


Seguidnos: estrategias de captura de tweets en catalán

[\[Versión castellana\]](#)

ANITA E. LOCHER
Facultat de Biblioteconomia i Documentació
Universitat de Barcelona
alocher@ub.edu

AINA GIONES-VALLS 
VTLS Europe
gionesa@vtlseurope.com

ELI RAMÍREZ 
Bibliotecaria freelance
eliramirez82@gmail.com

GRELDA ORTIZ
Real Academia de Ciencias Económicas y Financieras
biblioteca@racef.es

Opciones

 Imprimir  Recomandar  Citació  Estadístiques  <meta />  Similars

Resumen [\[Abstract\]](#) [\[Resum\]](#)

Twitter es una herramienta de *microblogging* utilizada en todo el mundo para generar y compartir información. La comunidad catalana la utiliza desde el Govern de la Generalitat hasta los medios de comunicación, pasando por personas con diferente formación y estilos de vida. Sin embargo, a día de hoy, en Catalunya, hay pocas actuaciones claras para preservarla desde una perspectiva de gestión bibliotecaria y documental de Catalunya. Este artículo propone una estrategia de captura de tuits para la Biblioteca de Catalunya, que incluye criterios de selección, metadatos de preservación y previsiones de crecimiento.

Metodología: Se analizan los aspectos legales y tecnológicos de Twitter, de la información del archivo Twitter de la Library of Congress, del portal con estadísticas Twitter'n'Català y de PADICAT (Patrimonio Digital de Catalunya), para describir el estado de la cuestión de las posibilidades de preservación los tuits que tiene la Biblioteca de Catalunya.

Resultados: Los resultados recogidos muestran que actualmente utilizar una única herramienta no es suficiente para filtrar, capturar y preservar los tuits. Los requerimientos de una institución que tiene como objetivo la preservación a largo plazo hacen parecer insuficientes las herramientas existentes en Internet. No obstante, la necesidad de actuar inmediatamente para no perder el patrimonio digital hace que se tengan que utilizar las herramientas y los recursos disponibles. Junto con una combinación de herramientas existentes se pueden mejorar las estrategias de captura de tuits actuales.

1 Introducción

[Twitter](#) es una herramienta de *microblogging* que sirve para publicar brevemente, con ciento cuarenta caracteres, un mensaje de estado –también conocido como *actualización*– para personas y entidades. Esta herramienta permite seguir las actualizaciones que interesan a los usuarios.

Como herramienta social (Boyd, Ellison, 2007) cumple los requisitos siguientes:

- Es un servicio basado en la web.
- Permite al usuario crear un perfil público o semipúblico.
- Establece una lista de usuarios con los que se comparte una relación.
- Ver las listas de las conexiones propias y de las de otros usuarios.

Cuando alguien se quiere dar de alta en Twitter necesario que proporcione al sistema algunos datos (nombre, correo electrónico y nombre de usuario). Estos datos forman parte del perfil, que puede complementar con una descripción, una imagen y un lugar que son públicos, así como los seguidores que tiene el usuario y los que sigue.¹

Cada mensaje publicado es un tuit o twitts, que puede contener (o no) los siguientes elementos:

- @ Usuario: sirve para mencionar un usuario (responder, citar, etc.).
- # Etiqueta (*hashtag*): identifica una palabra o una expresión significativa para un tuit y sirve para recuperarlas

- posteriormente.
- RT: republicar contenidos que ha publicado otro usuario.
- También se permite incluir texto, URL, fotografías, vídeos, etc.

Twitter es un recurso nacido digitalmente, publicado en abierto y en tiempo real y que se utiliza ampliamente en Catalunya.² Es una fuente de información de primera mano sobre eventos y noticias de actualidad de carácter no oficial, en el que la sociedad refleja hechos sociales, concentraciones, actividades, estados, de manera efímera (Bruns; Burgess, 2011). En este artículo queremos investigar cómo y qué se está preservando. En general, las entidades que se encargan de la preservación digital seleccionan, capturan y describen los archivos web y hay acceso dependiendo de la legislación de cada país.³ También cabe destacar que los problemas encontrados en la mayoría de centros son comunes a todas las instituciones, se trata de problemas tecnológicos y falta de recursos humanos (Castello; Priem, 2008).

Catalunya no es una excepción y desde hace tiempo preserva archivos web digitalmente a través de algunos proyectos, como la [Memoria Digital de Catalunya](#) y [ARCA](#), entre otros, desarrollados por la Biblioteca de Catalunya (BC) (Serra; Pérez; Lluca, 2011; Lluca *et al.* 2010). Con el fin de garantizar la perdurabilidad de los datos de diferentes proyectos digitales, se ha creado el sistema COFRE (COnservemos para el Futuro Recursos Electrónicos), un repositorio de preservación digital de alta seguridad (Serra; Pérez; Lluca, 2011). En el caso de las páginas web catalanas, [PADICAT](#) (Lluca *et al.*, 2010) captura el dominio .cat semestralmente, mientras que otras páginas web que tratan de temas seleccionados por los profesionales de la información se capturan periódicamente. En esta selección temática se incluyen las cuentas públicas de Twitter de algunos políticos y partidos (Lluca *et al.*, 2011). Es decir, que la BC forma parte de las bibliotecas en la vanguardia porque ya está capturando contenido de Twitter. Sin embargo, en este artículo proponemos dar un paso más en este proceso de captura.

