, però també d'altres que només afectaven un nombre concret de centres. A continuació, es presenten aquests problemes agrupats en: tecnològics, legals i ètics.

3.1.1 Problemes tecnològics

Els problemes tecnològics són múltiples. El principal és que Twitter continua canviant i migrant cap a noves tecnologies. Per exemple, abans s'utilitzava HTTP bàsic per a l'autenticació dins la xarxa, mentre que ara es fa servir OAuth.[15]

Heritrix, usat per PADICAT i la majoria de centres, funciona utilitzant un rastrejador web que recol·lecta la pàgina i els enllaços que conté. Les pàgines capturades són còpies de l'original. PADICAT captura els comptes en HTML ? uns vint tuits per pàgina?, i el pes varia d'un usuari a un altre; així, un compte institucional pesa 153 MB, mentre que un de personal pesa 20 kB.

Ara bé, si s'utilitza la tecnologia AJAX,[16] el robot no pot capturar tota la informació. En aquest cas, el resultat es visualitza d'una manera diferent de com el veu l'usuari d'Internet. Twitter utilitza AJAX quan mostra les respostes a un tuit, per exemple. Com en el cas següent: la figura 1 mostra els tuits fets per un usuari i que es poden capturar amb Heritrix; en canvi, la figura 2 mostra la informació d'aquests tuits que no queda guardada per Heritrix.

Figura 1. Captura de pantalla en què es veuen els tuits recol·lectats per Heritrix

Figura 2. Captura de pantalla de la informació que no s'està recol·lectant. Quan l'usuari clica sobre un tuit per veure'n les respostes, apareix informació que Heritrix no capturarà

La captura de pàgines de Twitter es complica per l'ús del símbol coixinet (#!) que s'introdueix a l'identificador uniforme de recursos (URI) quan el web utilitza AJAX, tal com es pot apreciar en la figura 3. La part de l'URI que segueix després del símbol no s'envia al servidor, sinó que és interpretada pel navegador. Això impedeix a Heritrix capturar les subpàgines d'un domini amb coixinet.

Figura 3. Captura de pantalla en què s'assenyala el símbol #! utilitzat per AJAX

Es presenta un problema addicional quan es vol capturar una etiqueta. Primer s'ha de fer una cerca per després capturar els tuits resultants. No obstant això, les pàgines de resultats estan bloquejades per a màquines com Heritrix amb fitxers *robots.txt*. Tot i així, algunes biblioteques ignoren els *robots.txt* i capturen la pàgina igualment.

Un altre problema consisteix en la identificació dels tuits com a catalans de manera automàtica. No es pot fer servir el domini, com es fa en alguns països per seleccionar pàgines web, perquè tots els perfils de Twitter tenen el domini .com. Es poden utilitzar els operadors "language" i "place" amb l'API de Twitter però, aleshores, queden exclosos els

tuits que no hagin afegit aquesta funcionalitat en els comptes.

També ens trobem que la freqüència d'actualització de Twitter és molt alta i Heritrix, de moment, només fa captures dos cops al dia. Si la freqüència d'actualització dels comptes que cal capturar augmenta, Heritrix hauria de poder augmentar també la freqüència de captura. A part, fa falta un sistema de monitoratge per evitar capturar dues vegades el mateix tuit en cas que la freqüència de captura sigui superior a la d'actualització, o poder, si cal, recuperar un tuit que no s'hagi capturat (Kelly *et al.*, 2010).

L'últim problema al qual farem referència és que l'API de Twitter només dóna com a resultats els tuits més recents i no pas tots els tuits que corresponen als criteris de la cerca. L'API només permet la recuperació dels últims 1.500 tuits publicats els darrers nou dies com a resultat d'una cerca, tal com es pot veure en la figura 4, en què es busca una etiqueta de la qual ja no es recuperen resultats, encara que n'hi hagin.

Figura 4. Captura de pantalla en què no es poden recuperar els tuits existents amb l'etiqueta #osr4 (4a Jornada Open Science Repositories, que va tenir lloc a Barcelona l'any 2010), ja que no són recents

3.1.2 Problemes legals

En el moment d'acceptar les condicions legals del servei, els usuaris de Twitter són els posseïdors dels drets d'autor sobre el contingut dels seus tuits. En cas de litigi, els usuaris accepten la jurisdicció dels Estats Units perquè és on hi ha la seu de l'empresa.[17] Els usuaris cedeixen drets d'explotació a Twitter Inc., de manera que aquesta empresa pot llicenciar i subllicenciar l'ús d'aquests tuits públics per defecte sense cap contraprestació econòmica a l'autor. De fet, Twitter Inc. ja comparteix aquestes dades amb empreses externes com ara Crimson Hexagon[18] o Mediasift.[19]

Segons l'API Terms of Service, només es pot exportar contingut de Twitter en PDF o en full de càlcul. L'exportació a una base de dades no està permesa, per la qual cosa el PADICAT no pot incloure les dades capturades amb l'API de Twitter.

