Revista Internacional de Ciencia y Tecnología de la Información Geográfica
International Review of Geographical Information Science and Technology

García Balboa, J. L., Ariza López, F. J., Ureña Cámara, M. A. y Ureña-López, A. (2012): "Análisis de métodos de etiquetado geográfico de documentos html en internet y su aplicación en España", GeoFocus (Artículos), nº 12, p. 122-146. ISSN: 1578-5157

ANÁLISIS DE MÉTODOS DE ETIQUETADO GEOGRÁFICO DE DOCUMENTOS HTML EN INTERNET Y SU APLICACIÓN EN ESPAÑA

JOSÉ LUIS GARCÍA BALBOA[1], FRANCISCO JAVIER ARIZA LÓPEZ[2], MANUEL ANTONIO UREÑA CÁMARA[3], ALFONSO UREÑA-LÓPEZ [4]
[1,2,3] Grupo de Investigación Ingeniería Cartográfica. Universidad de Jaén
[4] Grupo de Investigación Sistemas Inteligentes de Acceso a la Información. Universidad de Jaén
Campus Las Lagunillas s/n. 23071, Jaén, España.
[1] jlbalboa@ujaen.es; [2] fjariza@ujaen.es; [3] maurena@ujaen.es; [4] laurena@ujaen.es

RESUMEN

En los últimos años el etiquetado geográfico o geoetiquetado está cobrando una especial relevancia, sobre todo con la generalización del uso de dispositivos móviles con conexión inalámbrica a internet. En esta misma línea, está en pleno auge el desarrollo y uso de servicios basados en localización o LBS (*Location Based Services*). Por ello es importante que los documentos html vinculados a una determinada localización, sean geoetiquetados, y así facilitar la labor de los agentes automáticos de búsqueda. Existen varias alternativas para realizar este geoetiquetado. En este documento se presentan las diversas alternativas existentes, resumiendo sus características y realizando un análisis comparativo. Finalmente se analiza el grado de uso del geoetiquetado en páginas web en España.

Palabras clave: etiquetado geográfico, geoetiquetado, html, web

ANALYSIS OF METHODS OF GEOGRAPHIC TAGGING IN HTML DOCUMENTS IN INTERNET AND THEIR USE IN SPAIN

ABSTRACT

Geographic tagging, or geotagging, is becoming more important nowadays, especially because of the increased use of mobile devices with wireless internet access and the development and use of Location Based Services. Therefore, it is important that the html documents pointing to a location are geotagged, thus facilitating the task of automatic search agents. The available alternatives are presented in this paper and their main characteristics are summarized in a comparative analysis. Finally, the extent to which geotagging is used in web pages in Spain is also analyzed.

Keywords: geographic tagging, geotagging, html, web

Recibido: 18/7/2011
Aceptada versión definitiva: 31/5/2012

García Balboa, J. L., Ariza López, F. J., Ureña Cámara, M. A. y Ureña-López, A. (2012): "Análisis de métodos de etiquetado geográfico de documentos html en internet y su aplicación en España", GeoFocus (Artículos), nº 12, p. 122-146. ISSN: 1578-5157

1. Introducción

Scharl (2007) define el geoetiquetado, o etiquetado geográfico, como el proceso de asignar información contextual geoespacial, desde localizaciones puntuales específicas hasta regiones de geometría arbitraria. Por tanto, este proceso da como resultado una o varias etiquetas geográficas, o geoetiquetas, junto a alguna información publicada en cualquier formato (fotografías, vídeos, páginas web, etc.). Estas geoetiquetas son metadatos, cuyo objetivo es enriquecer la información al asociarle una localización espacial. Su empleo está en pleno auge, fundamentalmente para el caso de las fotografías, existiendo diferentes trabajos como los de Torniai *et al*. (2007) o Viana *et al*. (2007). La llegada de los teléfonos móviles inteligentes (o smartphones), con cámara de fotos digital, receptor GNSS y brújula integrados, con conexión a internet, ha impulsado aún más el enriquecimiento de datos para el caso de las fotografías, como ya se adelantaba en Viana *et al*. (2007). En cuanto a la información geoespacial contenida, la opción más sencilla opta por incluir las coordenadas geográficas de latitud y longitud del punto de disparo. No obstante también se pueden incluir otros campos de información sobre la altitud, la orientación, etc. Toda esta información puede ser capturada por aquel dispositivo que consta de los sensores anteriormente citados, por lo que el geoetiquetado puede hacerse de forma transparente al usuario, que sólo debe activarlo en el programa de captura de la imagen, y sin necesidad de ninguna formación específica.