En estos últimos dos años ha aumentado notablemente el número de usuarios de Twitter. Sólo en un día del mes de febrero del 2011 se crearon una media de 460.000 cuentas nuevas que llegaron a generar una media de 140.000 tuits al día.⁴

Desde julio de 2012, el catalán es un idioma oficial en Twitter. Actualmente tiene más de 52.000 usuarios.⁵

1.1 ¿Por qué es importante preservar tuits catalanes?

Algunas instituciones públicas, como la Generalitat de Catalunya,⁶ ayuntamientos o entidades culturales, han elegido crear un perfil en Twitter como un medio de comunicación más para interactuar con la gran cantidad de catalanes que cada vez más utilizan esta herramienta. La información que se genera a través de estos tuits puede ser interesante para futuros estudios sociológicos, históricos, lingüísticos, estadísticos o periodísticos, entre otros. Investigadores como Banks (2009) consideran que los tuits son literatura gris. Por este motivo, su preservación forma parte ya del patrimonio documental catalán.

Este artículo quiere llamar la atención sobre la rapidez con la que desaparecen los tuits, si se tiene en cuenta que Twitter sólo permite recuperar los últimos 3.200 tuits de un usuario y 1.500 tuits por cada búsqueda, ya sea etiqueta o texto libre. Por ejemplo, la Generalitat de Catalunya, en su perfil @gencat, publicó su primer tuit el 29 de marzo de 2009. Ahora, que tiene 5.181 tuits, ya no se pueden recuperar los primeros. Si la BC no actúa, deberá depender de la Library of Congress (LC) para acceder porque es quien los conserva todos por ahora.⁷

2 Metodología

Para conocer qué otras iniciativas llevan a cabo centros similares en la BC todo el mundo, se contactó con las diferentes instituciones que forman parte del [International Internet Preservation Consortium](#) (en adelante, IIPC). El IIPC fomenta el intercambio de experiencias en preservación web llevadas a cabo por las diferentes entidades: bibliotecas, archivos, universidades y entidades privadas.

Nos pusimos en contacto con veintiséis miembros del IIPC,⁸ de los cuales contestaron catorce, y de acuerdo con esta información hemos desarrollado las propuestas para la BC. El objetivo era averiguar cuáles eran las tecnologías en uso y los criterios de inclusión o exclusión de las redes sociales y especialmente de Twitter. En el primer correo electrónico se plantearon las siguientes preguntas:

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?*

3 Resultados

Según las respuestas recibidas, se ha podido observar que son pocos los centros que capturan tuits con regularidad. En los centros donde se han hecho pruebas, la captura no ha sido satisfactoria o ha presentado errores.

Los centros que capturan tuits lo hacen porque estos mensajes están relacionados con otras páginas y no para

hacer una muestra representativa de la red social y de aquello de lo que se habla. Las capturas, según estos criterios, no son representativas para la red, ya que son fruto de una captura de un momento concreto, y en Twitter la actualización es constante y la frecuencia de captura no está a la altura de las circunstancias.

Estas instituciones capturan los tuits o bien porque forman parte de páginas de un dominio seleccionado, o bien porque son de un usuario relevante.

La intención de los archivos no es, en ningún caso, crear un archivo de Twitter separado, desconectado del archivo web, más bien quieren capturar todas las formas de expresión relacionadas con un tema concreto y hacerlas accesibles en conjunto. En las políticas de colección no se mencionan explícitamente las redes sociales. Este hecho se puede explicar de dos maneras:

1. Las políticas se han hecho en momentos en que las redes sociales no existían o no tenían un papel muy importante.
2. Muchas de las políticas de colección son lo suficientemente generales como para incluir redes sociales.

De hecho, si se aplican los criterios de "nacionalidad" que muchas bibliotecas nacionales aplican a su fondo impreso, los tuits incluidos serían:

- Producidos en el país.
- Creados por un ciudadano del país.
- En el idioma del país.
- Sobre una temática del país.

Por tanto, no es necesario que las redes sociales estén incluidas explícitamente en las políticas de colección, aunque algunos centros tienen previsto incluirlas dentro de sus archivos de preservación web. Sólo el archivo web de la República Checa excluye intencionadamente las redes sociales por dos razones: en primer lugar, porque las consideran conversaciones privadas y, en segundo lugar, por los problemas técnicos. Los archivos que sólo capturan por dominio necesitarán nuevas estrategias para tener representadas las redes sociales de sus comunidades. En el caso de captura selectiva (en tanto que no captura todo el dominio ya que se siguen criterios temáticos, por ejemplo), el archivo puede elegir la parte que hay que preservar. El problema radica en saber qué es relevante para la institución.

Ante esta problemática se puede optar por, o bien seleccionar perfiles conocidos, o bien capturar exhaustivamente todas las publicaciones hechas en el idioma del país con métodos automáticos.

Por ejemplo, el archivo de la República Checa ha desarrollado un conector para reconocer automáticamente el idioma y para filtrar páginas de una temática concreta. Otros, como las bibliotecas nacionales de Austria y Dinamarca, capturan las páginas de perfil de Facebook y Twitter, a pesar de que quedan fuera de contexto debido a la rápida frecuencia de actualización de los usuarios.

La mayoría de las capturas se hacen una o dos veces al año, aunque en periodos electorales la frecuencia suele aumentar. Desgraciadamente, esta frecuencia resulta del todo insuficiente para convertirse en un resultado representativo del movimiento social. En ningún caso nos hemos encontrado con instituciones que detecten y capturen eventos espontáneos como el movimiento de los indignados con la etiqueta #15m⁹ o la lucha contra los precios de las autopistas catalanas con la etiqueta #novullpagar.¹⁰

A continuación, antes de exponer los problemas encontrados, explicamos algunas herramientas que han mencionado las bibliotecas consultadas o que aparecen en la bibliografía, hemos descartado las que no son de código abierto o que no permiten la exportación de los resultados. Entre estas herramientas, podemos diferenciar las de selección y las de captura.