Gomes, Freitas i Silva (2006) expliquen que és difícil per a un país exigir a un altre que li entregui dades que es troben en servidors externs com és el cas de Twitter. De totes les institucions consultades, la Biblioteca Nacional de Dinamarca té una llei de dipòsit legal des del 2004; en canvi, altres països, sense llei de dipòsit legal, creen i mantenen arxius web.

Twitter va fer una donació de tota la seva base de dades a la Library of Congress.[20] Per contracte, no se'n pot fer una distribució abans de sis mesos i s'ha d'avisar els usuaris del Twitter Archive de la LC que no se'n pot fer ús comercial ni redistribuir-los. Tot i que la donació es va fer l'abril del 2010, l'arxiu encara no és consultable. Tot i que els drets d'autor no siguin un problema, recordem que els tuits són públics; ara bé, sí que s'han de tenir en compte qüestions relacionades amb la privacitat. A més, hi ha el problema del fitxer *robots.txt*, el qual s'ha d'ignorar si es volen capturar els resultats d'una cerca mitjançant etiquetes (O'Keeffe, 2011).

Legalment parlant, la BC s'empara sota la Llei de biblioteques de Catalunya i la Llei del sistema bibliotecari de Catalunya,[21] que estableix que la seva missió és recopilar, conservar i difondre la producció bibliogràfica catalana incloent-hi la digital. D'acord amb aquesta normativa, es poden capturar perfils públics de Twitter d'usuaris catalans, que s'inclouen dins del patrimoni digital (Llueca *et al.*, 2010 i Serra; Pérez; Llueca, 2011).

3.1.3 Problemes ètics

La majoria de les persones no llegeixen els termes de servei i molts no saben que, quan accepten utilitzar Twitter, estan signant un contracte. Si algú vol llegir els termes de servei té dues opcions: d'una banda, llegir la versió vinculant en anglès amb possibles dificultats lingüístiques, o de l'altra llegir la traducció en castellà, no vinculant, i amb el risc de llegir una versió no actualitzada i, així, malinterpretar les normes d'ús de Twitter.

Alguns usuaris han mostrat el desacord que sigui el govern, a través de la LC, qui capturi i preservi els seus tuits. Això es manifesta a través de l'etiqueta no vinculant #noLoC, que utilitzen usuaris d'arreu del món.

D'altra banda, una pràctica comuna i poc ètica desenvolupada per algunes empreses consisteix a crear comptes falsos d'usuaris[22] per fer-los seguidors d'un compte de Twitter (marca, persona, etc.). Actualment, encara és freqüent valorar la presència en línia segons la quantitat de contactes o seguidors i no tant segons la qualitat de les relacions que s'hi desenvolupen. Això pot comportar criteris erronis a l'hora de seleccionar comptes de Twitter influents.

4 Estratègies generals

4.1 Estratègia de selecció

Hi ha tres tècniques de selecció per als tuits: la selecció exhaustiva, el mostreig aleatori i la selecció subjectiva.

En primer lloc, una selecció exhaustiva no és possible perquè no es pot saber el conjunt de població de Twitter d'un lloc geogràfic, ja que el fet d'afegir la localització al compte d'un usuari no és una dada obligatòria. Una manera de saber la localització seria mitjançant l'IP de l'ordinador, però hi ha gent que utilitza un servidor intermediari o que desactiva la memòria cau. Si es tenen en compte aquestes variables caldria calcular molt bé el marge d'error.

En segon lloc, es podria utilitzar una eina que filtrés els tuits per idioma, però llavors s'hi haurien de considerar els catalans que tuitegen en altres idiomes. Sense el conjunt de la població no es pot obtenir un mostreig aleatori. Es podria fer, però no seria representatiu.

En darrer lloc, amb la selecció subjectiva, en canvi, seria més fàcil de fer perquè cada centre podria triar, d'acord amb els seus criteris, quins usuaris s'han de capturar o no. Amb aquest tipus de mostreig cal tenir present que hi ha diferents variables que influeixen en la qualitat de la selecció. En cas que ens centrem en criteris basats en la quantitat de seguidors per seleccionar un perfil per capturar-ne els tuits s'ha de tenir en compte que pot ser que hagin comprat seguidors.