Actualmente, el principal motivo de geoetiquetar la información es el de permitir el desarrollo de los servicios basados en localización o LBS (*Location Based Services*). Los LBS son aplicaciones que integran la localización geográfica con el concepto general de servicio y que pueden ser definidos como (Schiller y Voisard, 2004): servicios que integran la localización o posición de un dispositivo móvil con otra información para proporcionar un valor añadido al usuario. Muchos de los documentos html existentes actualmente en internet contienen información que está vinculada a una localización (una página web de un hotel, por ejemplo). Para facilitar que esta información pueda ser correctamente integrada en servicios LBS es conveniente que estos documentos estén geoetiquetados. En este sentido existen trabajos que pretenden geoetiquetar de forma automática grandes lotes de páginas web indexadas, como Amitay *et al*. (2004) o Pyalling *et al*. (2006). También se han realizado numerosos trabajos centrados en los procesos de búsqueda basados en la posición, como Jones *et al*. (2004), Graupmann y Schenkel (2006), Delboni *et al*. (2007) o Tsai (2011). En la mayoría de ocasiones los motores de búsqueda utilizan procesos de recuperación de información geográfica o GIR (*Geographic Information Retrieval*), a partir de la información no estructurada que contienen las páginas web. Pero también hay trabajos como los de Hu y Ge (2008), basados en la búsqueda de geoetiquetas. Siempre es preferible que el autor incluya las geoetiquetas, y por tanto explicite la información geográfica para los motores de búsqueda, ya que es quien mejor entiende el contexto geoespacial del contenido de su web (McCurley, 2001). En este sentido, el trabajo de Fink *et al*. (2009) obtiene una discrepancia inferior a 100 millas en el 61% de blogs analizados, entre la posición aportada por el autor y la extraída a partir del contenido del blog, con un valor medio de 50 millas.

El principal problema del geoetiquetado de documentos html reside en la ausencia de un estándar al uso que determine unívocamente un método o formato. Al contrario, el panorama es bastante difuso, al existir diferentes alternativas, propuestas por diferentes autoridades de la red

García Balboa, J. L., Ariza López, F. J., Ureña Cámara, M. A. y Ureña-López, A. (2012): "Análisis de métodos de etiquetado geográfico de documentos html en internet y su aplicación en España", GeoFocus (Artículos), nº 12, p. 122-146. ISSN: 1578-5157

(IETF, W3C), por comunidades menores (como Microformats), o bien propuestas no formales que comenzaron a utilizarse cuando internet comenzaba a desarrollarse, y que con el paso del tiempo se han terminado aceptando como una alternativa más.

Dada esta situación, tampoco existen aportaciones que aborden un análisis comparativo entre estos estilos. Por ello se plantea el presente trabajo, cuyo objetivo principal es realizar una revisión de los diferentes métodos de geoetiquetado de documentos html que se proponen en la actualidad, resumiendo sus principales características y explicitando sus ventajas e inconvenientes. Como complemento a lo anterior, también se incluye un análisis estadístico sobre el grado de uso del geoetiquetado en España, con el objetivo de conocer el grado de penetración de los métodos de geoetiquetado. Existen pocos estudios internacionales a este respecto, pudiendo citarse a Fink *et al*. (2009), que indica que sólo el 0,11% de los blogs analizados (800000) incluyen geoetiquetas en la cabecera del documento html.

El documento se organiza en cinco apartados. El primero de ellos se corresponde con esta introducción y el segundo se dedica a presentar y describir las características de cada uno de los métodos de geoetiquetado analizados, a los que se denominarán estilos de geoetiquetado. El siguiente apartado realiza un análisis comparativo de todos ellos. El cuarto apartado presenta un estudio sobre el grado de aplicación de las geoetiquetas en España mediante la exploración de un conjunto de sitios webs. El quinto y último de los apartados se dedica a las conclusiones.

2. Estilos de etiquetado geográfico en páginas web

2.1. Estilos en la cabecera del html

Se incluyen en este apartado los estilos que se basan en la inclusión de información en la cabecera del archivo html, mediante una o varias etiquetas.