Herramientas de selección

En el caso de archivos selectivos se pueden utilizar herramientas de selección que permitan detectar perfiles, tuits o conversaciones con cierto grado de repercusión y, de este modo, poder automatizar el proceso de selección. Estas herramientas utilizan diferentes criterios para generar la lista los más populares. Por ello, cada centro puede utilizar la herramienta con los criterios que mejor se adapten a sus necesidades, por ejemplo, número de seguidores de un usuario, popularidad, frecuencia de las actualizaciones, fecha de la última actualización o número de retuits recibidos, entre otros. Destacamos dos:

- [TwitterGrader](#) y [Twitleve](#): herramientas que sirven para crear rankings de influencia de un perfil o de un tuit dentro de la red social Twitter.
- [Twitter 'n' Català](#): proyecto de [Data'n'press](#)¹¹ que toma el relevo del portal [Twit.cat](#)¹² y que da las estadísticas de uso de Twitter utilizando el operador "language" de la herramienta y otros factores como los tuiteadores que siguen su cuenta y un estudio de la red en que detecta posibles usuarios catalanes, y se crea así una temprana base de datos.

Herramientas de captura

Son las que permiten buscar, copiar y exportar los tuits de las cuentas de los usuarios. Destacamos las siguientes:

- **Twitter Api** (*application programming interface*, interfaz de programación de aplicaciones): herramientas creadas para Twitter que permiten a los programadores desarrollar las propias aplicaciones. La REST API permite publicar microentradas en las aplicaciones, seguir a alguien o crear listas. La Search API sirve para buscar tuits en un índice de tuits recientes, no se pueden recuperar tuits más allá de una semana y sólo se recuperan los tuits considerados relevantes dentro de la búsqueda. Las Streaming API las utilizan los desarrolladores que quieren toda la secuencia de tuits en tiempo real en el mismo momento en que se publican en Twitter. Encontramos la API para la secuencia de tuits públicos (*public streams*), para la secuencia de tuits de una cuenta de un usuario (*user streams*) y para la secuencia de muchas cuentas de diferentes usuarios (*site streams*). Esta última es una herramienta muy reciente y, por ello, no todas las aplicaciones están tienen acceso. Las dos herramientas siguientes utilizan el Twitter Api como base:
 - **TwapperKeeper**: Programa financiado por JISC (Joint Information Systems Committee) del Reino Unido. Permite crear un archivo de tuits en un servidor propio en cuatro formatos diferentes: HTML, RSS, XLS y [JSON](#).¹³
 - **Backupify**: Permite exportar los contenidos de los tuits en archivos PDF indexados o ficheros JSON. Archiva hasta 1 GB de tuits en la nube de forma gratuita utilizando la infraestructura Amazon S3 (*simple storage service*). En caso de que se sobrepase el gigabyte, el servicio pasa a ser de pago. Siguiendo los límites que impone Twitter Api, sólo permite capturar los 3.200 tuits más recientes.
- **Archive-it**: Servicio ofrecido por Internet Archive, entidad que trabaja en la preservación web desde 1996. Archive-it es una aplicación web de pago que permite crear, descargar y gestionar colecciones digitales con distintos tipos de contenidos y acceder a ellos, tales como: HTML, vídeos o audio¹⁴ Cabe destacar que puede capturar redes sociales, entre ellas Twitter. Esta herramienta permite exportar la colección y recibir una copia en un disco duro con los datos capturados.
- **Heritrix**: rastreador web (*crawler*) gratuito y de código abierto que creó Internet Archive con la colaboración del Nordic Web Archive en 2003 (Mohr *et al.*, 2004) implementado en lenguaje JAVA. El formato de captura es HTML, y el de almacenamiento de la información es ARC. Es la herramienta que utiliza PADICAT.
- **Web Analyzer**: aplicación que puede integrarse en uno de los módulos de Heritrix que permite identificar el idioma de la página web o filtrarse la por tema. Está desarrollada por la Biblioteca de la República Checa (Vlcek, 2008).

3.1 Problemas encontrados

Se han identificado problemas comunes a todas las bibliotecas o todos los centros de documentación, pero también de otros que sólo afectaban a un número concreto de centros. A continuación, se presentan estos problemas agrupados en: tecnológicos, legales y éticos.

3.1.1 Problemas tecnológicos

Los problemas tecnológicos son múltiples. Lo principal es que Twitter sigue cambiando y migrando hacia nuevas tecnologías. Por ejemplo, antes utilizaba HTTP básico para la autenticación en la red, mientras que ahora se usa OAuth.¹⁵

Heritrix, usado por PADICAT y la mayoría de centros, funciona utilizando un rastreador web que recolecta la página y los enlaces que contiene. Las páginas capturadas son copias del original. PADICAT captura las cuentas en HTML – unos veinte tuits por página–, y el peso varía de un usuario a otro, así, una cuenta institucional pesa 153 MB, mientras que una personal pesa 20 KB.

Ahora bien, si se utiliza la tecnología AJAX,¹⁶ el robot no puede capturar toda la información. En este caso, el resultado se visualiza de un modo diferente de como lo ve el usuario de Internet. Twitter utiliza AJAX cuando muestra las respuestas a un tuit, por ejemplo. Como en el caso siguiente: la figura 1 muestra los tuits hechos por un usuario y que se pueden capturar con Heritrix, mientras que la figura 2 muestra la información de estos tuits que no queda guardada por Heritrix.