4.2 Eina de captura

En els paràgrafs anteriors s'han analitzat breument les eines existents. Algunes s'haurien de millorar si es volen utilitzar, mentre que d'altres no són recomanables perquè d'una institució governamental les preservi a llarg termini.

Per poder arxivar els tuits dins del seu context, es necessita un connector que complementi Heritrix amb la funcionalitat de capturar informació emmagatzemada amb la tecnologia AJAX.

Tot i així, amb aquest connector es resol el problema només temporalment, ja que Twitter canvia de tecnologia amb molta freqüència. Per aquesta raó, les biblioteques consultades capturen només la pàgina principal dels perfils amb els últims tuits publicats i es perd la interacció i la relació entre tuits que segueixen un fil argumental.

4.3 Filtrar l'idioma

Una altra manera de filtrar per idioma, a part del Web Analyzer, és el Twitter Search API, que pot arribar a ser una solució per filtrar l'idioma gràcies als operadors "language" i "place" de l'eina. Així, és molt més probable que un tuit fet a Olot estigui escrit en català que no pas un d'escrit a Madrid. L'estratègia consistiria a fer la cerca de tots els tuits fets des de les diferents localitats de Catalunya i capturar-ne el resultat en HTML, com es pot veure en la figura 5.

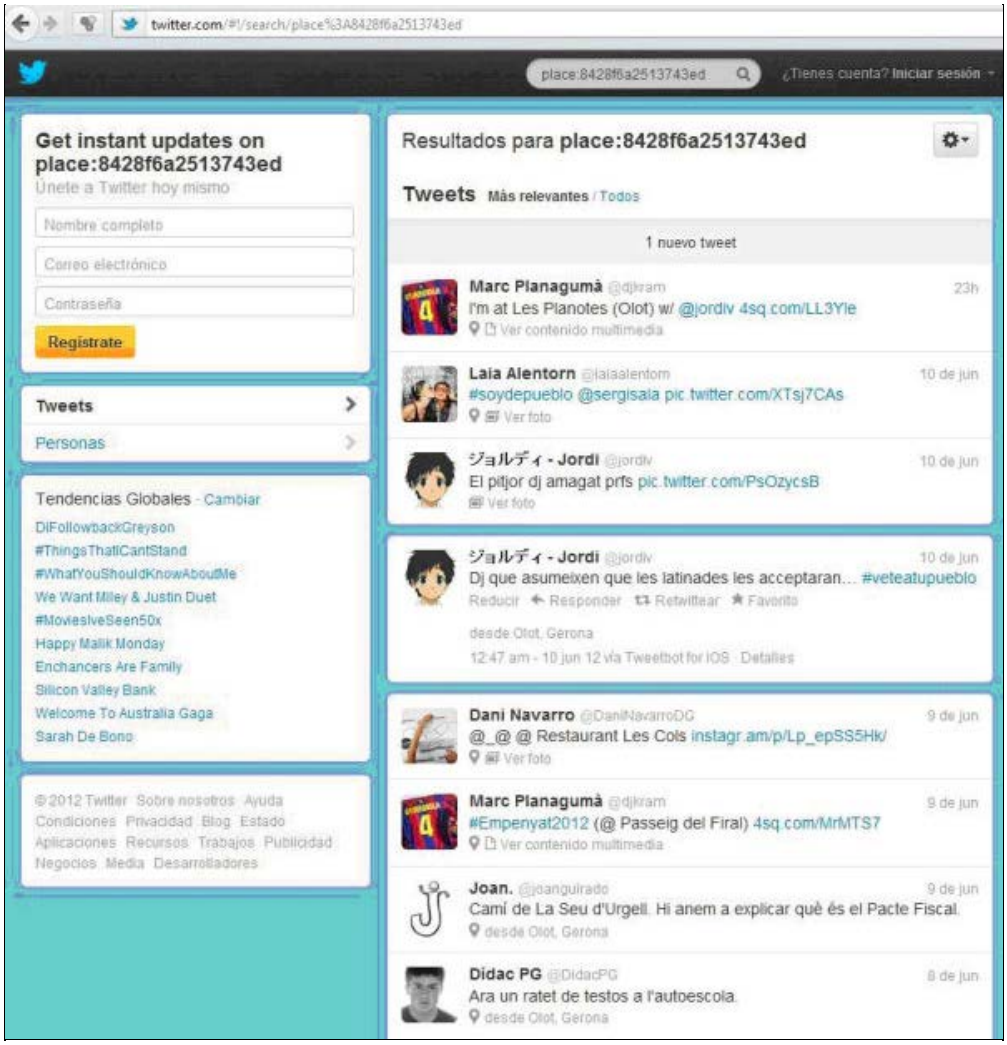

Figura 5. Captura de pantalla en què es poden veure els tuits obtinguts amb la cerca amb l'operador "places" corresponent a Olot (Girona) de la Search API de Twitter

4.4 Col·laboració i ús de serveis externs

Tal com s'ha comentat a priori en els problemes legals, Twitter ha donat la seva base de dades amb els tuits públics a la LC. Per contracte, la LC no la pot cedir ni sencera ni en part a tercers. Per tant, una altra de les propostes seria continuar treballant amb aquesta institució i a través d'IIPC per desenvolupar més projectes comuns per facilitar aquestes dades a l'usuari català.