2.1.1. Estilo de misil balístico intercontinental (ICBM)

El nombre del estilo de geoetiquetado ICBM procede de las siglas *InterContinental Ballistic Missile* (misil balístico intercontinental). Este estilo de geoetiquetado está ligado al concepto de dirección ICBM (*ICBM address*), que se utilizó en el antiguo proyecto *Usenet mapping project* (Raymond, 1996). Usenet es uno de los sistemas más antiguos de comunicación en red, y aún hoy está en uso, ya que es el sistema de discusión distribuida más conocido como grupos de noticias.

El estilo proporciona el par de coordenadas latitud, longitud, en formato decimal y en el sistema WGS84. La información se incluye en la cabecera del archivo html mediante un *metatag* o etiqueta html, cuyo nombre es ICBM y cuyo contenido es el par de coordenadas latitud, longitud, por este orden. A modo de ejemplo, las coordenadas de la Universidad de Jaén (España), concretamente las de su edificio de Rectorado, son: latitud igual a 37,78968ºN y longitud igual a 3,77952ºW; la etiqueta se introduciría según el siguiente formato:

García Balboa, J. L., Ariza López, F. J., Ureña Cámara, M. A. y Ureña-López, A. (2012): "Análisis de métodos de etiquetado geográfico de documentos html en internet y su aplicación en España", GeoFocus (Artículos), nº 12, p. 122-146. ISSN: 1578-5157

```
<meta name="ICBM" content="37.78968, -3.77952">
```

Hoy día éste es un método bastante habitual para geoetiquetar cualquier material en la red. Prueba de ello son dos proyectos que permiten poner en marcha servicios LBS a partir de las páginas web recogidas en sus bases de datos: GeoURL (http://geourl.org) y A2B (http://a2b.cc/). En ambos sitios web se recogen directorios que relacionan una gran cantidad de direcciones URL con localizaciones geográficas. Es decir, si los propietarios de las páginas web dan de alta su página web en A2B, los usuarios pueden realizar búsquedas basadas en la posición, de forma que accedan a las páginas web de lugares o establecimientos próximos a su posición actual.

2.1.2. Estilo GeoTag (GT)

El estilo de geoetiquetado *GeoTag* (en adelante GT) ha sido propuesto por la IETF (*Internet Engineering Task Force*), que es una comunidad abierta de la ASOC (*Internet Society*), y que proporciona documentación técnica para el desarrollo de internet. El estilo GT es actualmente un borrador o *internet-draft*, cuya última versión es la 8, documentada en Daviel y Kaegi (2007), que en caso de ser aprobado pasaría a ser un RFC (Request For Comments), que constituiría una propuesta formal de estándar. Aunque aún se trate de un borrador, este estilo se ha venido utilizando frecuentemente. La primera versión del borrador es de 1999, lo que da idea del tiempo transcurrido desde que se planteó su uso por primera vez.

Como se indica en Daviel y Kaegi (2007), este geoetiquetado pretende ser una forma concisa, no ambigua, simple de usar y compatible con herramientas de edición existentes, para proporcionar información sobre localización a los robots web que revisitan las páginas cada varias semanas. Los tipos de etiquetas son los siguientes[1]:

- Posición (*geo.position*). Se utiliza para proporcionar la latitud y longitud de una localización en el sistema WGS84, y opcionalmente la elevación, por este orden.
- País (*geo.country*). Se corresponde con un identificador del país, según la norma ISO 3166-1 (ISO, 2006a). En el caso de España el identificador es "ES".
- Dirección (*geo.a1*, *geo.a2*, etc.). Para la etiqueta geo.a1 se propone utilizar una lista controlada de subregiones, como la proporcionada por la norma 3166-2 (ISO, 2007). Por ejemplo, la etiqueta geo.a1 para el caso de la provincia de Jaén sería "J". El resto de etiquetas serían libres, pero su uso sería más descriptivo que de indexación, por la posible ambigüedad en su utilización, dadas las diferentes convenciones a la hora de nombrar un lugar. En resumen, puede considerarse como una estructura jerárquica para ir presentando la dirección postal de una ubicación.
- Hito (geo.lmk). Se trata de una etiqueta que se propone para incluir el nombre común del objeto que se está describiendo.

En cuanto al uso de estas etiquetas, en principio, como mínimo hay que utilizar la etiqueta de posición (*geo.position* con la latitud y longitud). No obstante, esto tendría poco sentido si se está

García Balboa, J. L., Ariza López, F. J., Ureña Cámara, M. A. y Ureña-López, A. (2012): "Análisis de métodos de etiquetado geográfico de documentos html en internet y su aplicación en España", GeoFocus (Artículos), nº 12, p. 122-146. ISSN: 1578-5157

haciendo referencia a un país o región. En ese caso se puede omitir la etiqueta de posición, y en su lugar se incluirían las etiquetas de país (*geo.country*) y dirección (*geo.a1*, etc.)