Figura 1. Captura de pantalla en la que se ven los tuits recolectados por Heritrix



Figura 2. Captura de pantalla de la información que no está recolectando. Cuando el usuario pincha sobre un tuit para ver las respuestas, aparece información que Heritrix no capturar

La captura de páginas de Twitter se complica por el uso del símbolo almohadilla (!) que se introduce en el identificador uniforme de recursos (URI) cuando el web utiliza AJAX, tal como se puede apreciar en la figura 3. La parte del URI que sigue después del símbolo no se envía al servidor, sino que es interpretada por el navegador. Esto impide a Heritrix capturar las subpáginas de un dominio con almohadilla.



Figura 3. Captura de pantalla en la que se señala el símbolo #! utilizado para AJAX

Se presenta un problema adicional cuando se quiere capturar una etiqueta. Primero hay que hacer una búsqueda para después capturar los tuits resultantes. Sin embargo, las páginas de resultados están bloqueadas para máquinas como Heritrix con ficheros robots.txt. Aún así, algunas bibliotecas ignoran los robots.txt y capturan la página igualmente.

Otro problema consiste en la identificación de los tuits como catalanes de manera automática. No se puede utilizar el dominio, como se hace en algunos países para seleccionar páginas web, para que todos los perfiles de Twitter

tienen el dominio.com. Se pueden utilizar los operadores "language" y "place" con la API de Twitter pero, entonces, quedan excluidos los tuits que no hayan añadido esta funcionalidad en los cuentas.

También nos encontramos que la frecuencia de actualización de Twitter es muy alta y Heritrix, de momento, sólo hace capturas dos veces al día. Si la frecuencia de actualización de las cuentas que hay que capturar aumenta, Heritrix debería poder aumentar también la frecuencia de captura. A parte, hace falta un sistema de monitorización para evitar capturar dos veces el mismo tuit en caso de que la frecuencia de captura sea superior a la de actualización, o poder, si es necesario, recuperar un tuit que no se haya capturado (Kelly *et al.*, 2010).

El último problema al que haremos referencia es que la API de Twitter sólo da como resultados los tuits más recientes y no todos los tuits que corresponden a los criterios de la búsqueda. La API sólo permite la recuperación de los últimos 1.500 tuits publicados en los últimos nueve días como resultado de una búsqueda, tal como se puede ver en la figura 4, en la que se busca una etiqueta de la que ya no se recuperan resultados, aunque que las haya.



Figura 4. Captura de pantalla en la que no se pueden recuperar los tuits existentes con la etiqueta #osr4 (4ª Jornada Open Science Repositories, que tuvo lugar en Barcelona en 2010), ya que no son recientes

3.1.2 Problemas legales

En el momento de aceptar las condiciones legales del servicio, los usuarios de Twitter son los poseedores de los derechos de autor sobre el contenido de sus tuits. En caso de litigio, los usuarios aceptan la jurisdicción de los Estados Unidos porque es donde está la sede de la empresa.¹⁷ Los usuarios ceden derechos de explotación en Twitter Inc., de modo que esta empresa puede licenciar y sublicenciar el uso de estos tuits públicos por defecto sin ninguna contraprestación económica al autor. De hecho, Twitter Inc. ya comparte estos datos con empresas externas como ahora [Crimson Hexagon](#)¹⁸ o [Mediasift](#).¹⁹

Según el [API Terms of Service](#), sólo se puede exportar contenido de Twitter en PDF o en hoja de cálculo. La exportación a una base de datos no está permitida, por lo que PADICAT no puede incluir los datos capturados con la API de Twitter.

Gomas, Freitas y Silva (2006) explican que es de difícil para un país exigir a otro que le entregue datos que se encuentren en servidores externos como es el caso de Twitter. De todas las instituciones consultadas, la Biblioteca Nacional de Dinamarca tiene una ley de Depósito legal desde 2004, en cambio, otros países, sin ley de Depósito legal, crean y mantienen archivos web.

Twitter hizo una donación de toda su base de datos a la Library of Congress.²⁰ Por contrato, no se puede hacer una distribución antes de seis meses y se debe avisar a los usuarios de Twitter Archive de la LC que no se puede hacer uso comercial ni redistribuirlos. Aunque la donación se hizo en abril de 2010, el archivo no es consultable. Aunque los derechos de autor no sean un problema, recordemos que los tuits son públicos, ahora bien, sí que se deben tener en cuenta cuestiones relacionadas con la privacidad. Además, está el problema del fichero *robots.txt*, el cual se debe ignorar si se quieren capturar los resultados de una búsqueda mediante etiquetas (O'Keeffe, 2011).

Legalmente hablando, la BC ampara bajo la Ley de bibliotecas de Catalunya y la Ley del sistema bibliotecario de Catalunya,²¹ establecidos que su misión es recopilar, conservar y difundir la producción bibliográfica catalana incluyendo la digital. De acuerdo con esta normativa, se pueden capturar perfiles públicos de Twitter de usuarios catalanes, que se incluyen dentro del patrimonio digital (Lluca *et al.*, 2010 y Serra; Pérez; Lluca, 2011).

3.1.3 Problemas éticos

La mayoría de las personas no leen los términos de servicio y muchos no saben que, cuando aceptan utilizar Twitter, están firmando un contrato. Si alguien quiere leer los términos de servicio tiene dos opciones: por un lado, leer la

versión vinculante en inglés con posibles dificultades lingüísticas, u otra leer la traducción en castellano, no vinculante, y con el riesgo de leer una versión no actualizada y, así, malinterpretar las normas de uso de Twitter.