4.5 Format de preservació

En el moment de triar el format de preservació tenim a disposició PDF, JSON o HTML.

- PDF: és un dels formats considerats sostenibles que manté els enllaços actius. El desavantatge és que es tracta d'un format propietari i, a més, com que està generat per Backupify, es perden part de les metadades i l'aspecte visual de Twitter.

- JSON: és un format obert de JAVA que es pot guardar en format de text pla. L'avantatge principal és que pesa poc, és flexible i no es perden les metadades de Twitter. Com a desavantatge principal, trobem que és un format massa nou i encara no està prou analitzat. El format ni consta en la base de dades d'informació tècnica de formats de fitxers per a la preservació PRONOM ni en la llista de formats analitzats per la Library of Congress, encara que Heritrix pot capturar fitxers de tipus aplication/x-JavaScript i fitxers text/plain (Llueca *et al.*, 2010) i que COFRE accepta aquest format. Per tant, aquest és un format estàndard[23] que pot ser accessible en el futur, i en cas necessari se'n podrà fer una migració de les dades.

- HTML: és un format estàndard àmpliament utilitzat arreu del món. Com a avantatge, trobem que és el format per defecte de captura de Heritrix i és capaç de guardar l'aspecte visual del compte en Twitter. No obstant això, el programari actual de visualització presenta problemes a l'hora de mostrar les captures de Twitter en HTML. És probable que, en un futur, es desenvolupi una tecnologia que millori la visualització i interacció amb les captures d'HTML, que és un format interactiu per naturalesa.

4.6 Freqüència de captura

A l'hora de triar una freqüència de captura s'ha d'adaptar al ritme d'actualització dels usuaris. Per exemple, hi ha alguns polítics que tuitegen molt (a dia d'avui més de 15.000 tuits) i d'altres que potser no arriben als 1.000 tuits. Per aquest motiu, la freqüència de captura no hauria de ser la mateixa, i caldria trobar un punt intermedi per poder ser prou representatius en tots els casos possibles. Una proposta és augmentar la freqüència abans dels actes previsibles, com ara eleccions, congressos científics o esdeveniments socials.

És molt important disposar també de prou recursos que permetin la flexibilitat necessària per capturar esdeveniments no previstos, com ara el #15m o el #novullpagar.

5 Estratègies proposades

El punt de partida d'aquest article va ser l'observació de l'ús creixent de Twitter a Catalunya. Es va voler conèixer quina era la postura de la BC sobre aquesta qüestió: si s'estaven preservant o no tuits i de quina manera. Partint d'aquesta premissa, volíem saber com es podien preservar els tuits i fer una proposta a la BC que inclogués només la selecció i la captura de tuits. Pel que fa a la descripció, la validació de les dades capturades i indexades (control de qualitat), la migració de formats i suports, la seguretat d'aquestes dades (que no es corrompin, que no desapareguin per errades humanes) i el control de processos, són temes que es podrien desenvolupar en un altre article.

S'han analitzat quines eines i quins serveis servirien per oferir solucions al problema de Twitter. Un cop feta l'anàlisi, s'ha vist que totes presenten mancances des de la perspectiva de la preservació. Juntament amb els problemes tecnològics i legals que es trobin les institucions que preserven tuits, es poden fer les recomanacions que es proposen a continuació.

S'ha de tenir en compte que actualment s'estan fent proves i encara no s'ha trobat cap solució ideal per preservar tuits. La que nosaltres proposem està condicionada per la situació econòmica actual del nostre país i per la urgència de preservació d'aquesta eina nascuda digitalment i de la qual "desapareix" la informació de manera vertiginosa. També proposem automatitzar al màxim aquest procés, per poder optimitzar els recursos existents. Per aquest motiu, escollim fer-ne una mostra exhaustiva abans d'una de selectiva.