Siguiendo con el ejemplo de la Universidad de Jaén, las etiquetas a incluir serían las siguientes:

```
<meta name="geo.position" content="37.78968; -3.77952">
<meta name="geo.country" content="ES">
<meta name="geo.a1" content="J">
<meta name="geo.a2" content="Jaén">
<meta name="geo.lmk" content="Universidad de Jaén">
```

2.2. Estilo de microformato Geo (MFG)

Los microformatos son un conjunto de formatos de datos simples y abiertos, elaborados a partir de estándares existentes y ampliamente adoptados (Microformats, 2011a). El término microformato procede de la idea de reutilizar lo que ya está en uso, de forma que mediante pequeñas adaptaciones, se puedan resolver problemas específicos. El origen de los microformatos no está vinculado a ninguna organización de normalización, sino a una comunidad de usuarios que van desarrollando propuestas sencillas de codificación, colaborando a través de diferentes canales, como una lista de correo, un chat y una wiki (Microformats, 2011a), donde se encuentra toda la información actualizada sobre los microformatos (aprobados y en borrador). Como fecha de origen de los microformatos suele establecerse la del 25 de junio de 2005, en la que se lanzó el sitio microformats.org.

Uno de los microformatos, que actualmente se clasifica como borrador, aunque está en uso, es el Geo (Microformats, 2011b) (en adelante se denominará MFG). Este microformato permite etiquetar la latitud y longitud en grados decimales y en el sistema WGS84 (actualmente está bajo estudio la inclusión de una extensión del microformato para incluir información sobre elevación). A diferencia de los estilos ICBM y GT, en este caso se pueden incluir en un archivo html tantas localizaciones como se desee, ya que el microformato no se incluye en la cabecera, sino que forma parte del cuerpo del documento.

En línea con la idea de utilizar estándares existentes, la información que contiene un microformato se añade a una página web mediante los mecanismos genéricos que existen en las especificaciones de html (WC3, 1999) para añadir estructuras a los documentos. Por ejemplo, se puede crear una estructura utilizando los habituales elementos de html *div* y *span*, junto con el atributo *class*. Siguiendo con el ejemplo de la Universidad de Jaén, la información a incluir sería la siguiente:

```
<div class="geo">
     <span class="latitude">37.78968</span>;
     <span class="longitude">-3.77952</span>
</div>
```

García Balboa, J. L., Ariza López, F. J., Ureña Cámara, M. A. y Ureña-López, A. (2012): "Análisis de métodos de etiquetado geográfico de documentos html en internet y su aplicación en España", GeoFocus (Artículos), nº 12, p. 122-146. ISSN: 1578-5157

En ejemplo anterior, la información que lee el usuario y la que lee cualquier agente automático de búsqueda es la misma. Si se desea presentar la información de latitud y longitud en otro formato, para que sea más entendible por el ser humano (por ejemplo, como grados, minutos y segundos), existe un patrón de diseño que permite hacerlo. En concreto, a este patrón se le denomina *abbr-design-pattern* (http://microformats.org/wiki/abbr-design-pattern), y permite utilizar el elemento *abbr* en lugar de *span*, y añadir tras el atributo *class* el atributo *title*, que es el que contiene la información estandarizada y compatible con el microformato. Así, siguiendo con el ejemplo que se está manejando se podría incluir la siguiente información:

```
<div class="geo">
   <abbr class="latitude" title="37.78968">N 37° 47′ 22.8′′</abbr>
   <abbr class="longitude" title="-3.77952">W 3° 46′ 46.3′′</abbr>
</div>
```

Es conveniente aclarar que el MFG nace a partir del microformato *hCard* (Microformats, 2011c), el cual engloba al primero. Por ello, si se desea añadir el nombre de la localización, o la dirección postal, es posible recurrir al microformato *hCard*, más rico, que pone a disposición del usuario una mayor variedad de atributos de tipo *class* (entre ellos los de atributos geo, relativos al geoetiquetado). La Universidad de Jaén podría incluirse de la siguiente forma si se quiere incluir el nombre y la dirección postal de la misma:

```
<div class="hcard">
   <span class=”fn org”>Universidad de Jaén</span>
   <span class=”adr”>
      <span class="street-address">Paraje Las Lagunillas</span>
      <span class="locality">Jaén</span>,
      <span class="region">J</span>
      <span class="postal-code">23071</span>
      <span class="country-name">Spain</div>
   </span>
   <span class="geo">
      <abbr class="latitude" title="37.78968">N 37° 47′ 22.8′′</abbr>
      <abbr class="longitude" title="-3.77952">W 3° 46′ 46.3′′</abbr>
   </span>
</div>
```