Algunos usuarios han mostrado el desacuerdo que sea el gobierno, a través de la LC, quien capture y preserve sus tuits. Esto se manifiesta en través de la etiqueta no vinculante [#noLoC](#), que utilizan usuarios de todo el mundo.

Por otra parte, una práctica común y poco ética desarrollada por algunas empresas consiste en crear cuentas falsas de usuarios²² por hacerles seguidores de una cuenta de Twitter (marca, persona, etc.). Actualmente, aún es frecuente valorar la presencia en línea según la cantidad de contactos o seguidores y no tanto según la calidad de las relaciones que se desarrollan. Esto puede conllevar criterios erróneos a la hora de seleccionar cuentas de Twitter influyentes.

4 Estrategias generales

4.1 Estrategia de selección

Hay tres técnicas de selección para los tuits: la selección exhaustiva, el muestreo aleatorio y la selección subjetiva.

En primer lugar, una selección exhaustiva no es posible porque no se puede saber el conjunto de población de Twitter de un lugar geográfico, ya que el hecho de añadir la localización en la cuenta de un usuario no es un dato obligatorio. Una manera de saber la localización sería mediante la IP del ordenador, pero hay gente que utiliza un servidor proxy o desactiva la caché. Si se tienen en cuenta estas variables habría calcular muy bien el margen de error.

En segundo lugar, se podría utilizar una herramienta que filtrara los tuits por idioma, pero entonces habría que considerar los catalanes que tuitean en otros idiomas. Sin el conjunto de la población no se puede obtener un muestreo aleatorio. Se podría hacer, pero no sería representativo.

Por último, con la selección subjetiva, en cambio, sería más fácil de hacer porque cada centro podría elegir, de acuerdo con sus criterios, qué usuarios deben capturar o cuáles no. Con este tipo de muestreo hay que tener presente que hay diferentes variables que influyen en la calidad de la selección. En caso de que nos centramos en criterios basados en la cantidad de seguidores para seleccionar un perfil para capturar los tuits se debe tener en cuenta que puede que hayan comprado seguidores.

4.2 Herramienta de captura

En los párrafos anteriores se han analizado brevemente las herramientas existentes. Algunas deberían mejorar si se quieren utilizar, mientras que otros no son recomendables para que una institución gubernamental las preserve a largo plazo.

Para poder archivar los tuits dentro de su contexto, se necesita un conector que complemente Heritrix con la funcionalidad de capturar información almacenada con la tecnología AJAX.

Sin embargo, con este conector se resuelve el problema sólo temporalmente, ya que Twitter cambia de tecnología con mucha frecuencia. Por esta razón, las bibliotecas consultadas capturan sólo la página principal de los perfiles con los últimos tuits publicados y se pierde la interacción y la relación entre tuits que siguen un hilo argumental.

4.3 Filtrar el idioma

Otra manera de filtrar por idioma, aparte del Web Analyzer, es el Twitter Search API, que puede llegar a ser una solución para filtrar el idioma gracias a los operadores "language" y "place" de la herramienta. Así, es mucho más probable que un tuit hecho en Olot esté escrito en catalán que no uno escrito en Madrid. La estrategia consistiría en realizar la búsqueda de todos los tuits hechos desde las diferentes localidades de Catalunya y capturar el resultado en HTML, como se puede ver en la figura 5.