La mostra exhaustiva planteja el problema següent: com es pot delimitar el *tuitvers* (univers català a Twitter)? Com es pot saber quina és la població catalana a Twitter? Ara per ara, no hi ha manera de conèixer aquesta dada. Una possible estratègia és l'ús dels operadors "language" i "place" de la Search API. S'hauria de desenvolupar una aplicació que alimenti Heritrix amb les URL dels resultats de la cerca de tuits per a cadascuna de les localitats catalanes. També es podrien capturar els comptes de diverses institucions catalanes i els seus seguidors:

- Tuits públics de l'Administració
- Entitats catalanes
- Portals
- Mitjans de comunicació d'àmbit geogràfic català
- Museus catalans
- Biblioteques i arxius
- Partits polítics catalans

Aquesta selecció es podria complementar amb un compte de Twitter específic de preservació de la BC (@BCtuitscatalans, @PADICAT o un compte semblant), i que es capturin els comptes dels seus seguidors. Aquest compte s'hauria de difondre entre totes aquestes entitats i els seus seguidors; a més, hauria de quedar clar entre els seguidors que seguint aquest compte els seus tuits es capturen per preservar-los. Això resoldria també un dels problemes ètics plantejats en el punt 3.1.3.

Un cop elaborada la llista, el primer que s'ha de fer és analitzar-la automàticament amb eines que evitin capturar dues vegades el mateix usuari, i també que permetin identificar els tuits escrits en llengua catalana. Ara mateix, Heritrix no permet un control de fitxers duplicats, però en la propera versió del programari s'espera que ja estigui implementat (actualment es treballa amb la versió 1.14.4).

Una vegada els tuits dels comptes seleccionats s'han capturat i integrat dins la base de dades de la BC (que no està limitada per fitxers robot.txt), dependrà de l'estructura d'aquesta base de dades i de la seva indexació que es puguin recuperar els tuits per etiqueta. Si la base de dades conté els tuits en JSON, la cerca no hauria de ser un problema. No obstant això, la manera com es capturen actualment (en HTML) i amb els recursos de cerca a disposició de PADICAT, no es poden distingir els comptes de Twitter dels d'altres pàgines web ni tampoc les etiquetes de les paraules clau en general. A més, si la BC té el màxim de tuits capturats, ja no és necessari que els bibliotecaris segueixin els esdeveniments i, per tant, es pot deixar aquesta feina als investigadors que poden descobrir tendències a posteriori. Amb aquesta informació es podria, més endavant, fer una anàlisi lingüística, política o cultural, entre d'altres.

El punt següent seria valorar quina freqüència de captura s'hauria d'establir. Heritrix permet capturar els 20 últims tuits de cada usuari, que són els que apareixen en la primera càrrega de cada perfil. La Generalitat de Catalunya, per exemple, recomana un màxim de deu tuits al dia (Generalitat de Catalunya, 2012) i, per tant, el que es podria fer

és una captura cada dos dies (dos dies són, aproximadament, vint tuits), ja que així es capturaria el màxim possible d'informació. S'entén que no tothom ha de seguir les recomanacions de la Generalitat, però ens permeten fer un càlcul aproximat.

El volum d'aquests tuits proposats és molt gran tenint en compte els caràcters i les metadades de JSON.[24] Hem calculat que el pes d'un tuit és d'1,5 kB, aproximadament. Fent un càlcul aproximat amb els comptes abans mencionats, ens trobem que capturar els comptes de 2 milions d'usuaris, en format JSON, té un pes de 60 GB per captura. Si es programa una captura cada dos dies, es tindrien 10,8 TB l'any de tuits capturats.

Si es captura tot amb Heritrix, PADICAT ho tindrà tot en una mateixa base de dades i alguns enllaços a webs externs es podran mantenir actius. A llarg termini, la BC podria crear una eina similar al Web Analyzer per filtrar els tuits escrits en català i a Heritrix per capturar-los.

Atès que el Web Analyzer no és programari lliure sinó un producte intern de la Biblioteca de la República Txeca, no es pot descarregar. Per això, idealment, es necessitaria un informàtic per poder adaptar aquest codi a les necessitats del centre o per negociar amb els productors de l'aplicació. Probablement seria més fàcil crear un programa des de zero i tenir un informàtic a la plantilla que el desenvolupés.

Amb la traducció de la interfície de Twitter al català, s'hi afegeix la metadada de l'idioma o el lloc. La BC es podria estalviar el desenvolupament d'una eina de filtratge d'idioma, encara que el resultat no seria el mateix. S'ha de tenir en compte que tant Twitter com Data'n'press són empreses privades i que, en qualsevol moment, poden decidir deixar de donar suport a qualsevol de les seves API o dels seus projectes. Dependre tant d'una empresa no és el més recomanable.