2.3. Estilo RDF Geo (RDFaG)

El RDF (*Resource Description Framework*, Marco de Descripción de Recursos) es un modelo de datos que fue creado por el W3C (*World Wide Web Consortium*), dentro del concepto de Web Semántica. La idea de la Web Semántica fue popularizada por los mismos creadores del formato html y de la WWW, en Berners-Lee, Hendler y Lassila (2001), donde es descrita de forma sencilla. Su objetivo es añadir información adicional a las páginas web, en forma de metadatos y de una forma estructurada, para que pueda ser procesada de forma automática por cualquier agente automático de búsqueda de información en la red.

García Balboa, J. L., Ariza López, F. J., Ureña Cámara, M. A. y Ureña-López, A. (2012): "Análisis de métodos de etiquetado geográfico de documentos html en internet y su aplicación en España", GeoFocus (Artículos), nº 12, p. 122-146. ISSN: 1578-5157

El *RDF Interest Group*, que a partir de 2004 pasó a denominarse *Semantic Web Interest Group* (SWIG), seleccionó en primer lugar varias ontologías[2], cuyo uso podría generalizarse con mayor facilidad, es decir, podrían ser utilizadas por una mayor cantidad de aplicaciones. Entre estas ontologías el SWIG seleccionó la referente a la información geoespacial, que queda representada mediante un vocabulario básico, denominado *SWIG Basic Geo Vocabulary* (BGV) (W3C, 2006). El vocabulario BGV permite describir puntos con las coordenadas de latitud, longitud y altitud. No obstante hay que aclarar que aún no forma parte de las estandarizaciones elaboradas por el W3C.

Para introducir los metadatos ajustados al modelo RDF en una página web se puede utilizar RDFa, que es una sintaxis para incorporar la información RDF en un documento html a través de atributos (véase W3C, 2008). En el caso del vocabulario BGV, los atributos de RDFa que se han de utilizar para la información geográfica son los siguientes:

- *typeof*. Para indicar de qué tipo es la información descrita. Aquí se indicaría que se trata de una información de tipo *geo:Point*, que quiere decir que se va a describir la localización de un punto.
- *property*. Se refiere a la propiedad (o atributo) que se va a describir. Serían tres las propiedades que se pueden incluir: *geo:lat* para la latitud, *geo:long* para la longitud y *geo:alt* para la altitud (ésta es opcional).
- *content*. Se incluye a continuación de la propiedad y contiene el valor del atributo, que sería el valor de latitud, longitud o altitud, en cada caso. Las dos primeras se introducen en grados decimales sobre WGS84 y la tercera en metros sobre el elipsoide de referencia local (así se indica en la documentación).

En adelante, se denominará RDFaG a la utilización de RDFa para representar información según el vocabulario BGV. Siguiendo con el ejemplo de la Universidad de Jaén, la información a incluir sería la siguiente:

```
<div typeof="geo:Point">
   <span property="geo:lat" content="37.78968"> N 37° 47′ 22.8′′</span>
   <span property="geo:long" content="-3.77952"> W 3° 46′ 46.3′′</span>
</div>
```

Como se puede comprobar, la información que lee el usuario y la contenida en los metadatos están en formatos diferentes. La primera estaría en formato sexagesimal y la segunda en formato decimal, que es el que se ajusta al vocabulario BGV y es capaz de ser entendida por cualquier agente automático de búsqueda de información. Es lo mismo que se ha indicado en el estilo MFG en cuanto al patrón *abbr-design-pattern*.