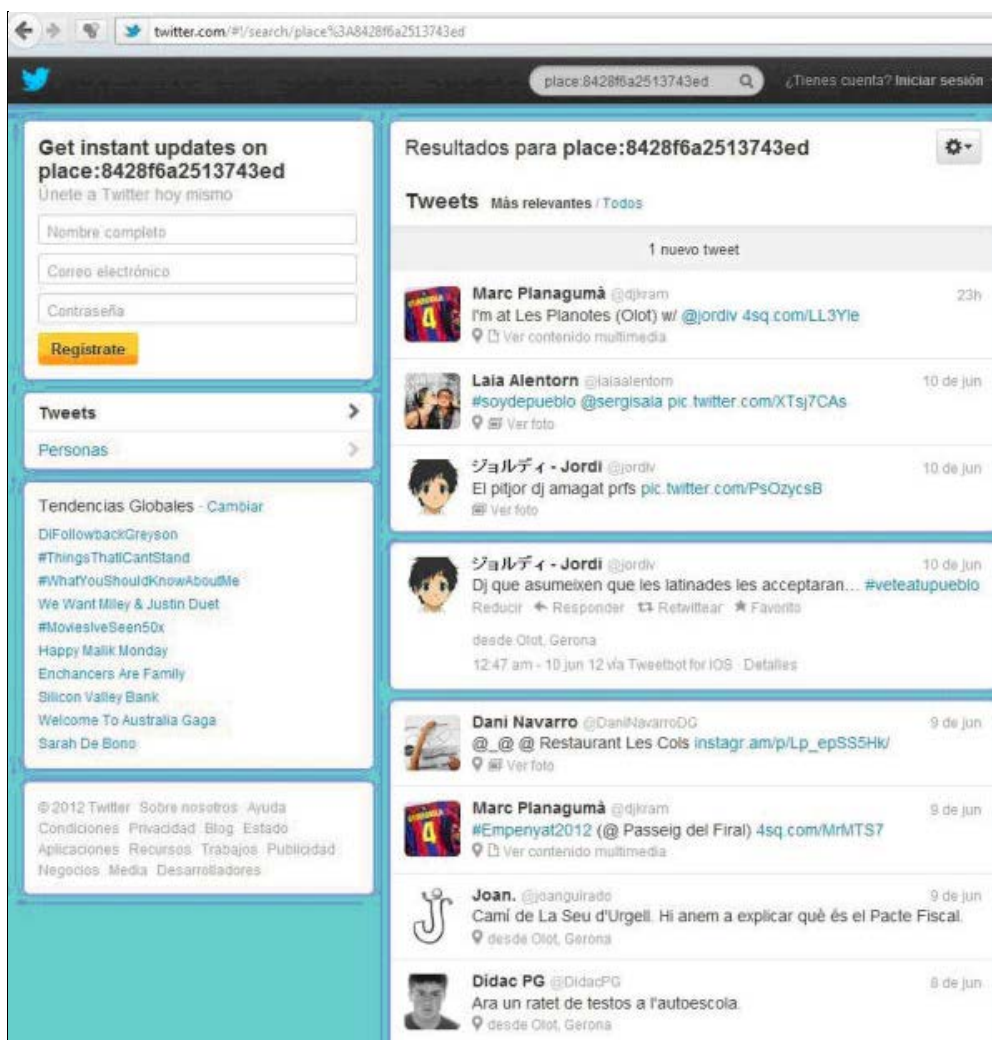


Figura 5. Captura de pantalla en la que se pueden ver los tuits obtenidos con la búsqueda con el operador "place" correspondiente a Olot (Girona) de la Search API de Twitter

4.4 Colaboración y uso de servicios externos

Tal como se ha comentado a priori en los problemas legales, Twitter ha dado su base de datos con los tuits públicos a la LC. Por contrato, la LC no puede ceder ni entera ni en parte a terceros. Por lo tanto, otra de las propuestas sería continuar trabajando con esta institución ya través de IIPC para desarrollar más proyectos comunes para facilitar estos datos al usuario catalán.

4.5 Formato de preservación

En el momento de elegir el formato de preservación tenemos a disposición PDF, JSON o HTML.

- PDF: es uno de los formatos considerados sostenibles que mantiene los enlaces activos. La desventaja es que se trata de un formato propietario y, además, al estar generado por Backupify, se pierden parte de los metadatos y el aspecto visual de Twitter.
- JSON: es un formato abierto de JAVA que se puede guardar en formato de texto plano. La ventaja principal es que pesa poco, es flexible y no se pierden los metadatos de Twitter. Como desventaja principal, encontramos que es un formato demasiado nuevo y todavía no está suficientemente analizado. El formato ni consta en la base de datos de información técnica de formatos de archivos para la preservación [PRONOM](#) ni en la lista de formatos analizados por la [Library of Congress](#). Aunque Heritrix puede capturar archivos de tipo Application / x-JavaScript y archivos text / plain (Llueca *et al.*, 2010) y que COFRE acepta este formato. Por tanto, este es un formato estándar²³ que puede ser accesible en el futuro, y en caso necesario se podrá hacer una migración de los datos.
- HTML: es un formato estándar ampliamente utilizado en todo el mundo. Como ventaja, encontramos que es el formato por defecto de captura de Heritrix y es capaz de guardar el aspecto visual de la cuenta en Twitter. Sin embargo, el software actual de visualización presenta problemas a la hora de mostrar las capturas de Twitter en HTML. Es probable que, en un futuro, se desarrolle una tecnología que mejore la visualización e interacción con las capturas de HTML, que es un formato interactivo para naturaleza.

4.6 Frecuencia de captura

A la hora de elegir una frecuencia de captura debe adaptarse al ritmo de actualización de los usuarios. Por ejemplo, hay algunos políticos que tuitean mucho (a día de hoy más de 15.000 tuits) y otros que quizá no llegan a los 1.000 tuits. Por este motivo, la frecuencia de captura no debería ser la misma, y habría que encontrar un punto intermedio para poder ser suficientemente representativos en todos los casos posibles. Una propuesta es aumentar la frecuencia antes de los actos previsibles, tales como elecciones, congresos científicos o eventos sociales.

Es muy importante disponer también de recursos suficientes que permitan la flexibilidad necesaria para capturar eventos no previstos, como por ejemplo el #15m o #novullpagar.

5 Estrategias propuestas

El punto de partida de este artículo fue la observación del uso creciente de Twitter en Catalunya. Se quiso conocer cuál era la postura de la BC sobre esta cuestión: si se estaban preservando o no tuits y de qué manera. Partiendo de esta premisa, queríamos saber cómo se podían preservar los tuits y hacer una propuesta a la BC que incluyera sólo la selección y la captura de tuits. En cuanto a la descripción, la validación de los datos capturados e indexados (control de calidad), la migración de formatos y soportes, la seguridad de estos datos (que no se corrompan, que no desaparezcan por errores humanos) y el control de procesos, son temas que se podrían desarrollar en un otro artículo.

Se han analizado qué herramientas y qué servicios servirían para ofrecer soluciones al problema de Twitter. Una vez hecho el análisis, se ha visto que todas presentan carencias desde la perspectiva de la preservación. Junto con los problemas tecnológicos y legales que se encuentren las instituciones que preservan tuits, se pueden hacer las recomendaciones que se proponen a continuación.

Se debe tener en cuenta que actualmente se están haciendo pruebas y todavía no se ha encontrado ninguna solución ideal para preservar tuits. La que nosotros proponemos está condicionada por la situación económica actual de nuestro país y por la urgencia de preservación de esta herramienta nacida digitalmente y de la que "desaparece" la información de manera vertiginosa. También proponemos automatizar al máximo este proceso, para poder optimizar los recursos existentes. Por este motivo, elegimos hacer una muestra exhaustiva antes que una de selectiva.