Des del 2011 Twitter facilita l'opció d'afegir vídeos, imatges i documents i, des del 2012, permet fins i tot incrustar-los en el mateix tuit. Per això, es podria dir que si es volgués que aquests fitxers fossin capturats per la BC, aquesta última hauria de preveure un espai d'emmagatzematge vint cops més gran, com a mínim.

Creiem que la nostra proposta s'hauria d'incloure en els processos ja existents de selecció i captura de pàgines web dins de PADICAT. I d'aquesta manera continuar amb el flux de treball de preservació que ja s'està duent a terme a la BC dins del sistema COFRE.

Finalment, volem puntualitzar que l'ideal seria que la BC oferís un servei similar al d'Archive-it a les institucions catalanes, per crear les pròpies col·leccions web amb les xarxes socials incloses. Les institucions podrien utilitzar aquest nou servei ofert per una institució fiable i local per crear i gestionar via web i sense descarregar-se cap programari ni mantenir cap servidor o emmagatzematge. A més, seria un ingrés econòmic per a la mateixa biblioteca.

Agraïments

Agraïm la col·laboració de totes les institucions consultades. Especialment, de Ciro Llueca (PADICAT), Maria del Mar Pérez Almenta (Seidor) i Jordi Linares (UPC), per la seva paciència a l'hora de donar resposta a tots els nostres dubtes, que no han estat pocs.

Bibliografia

Banks, Markus (2009). "Chapter 14. Blogs posts and tweets: the next frontier for grey literature". *Future Trends*. <http://eprints.rclis.org/bitstream/10760/15411/9/5%2014%20Banks.pdf>. [Consulta: 17/05/2012].

Boyd, Danah M.; Ellison, Nicole B. (2007). "Social network sites: definition, history, and scholarship". *Journal of computer-mediated communication*, vol. 13, no. 1.

Bruns, Axel; Burgess, Jean (2011). *New methodologies for researching news discussion on Twitter*. <http://snurb.info/files/2011/New%20Methodologies%20for%20Researching%20News%20Discussion%20on%20Twitter%20(final).pdf>. [Consulta: 17/05/2012].

Castello, Kaitlin L.; Priem, Jason (2008). "Archiving scholars' tweets". *Society of American Archivist ? 2010 Research Forum*. <http://www2.archivists.org/sites/all/files/KCFinal.pdf>. [Consulta: 17/05/2012].

Generalitat de Catalunya (2012). *Guia d'usos i estil a les xarxes socials de la Generalitat de Catalunya*. 5a ed. <http://www.gencat.cat/web/meugencat/documents/guia_usos_xarxa.pdf>. [Consulta: 01/06/2012].

Gomes, Daniel; Freitas, Sérgio; Silva, Mário J. (2006). "Design and selection criteria for a national web archive". *ECDL?06 Proceedings of the 10th European conference on research and advanced technology for digital libraries*. Berlin: Springer-Verlag, p. 196?207. <http://dl.acm.org/citation.cfm?id=2111175>. [Consulta: 17/05/2012].

Java, Akshay *et al.* (2007). "Why we Twitter: understanting microblooging usage and communities". <http://aisl.umbc.edu/resources/369.pdf>. [Consulta: 17/05/2012].

Kelly, B. *et al.* (2010). "Twitter archiving using Twapper Keeper: technical and policy challenges". Poster presented on September 20, 2010 at the *7th International conference on preservation of digital objects (iPRES2010)*, Vienna,

Austria. <http://es.scribd.com/doc/36393115/Twitter-Archiving-Using-Twapper-Keeper-Technical-And-Policy-Challenges>. [Consulta: 17/05/2012].

Llueca, Ciro *et al.* (2010). "El PADICAT: l'experiència catalana en l'arxiu d'Internet". *Lligall*, núm. 31, p. 143?161. <http://eprints.rclis.org/bitstream/10760/16246/1/llueca_lligall_31_2010_padicat.pdf>. [Consulta: 17/05/2012].

Llueca, Ciro *et al.* (2011). "A ritmo de tweet: archivando elecciones 2.0". *El profesional de la información*, vol. 20, n.° 3 (mayo?junio), p. 309?314. <http://eprints.rclis.org/handle/10760/15764>. [Consulta: 17/05/2012].

Mohr, G. *et al.* (2004). "An introduction to Heritrix: an open source archival quality web crawler". *4th International web archiving workshop*, p. 1?15. <http://project.management6.com/An-Introduction-to-Heritrix-download-w17935.pdf>. [Consulta: 17/05/2012].

Niu, Jinfang (2012). "An overview of web archiving". *D-Lib magazine*, vol. 18, no. 3?4 (March?April). <http://www.dlib.org/dlib/march12/niu/03niu1.html>. [Consulta: 17/05/2012].