Con posterioridad al desarrollo de vocabulario BGV se crearon en el W3C los denominados *Incubator Groups* (XGs) con la idea de acelerar el desarrollo de nuevas propuestas relacionadas con la web, ya que el desarrollo de estándares suele ser lento. Uno de estos XGs fue el *W3C Geospatial Incubator Group* (GeoXG, ya desaparecido), siendo uno de sus primeros informes la elaboración del *W3C Geospatial Vocabulary* (W3C, 2007), que puede considerarse como una propuesta de actualización del vocabulario BGV. El GeoXG planteó inicialmente varios ejemplos prácticos para

García Balboa, J. L., Ariza López, F. J., Ureña Cámara, M. A. y Ureña-López, A. (2012): "Análisis de métodos de etiquetado geográfico de documentos html en internet y su aplicación en España", GeoFocus (Artículos), nº 12, p. 122-146. ISSN: 1578-5157

evidenciar la necesidad de fortalecer la web en la componente geoespacial y decidió adoptar el modelo de datos GeoRSS (http://www.georss.org/gml) de geosindicación de contenidos. Este modelo permite describir no sólo puntos, sino también rectángulos, líneas y polígonos. Se trata de un modelo ligeramente diferente y reducido respecto al modelo de entidades simples de OGC (http://www.opengeospatial.org/standards/sfs). No obstante, La falta de consenso provocó el cierre de este grupo de trabajo. También mencionar, relacionado con lo anterior, que en la wiki de W3C (W3C, 2011) existe una entrada dedicada al denominado GeoRDF, en la que se propone una estructura similar a la del GeoXG, pero esta propuesta por ahora no va más allá de su inclusión en esta wiki.

Otra propuesta a reseñar es la denominada GeoJSON (*Geo JavaScript Objetc Notation*). El JSON es un formato de texto que permite intercambiar información estructurada, derivado del lenguaje de javascript y que fue propuesto en Crockford (2006), aunque aún es sólo de carácter informativo, y por tanto no es una propuesta de estándar. En este marco, GeoJSON es una propuesta para intercambiar información geográfica estructurada, y también acepta estructuras más complejas que el punto. Sus especificaciones se encuentran en GeoJSON (2008). Una información textual en este formato podría convertirse al formato de alguno de los vocabularios anteriormente expuestos.

3. Análisis comparativo de estilos

En el apartado 2 se han resumido cuáles son los principales estilos para incluir información geográfica en una página web. Hasta la fecha, no se ha declarado ninguno de ellos como estándar que haga decantarse por uno u otro. Incluso en muchas ocasiones se opta por incluir la información geográfica en varios estilos, de forma que se maximice la posibilidad de ser localizada por los agentes automáticos de búsqueda. No obstante es conveniente realizar un análisis comparativo que permita vislumbrar las ventajas, inconvenientes, semejanzas y diferencias que hay entre ellos, queriendo destacar la ausencia de trabajos publicados con anterioridad en esta línea. La información resultante del análisis podrá servir de orientación al gestor de una web a la hora de elegir los estilos a utilizar. También sería un punto de partida para futuros trabajos sobre propuestas generales de uso de los diferentes estilos, según las características de la web (temática, propósito, estructura, etc.). Incluso podría ser útil para futuras revisiones de los estilos. El análisis prestará especial atención a algunos aspectos relevantes en el ámbito de la cartografía y SIG, como es el tratamiento dado a los sistemas de referencia y a la incertidumbre del dato.

Para facilitar el repaso a las principales características de los estilos de geoetiquetado descritos en el apartado 2, en la tabla 1 se presenta un resumen. Estas características se irán analizando en los siguientes apartados.

3.1. Existencia de documentación sobre el estilo y de herramientas

Los estilos GT, MFG y RDFaG sí están establecidos formalmente, respectivamente por IETF (aunque actualmente se trata de un borrador o *internet-draft*), *microformats community* y por W3C. Esto conlleva la existencia de páginas web en estas entidades, dedicadas a la descripción de

García Balboa, J. L., Ariza López, F. J., Ureña Cámara, M. A. y Ureña-López, A. (2012): "Análisis de métodos de etiquetado geográfico de documentos html en internet y su aplicación en España", GeoFocus (Artículos), nº 12, p. 122-146. ISSN: 1578-5157

estos estilos, indicando las sucesivas revisiones realizadas, y señalando si existe alguna discusión abierta, problemas pendientes de resolver, etc. Por el contario, el geoetiquetado ICBM ha venido utilizándose conforme se difundía su uso por la red, pero sin que ninguna organización lo describiera formalmente.

Sería cómodo para los usuarios que los propios editores html facilitarán la inclusión del geoetiquetado, respetando lo especificado en los documentos que definen los distintos estilos. No obstante, dado que las geoetiquetas son estructuras de código bastante simples, hasta ahora se introducen editando el código html directamente, con las herramientas genéricas habituales de los distintos editores de código.