La muestra exhaustiva plantea el problema siguiente: ¿cómo se puede delimitar el *tuitverso* (universo catalán en Twitter)? ¿Cómo se puede saber cuál es la población catalana en Twitter? Hoy por hoy, no hay forma de conocer este dato. Una posible estrategia es el uso de los operadores "language" y "place" de la Search API. Debería desarrollar una aplicación que alimente Heritrix con las URL de los resultados de la búsqueda de tuits para cada una de las localidades catalanas. También se podrían capturar las cuentas de diversas instituciones catalanas y sus seguidores:

- Tuits públicos de la Administración
- Entidades catalanas
- Portales
- Medios de comunicación de ámbito geográfico catalán
- Museos catalanes
- Bibliotecas y archivos
- Partidos políticos catalanes

Esta selección se podría complementar con una cuenta de Twitter específico de preservación de la BC (@BCtuitscatalans, @PADICAT o una cuenta similar), y que se capturen las cuentas de sus seguidores. Esta cuenta debería difundir entre todas estas entidades y sus seguidores, además, debería quedar claro entre los seguidores que siguiendo esta cuenta sus tuits se capturan para preservarlos. Esto resolvería también uno de los problemas éticos planteados en el punto 3.1.3.

Una vez elaborada la lista, lo primero que hay que hacer es analizarla automáticamente con herramientas que eviten capturar dos veces el mismo usuario, y también que permitan identificar los tuits escritos en lengua catalana. Ahora mismo, Heritrix no permite un control de archivos duplicados, pero en la próxima versión del software se espera que ya esté implementado (actualmente se trabaja con la versión 1.14.4).

Una vez los tuits de las cuentas seleccionados han capturado e integrado en la base de datos de la BC (que no está limitada por archivos robot.txt), dependerá de la estructura de esta base de datos y de su indexación que se puedan recuperar los tuits por etiqueta. Si la base de datos contiene los tuits en JSON, la búsqueda no debería ser un problema. Sin embargo, la manera como se capturan actualmente (en HTML) y con los recursos de búsqueda a disposición de PADICAT, no se pueden distinguir las cuentas de Twitter de los de otras páginas web ni tampoco las etiquetas de las palabras clave en general. Además, si la BC tiene el máximo de tuits capturados, ya no es necesario que los bibliotecarios sigan los acontecimientos y, por tanto, se puede dejar este trabajo a los investigadores que pueden descubrir tendencias a posteriori. Con esta información se podría, más adelante, hacer un análisis lingüístico, político o cultural, entre otros.

El punto siguiente sería valorar qué frecuencia de captura debería establecerse. Heritrix permite capturar los 20 últimos tuits de cada usuario, que son los que aparecen en la primera carga de cada perfil. La Generalitat de Catalunya, por ejemplo, recomienda un máximo de diez tuits al día (Generalitat de Catalunya, 2012) y, por tanto, lo que se podría hacer es una captura cada dos días (dos días son, aproximadamente, veinte tuits), ya que así se capturaría el máximo posible de información. Se entiende que no todo el mundo debe seguir las recomendaciones de la Generalitat, pero nos permiten hacer un cálculo aproximado.

El volumen de estos tuits propuestos es muy grande teniendo en cuenta los caracteres y los metadatos de JSON.²⁴ Hemos calculado que el peso de un tuit es de 1,5 KB, aproximadamente. Haciendo un cálculo aproximado con las cuentas antes mencionados, nos encontramos con que capturar las cuentas de 2 millones de usuarios, en formato JSON, tiene un peso de 60 GB por captura. Si se programa una captura cada dos días, se tendrían 10,8 TB al año de tuits capturados.

Si se captura todo con Heritrix, PADICAT lo tendrá todo en una misma base de datos y algunos enlaces a webs externas se podrán mantener activos. A largo plazo, la BC podría crear una herramienta similar en la Web Analyzer para filtrar los tuits escritos en catalán y en Heritrix para capturarlos.

Dado que la Web Analyzer no es software libre sino un producto interno de la Biblioteca de la República Checa, no se puede descargar. Por ello, idealmente, se necesitaría un informático para poder adaptar este código a las necesidades del centro o para negociar con los productores de la aplicación. Probablemente sería más fácil crear un programa desde cero y tener un informático en la plantilla que lo desarrollara.

Con la traducción de la interfaz de Twitter al catalán, se añade el metadato del idioma o el lugar. La BC se podría ahorrar el desarrollo de una herramienta de filtrado de idioma, aunque el resultado no sería el mismo. Se debe tener en cuenta que tanto Twitter como Data'n'press son empresas privadas y que, en cualquier momento, pueden decidir dejar de apoyar cualquiera de sus API o de sus proyectos. Depender tanto de una empresa no es lo más recomendable.

Desde 2011 Twitter facilita la opción de añadir vídeos, imágenes y documentos y, desde 2012, permite incluso incrustarse en el mismo tuit. Por ello, se podría decir que si se quisiera que estos ficheros fueran capturados por la BC, esta última debería prever un espacio de almacenamiento veinte veces mayor, como mínimo.

Creemos que nuestra propuesta debería incluirse en los procesos ya existentes de selección y captura de páginas web dentro de PADICAT. Y de esta manera continuar con el flujo de trabajo de preservación que ya se está llevando a cabo en la BC dentro del sistema COFRE.