O'Keeffe, Hope (2011). "Legal issues in building social media collections". Association of Research Libraries, May 2011. <http://www.arl.org/bm~doc/mm11sp-okeeffe.pdf>. [Consulta: 17/05/2012].

Pérez, Karibel; Serra, Eugènia (2010). "Repositori de preservació digital de la Biblioteca de Catalunya: informe descriptiu i de situació". <http://www.recercat.net/handle/2072/97251>. [Consulta: 05/06/2012].

Serra, Eugènia; Pérez, Karibel; Llueca, Ciro (2011). "La Biblioteca de Catalunya i l'accés al patrimoni digital". *Métodos de información (MEI)*, II època, vol. 2, núm. 2, p. 5?20. <http://eprints.rclis.org/handle/10760/16003>. [Consulta: 17/05/2012].

Vlcek, Ivan (2008). "Identification and archiving of the Czech web outside the national domain". *IWAW '08: 8th International workshop for web archiving*, September 18?19, 2008, Aarhus, Denmark. <http://iwaw.europarchive.org/08/IWAW2008-Vlcek.pdf>. [Consulta: 17/05/2012].

Data de recepció: 11/06/2012. Data d'acceptació: 03/10/2012.

Notes

[1] S'entén per *seguidor* una persona que escull llegir els tuits d'un compte; és similar als admiradors o amics d'altres xarxes socials. Les relacions es poden donar en un únic sentit o ser recíproques.

[2] Ens basem en dos indicadors que són: els seixanta-tres comptes de la Generalitat de Catalunya <http://www.gencat.cat/xarxessocials/ca/directori-xarxes-gencat-departaments.html>, i les dades publicades pel portal Twitter en Català que dóna estadístiques del nombre de tuitaires que des de juliol del 2012 fan servir l'eina en català.

[3] Un estudi breu sobre els arxius web es pot trobar a Niu, 2012.

[4] Informació extreta del blog del Twitter. <http://blog.twitter.com/2011/03/numbers.html>. [Consulta: 26/04/2012].

[5] Twitter en Català.

[6] Graells, Jordi; Xaudiera, Sergi (2011). *La Generalitat de Catalunya a les xarxes socials*. Curs al Departament de la Presidència de la Generalitat de Catalunya, 27 i 29 de juny del 2011. <http://www.slideshare.net/jordigraells/la-generalitat-a-les-xarxes-socials-8498924>. [Consulta: 23/09/2012]. A partir de la diapositiva 84 es pot extreure el nombre de tuits, retuits i seguidors de tots els comptes de la Generalitat el juny del 2011.

[7] Raymond, Matt (2010). "How tweet it is!: library acquires entire Twitter Archive". *Library of Congress Blog*. <http://blogs.loc.gov/loc/2010/04/how-tweet-it-is-library-acquires-entire-twitter-archive/>. [Consulta: 23/09/2012].

[8] Vegeu la relació de centres consultats en l'apèndix.

[9] El 15M, també denominat *moviment dels indignats*, fa referència a la data del 15 de maig del 2011, quan es van iniciar un seguit de protestes populars en contra de les actuacions del Govern espanyol. <http://movimiento15m.org/ >. [Consulta: 23/09/2012].

[10] No vull pagar és una campanya per denunciar el pagament dels peatges a Catalunya malgrat haver finalitzat el pagament del cost de construcció; consisteix a passar per les autopistes sense pagar. <http://www.novullpagar.cat/p/don-neix-aquesta-iniciativa.html>. [Consulta: 23/09/2012].

[11] Data'n'press és una petita empresa que treballa en l'anàlisi de dades en l'àmbit del periodisme; va sorgir per poder recollir dades sobre els catalans a Twitter.

[12] Twit.cat era un portal, creat per iniciativa de l'empresa Initec, que aplegava els tuits dels seus seguidors i que calculava les etiquetes més utilitzades i els seguidors més seguits. Des de la traducció de Twitter al català, ha deixat de donar aquesta informació. "El nou portal Twit.cat aplega els missatges a Twitter dels usuaris catalans". <http://www.324.cat/noticia/702886/societat/El-nou-portal-Twitcat-aplega-els-missatges-a-Twitter-dels-usuaris-catalans>.

[Consulta: 25/09/2012].

[13] JSON és el format d'objectes en JAVA. Un dels seus usos és el d'intercanvi entre client i servidor que suporta l'HTML, i és el format utilitzat per Twitter.

[14] Archive-it-it.org utilitza les mateixes eines de codi obert que el PADICAT: Heritrix, Wayback Machine, NutchWax i Solr.