En la red existen algunas herramientas sencillas (consulta 1/3/2012) que facilitan la incorporación del fragmento de código html relativo al geoetiquetado, principalmente en relación a los estilos ICBM y GT. Este es el caso de *HTML Geo-Tag Generator*[3], de *MyGeoposition.com*[4] o de *Geo-Tag your Web Pages!*[5], que facilitan el código en estos dos estilos a partir de la dirección postal introducida. La herramienta *Address Fix*[6] añade también el código para estilo MFG y en GeoRSS. También existen herramientas de validación como es el caso de *HTML Geo-Tag Validator*[7], que inspecciona la dirección URL introducida buscando las etiquetas según los estilos ICBM y GT, y realizando alguna inspección sencilla. No obstante es necesario indicar que las herramientas reseñadas siguen versiones antiguas del estilo GT, por lo que necesitarían ser actualizadas.

3.2. Información contenida en el estilo

La mayoría de los estilos contemplan únicamente el etiquetado asociado a una geometría de carácter puntual. Por tanto, cualquier fenómeno del mundo real, independientemente de su tamaño o extensión, ha de ser georreferenciado mediante un punto mediante sus coordenadas geográficas (latitud y longitud). Solamente las opciones basadas en el vocabulario del GeoXG, que podría usarse en el RDFaG, admitirían geometrías de carácter lineal o areal.

Aparte de lo anterior, sólo el estilo GT permite el posicionamiento denominado "indirecto" o por identificadores geográficos, es decir, basado en una relación con una localización dada por uno o varios fenómenos geográficos. De este modo, permite incluir información sobre el país, región, dirección postal y nombre común del objeto. Incluso la información sobre el país y región está estandarizada. Esto puede facilitar la recuperación de información no basada exclusivamente en la coordenada o realizar búsquedas basadas en la dirección postal. No obstante, este posicionamiento indirecto no sigue formalmente ningún sistema de referencia espacial por identificadores geográficos. En este sentido, sería muy positivo fijar un sistema de referencia global en conformidad con la norma ISO 19112 (ISO, 2003).

Merece un comentario aparte la información sobre la coordenada de altitud, ya que dos estilos, ICBM y MFG, no la consideran, aunque sobre éste último se está considerando introducir una extensión del microformato para poder incluir esta información.

García Balboa, J. L., Ariza López, F. J., Ureña Cámara, M. A. y Ureña-López, A. (2012): "Análisis de métodos de etiquetado geográfico de documentos html en internet y su aplicación en España", GeoFocus (Artículos), nº 12, p. 122-146. ISSN: 1578-5157

3.3. Ubicación de las etiquetas en el archivo html

Dos estilos consideran la inclusión de la información sobre el geoetiquetado en la cabecera del archivo html: el ICBM y GT. Por tanto son útiles para páginas web cuyo contenido aluda a una única localización. Es decir, si un documento contiene información sobre varias ubicaciones, sería más adecuado dividirlo en varios html, de forma que se incluyera la información sobre la posición en cada una de las cabeceras de los nuevos documentos (por ejemplo, una página por cada restaurante que se quiera georreferenciar). Al ubicarse en las cabeceras, esta información no es directamente accesible por el usuario que lee la página, aunque puede ser encontrada sin problema por agentes de búsqueda. Si se quisiera que esta información sobre la localización geográfica fuera explícita para el usuario, debería también incluirse como texto simple en el cuerpo del documento html.

Dos estilos consideran incluir la información sobre el geoetiquetado en el propio cuerpo del documento, el MFG y el RDFaG. Es decir, permiten ir incluyendo atributos al texto conforme se va redactando el documento html, de forma que la información sobre la localización que se pone de forma explícita a disposición del usuario (por ejemplo, para que pueda introducirla en su navegador GNSS) queda enriquecida y puede ser interpretada por un agente automático. De este modo en una misma página web se pueden incluir geoetiquetas sobre diferentes ubicaciones (por ejemplo, todos los restaurantes de una ciudad en el mismo documento html). Aunque la redacción explícita de la información geográfica para el usuario sigue en principio un formato libre, es interesante tener en cuenta el Anexo D, de carácter no obligatorio sino informativo, de la norma ISO 6709 (ISO, 2008), sobre la representación de la longitud y la longitud en la interfaz de usuario, en la que se dan una serie de sugerencias: la latitud precede a la longitud, uso de grados sexagesimales, uso de símbolos de grados, minutos y segundos, etc.