Finalmente, queremos puntualizar que lo ideal sería que la BC ofreciera un servicio similar al de Archive-it a las instituciones catalanas, para crear las propias colecciones web con las redes sociales incluidas. Las instituciones podrían utilizar este nuevo servicio ofrecido para una institución confiable y local para crear y gestionar vía web y sin descargarse ningún software ni mantener ningún servidor o almacenamiento. Además, sería un ingreso económico para la misma biblioteca.

Agradecimientos

Agradecemos la colaboración de todas las instituciones consultadas. Especialmente, de Ciro Lluca (PADICAT), Maria del Mar Pérez Almenta (Seidor) y Jordi Linares (UPC), por su paciencia a la hora de dar respuesta a todas nuestras dudas, que no han sido pocas.

Bibliografía

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](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: understanding microblogging usage and communities". <<http://ais1.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.º 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'Keefe, 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.

Notas

¹ Se entiende por seguidor una persona que escoge leer los tuits de una cuenta; es similar a los administradores o amigos en una red social. Las relaciones se pueden dar en un único sentido o recíprocamente.

² Nos basamos en dos indicadores que son: las sesenta y tres cuentas de la Generalitat de Catalunya <<http://www.gencat.cat/xarxessocials/ca/directori-xarxes-gencat-departaments.htm>>, y los datos publicados por el portal Twitter en Català que da estadísticas del número de tuiteros desde el julio del 2012 que utilizan la herramienta en catalán.

³ Un estudio sobre los archivos se puede encontrar en Niu, 2012.

⁴ Información del blog de 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 y 29 de junio del 2011. <<http://www.slideshare.net/jordigraells/la-generalitat-a-les-xarxes-socials-8498924>>. [Consulta: 23/09/2012]. A partir de la diapositiva 84 se puede extraer el número de tuits, retuits y seguidores de todas las cuentas de la Generalitat de Catalunya el junio del 2011.

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

⁸ Véase la relación de centros consultados en el apéndice.

⁹ El 15M, también denominado movimiento de os indignados, hace referencia a la fecha del 15 de mayo del 2011, cuando se iniciaron una serie de protestas populares contra las actuaciones del Gobierno español. [Consulta: 23/09/2012].

¹⁰ No vull pagar es una campaña para denunciar el pagamiento de los peajes en Catalunya ya habiendose finalizado el pagamiento del coste de construcción; consiste en pasar por las autopistas sin pagar <<http://www.novullpagar.cat/p/don-neix-aquesta-iniciativa.html>>. [Consulta:23/09/2012].

¹¹ Data'n'press es una pequeña empresa que trabaja en el análisis de datos en el ámbito del periodismo; surgió para recoger datos sobre los catalanes al Twitter.

¹² Twit.cat era un portal, creado por iniciativa de la empresa Initec, donde se podían ver tuits de sus seguidores y calculaba las etiquetas más utilizadas y los seguidores más seguidos. Desde la traducción del Twitter en catalán, ha parado de dar

esta 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 es el formato de objetos en JAVA. Uno de sus usos es el de intercambio entre cliente y servidor que soporta HTML, y es el formato utilizado por Twitter.

¹⁴ Archive-it.org utiliza las mismas herramientas de código abierto que PADICAT: Heritrix, Wayback Machine, NutchWax y Solr.

¹⁵ Basic HTTP Authentication (procedimiento para pedir el usuario y la contraseña que se envían al servidor en base64) y OAuth (protocolo que puede delegar la autenticación a una interfaz de programación de aplicaciones de manera que el usuario pueda autenticarse con bits aleatorios suministrados por el servidor), dan una solución abierta y estándar a la implementación de la autenticación de los usuarios por aplicaciones web, y es segura sólo si se hace servir mediante el protocolo HTTPS.

¹⁶ AJAX (Asynchronous JavaScript and XML) es una tecnología para desarrollar aplicaciones web interactivas que permiten enviar peticiones al servidor tanto de manera síncrona como asíncrona, sin interferir en el comportamiento de la página web, es decir, a pesar de que la página web carga datos parece que la página sea estática. Ejemplo: Twitter sólo nos muestra los X primeros tuits, pero si se continúa bajando en carga más pero nunca se pierden de vista los primeros mensajes cargados. Esta tecnología se aplica mediante el lenguaje JavaScript utilizando la XMLHttpRequest del DOM (Document Object Model es una interfaz de programación de aplicaciones (API) para acceder al contenido estructurado en documentos de lenguajes estándar ISO16262, lenguaje más utilizado en JavaScript, insertar lo y cambiarlo dinámicamente) para enviar peticiones HTTP o HTTPS directamente al servidor. La respuesta del servidor se recupera en JavaScript como texto plano o como documento XML.

¹⁷ Se está creando jurisprudencia en este sentido a partir del caso del ciudadano Harris: "[Twitter turns over user's messages in occupy Wall Street Protest Case](#)" (noticia aparecida en *The New York Times* el 14 de septiembre 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 Conforme con la [Llei 4/1993, de 18 de març, del sistema bibliotecari de Catalunya](#), la Biblioteca de Catalunya recoge, conserva y difunde la producción bibliográfica catalana y la relacionada con el ámbito lingüístico catalán y vela por la conservación y difusión del patrimonio bibliográfico. La misma biblioteca interpreta la ley de forma que incluye los documentos digitales.

²² Esta práctica es conocida y comentada en los medios de comunicación, en artículos como "[Compro seguidores](#)" (*La Vanguardia*, 29 d'abril del 2012) o "[Perfiles con muchos huevos](#)" (*El País*, 20 d'abril del 2012).

²³ Dentro del estándar tecnológico Ecma-262, equivalente al estándar internacional ISO / IEC 16262:2011. Ecma International, encontramos "[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>>.