[15] Basic HTTP Authentication (procediment per demanar l'usuari i la contrasenya que s'envien al servidor codificats en base64) i OAuth (protocol que pot delegar l'autenticació a una interfície de programació d'aplicacions de manera que l'usuari pot autenticar-se amb uns bits aleatoris subministrats pel servidor), donen una solució oberta i estàndard a la implementació de l'autenticació d'usuaris per a aplicacions web, i és segura només si es fa servir mitjançant el protocol HTTPS.

[16] AJAX (Asynchronous JavaScript and XML) és una tecnologia per desenvolupar aplicacions web interactives que permeten enviar peticions al servidor tant de manera síncrona com asíncrona, sense interferir en el comportament de la pàgina web; és a dir, tot i que la pàgina web carrega dades sembla que la pàgina sigui estàtica. Exemple: Twitter només ens mostra els X primers tuits, però si es continua baixant, en carrega més però mai no es perden de vista els primers missatges carregats. Aquesta tecnologia s'aplica mitjançant el llenguatge JavaScript utilitzant l'XMLHttpRequest del DOM (Document Object Model és una interfície de programació d'aplicacions (API) per accedir a contingut estructurat en documents de llenguatges estàndard ISO16262, llenguatge més utilitzat a JavaScript, inserir-lo i canviar-lo dinàmicament) per enviar peticions HTTP o HTTPS directament al servidor. La resposta del servidor es recupera en JavaScript com a text pla o com a document XML.

[17] S'està creant jurisprudència en aquest sentit a partir del cas del ciutadà Harris: "Twitter turns over user's messages in occupy Wall Street Protest Case" (notícia apareguda en *The New York Times* el 14 de setembre del 2012).

[18] Costa, Jason (2011). "Platform partner spotlight: mass relevance and Crimson Hexagon". *Build with Twitter* (blog). <https://dev.twitter.com/blog/platform-partner-spotlight-mass-relevance-and-crimson-hexagon>. [Consulta: 17/05/2012].

[19] Tsotsis, Alexia (2011). "Twitter and Mediasift partner to resell firehose data". *Tech crunch* (blog). <http://techcrunch.com/2011/04/04/twitter-and-mediasift-announce-partnership/>. [Consulta: 17/05/2012].

[20] Raymond, Matt (2010). "The Library and Twitter: an FAQ". *Library of Congress Blog*. <http://blogs.loc.gov/loc/2010/04/the-library-and-twitter-an-faq/>. [Consulta: 17/05/2012].

[21] D'acord amb la Llei 4/1993, de 18 de març, del sistema bibliotecari de Catalunya, la Biblioteca de Catalunya recull, conserva i difon la producció bibliogràfica catalana i la relacionada amb l'àmbit lingüístic català, i vetlla per la conservació i la difusió del patrimoni bibliogràfic. La mateixa biblioteca interpreta la llei de manera que inclou documents digitals.

[22] Aquesta pràctica és coneguda i comentada en els mitjans de comunicació, en articles com ara "Compro seguidors" (*La Vanguardia*, 29 d'abril del 2012) o "Perfiles con muchos huevos" (*El País*, 20 d'abril del 2012).

[23] Dins de l'estàndard tecnològic Ecma-262, equivalent a l'estàndard internacional ISO/IEC 16262:2011. Ecma International, trobem l'ECMAScript Language Specification (5.1 Edition. June 2011).

[24] Krikorian, Raffi (2010). "Map of a Twitter Status Object". <http://mehack.com/map-of-a-twitter-status-object>.

Apèndix

Relació de biblioteques i centres contactats. Els centres que ens van respondre estan assenyalats en negreta.

1 Bibliothèque Nationale de France
2 Library and Archives Canada
3 Institut National de l'Audiovisuel (Ina)
4 PADICAT (Biblioteca de Catalunya)
5 British Library
6 California Digital Library
7 The National Library of Finland
8 Netarchive.dk (Dinamarca)
9 The National Library of Norway
10 Swiss National Library
11 Austrian National Library
12 National and University Library of Slovenia
13 Alexandrina (Egipte)
14 National and University Library of Croatia
15 National Library of Korea
16 The National Archives (Regne Unit)
17 National Diet Library (Japó)
18 National Library of the Czech Republic
19 Columbia University Libraries
20 Harvard University Library
21 National and University Library of Iceland
22 National Library of New Zealand
23 National Library of Australia
24 Koninklijke Bibliotheek (Holanda)
25 Library of Congress (Estats Units)

Facultat de Biblioteconomia i Documentació
Universitat de Barcelona
Barcelona, desembre de 2012
http://www.ub.edu/biblio • Comentaris