3.4. Formato de las coordenadas geográficas y sistema de referencia

Todos los estilos requieren introducir la información sobre la latitud y la longitud en un mismo formato, que es el de grados decimales. No obstante, el usuario puede estar habituado a otros formatos, como el de grados en formato sexagesimal o coordenadas UTM. En el caso de los estilos ICBM y MFG no se trata de un problema, ya que la información reside en la cabecera del archivo html, la cual no es directamente accesible por el usuario. En el caso de los estilos MFG y RDFaG, dado que trabajan mediante atributos de la información vertida en el cuerpo del documento, supondría una limitación a la hora de presentar las coordenadas geográficas al usuario. En ambos casos esto se ha solventado de forma que se pueda presentar al usuario una información diferente a la contenida en los atributos.

En cuanto al sistema de referencia, se propone siempre la utilización de WGS84 en lugar de otros sistemas de referencia más antiguos y locales (como ED50 en España o NAD27 en EE.UU.). Junto con la utilización del formato de coordenadas en grados decimales, facilita la estandarización y el intercambio de información.

García Balboa, J. L., Ariza López, F. J., Ureña Cámara, M. A. y Ureña-López, A. (2012): "Análisis de métodos de etiquetado geográfico de documentos html en internet y su aplicación en España", GeoFocus (Artículos), nº 12, p. 122-146. ISSN: 1578-5157

No obstante es importante que el usuario sepa qué supone tomar información georreferenciada en antiguos sistemas locales. Es habitual que un usuario sin formación específica considere que la coordenada de latitud y de longitud es única, desconociendo que existen diferentes sistemas de referencia. Por ello es conveniente que siempre que se documente algún tipo de estilo de geoetiquetado (unas especificaciones, un manual, una guía de uso, etc.) se remarque la importancia de este aspecto.

Respecto al sistema de referencia empleado en altitud, para el estilo GT se solicita que se utilice el sistema WGS84. Aunque no es común en cartografía el utilizar altitudes elipsoidales, la difusión de la utilización de navegadores GNSS, y la necesidad de utilizar un sistema de referencia único, favorece su uso. No obstante, es importante informar al usuario sobre qué supone el uso de la altitud sobre WGS84. En este sentido, Daviel y Kaegi (2007) advierten de la diferencia que puede provocar respecto a un sistema de referencia altimétrico local (por ejemplo, referido al nivel medio del mar y utilizando altitudes ortométricas). En relación al RDFaG, este asunto no está tan claro; la documentación indica que la altitud se proporciona "sobre el elipsoide de referencia local". Esto contradice la idea de utilizar el sistema WGS84, que utiliza el elipsoide global GRS80. Podría pensarse que en realidad quería referirse al sistema de referencia altimétrico local (por ejemplo, el sistema de altitudes ortométricas sobre el geoide, utilizado en España). Esto último provocaría problemas a la hora de que los usuarios utilicen navegadores GNSS que suelen ofrecer la altitud elipsoidal sobre GRS80, sin aplicar ningún modelo de geoide para obtener altitudes ortométricas.

En cualquier caso, la concreción del sistema de referencia considerado sólo es importante cuando las coordenadas proporcionadas tienen una incertidumbre de carácter métrico (como la que puede ofrecer un navegador GNSS) y se refieren a objetos del mundo real de poca extensión (un pequeño edificio por ejemplo). A modo de ejemplo, tiene poca o ninguna importancia conocer el sistema de referencia si se van a proporcionar las coordenadas de una ciudad; igual sucedería si se quieren dar las coordenadas de un edificio de poca extensión pero con una incertidumbre de varias centenas de metros. Esto es así porque el cambio en el valor de las coordenadas al cambiar de sistema de referencia es inferior a la incertidumbre de las coordenadas y/o a la extensión del objeto del mundo real que se está geoetiquetando. En este sentido, en Daviel y Kaegi (2007), se indica que siempre que la incertidumbre sea inferior a 1 km, las coordenadas deben convertirse al sistema WGS84. Esto es consistente con la norma ISO 6709 (ISO, 2008), que indica que hay que especificar el CRS (*Coordinate Reference System*, Sistema de Referencia de Coordenadas) para aplicaciones que requieren una exactitud mayor a 1 km.

Es interesante analizar la compatibilidad de los estilos de geoetiquetado con la Norma ISO 6709 (ISO, 2008), cuyo objetivo es la normalización de la representación de localizaciones geográficas puntuales mediante coordenadas. Esta norma establece que una localización siempre ha de constar de una tupla de coordenadas y una identificación del CRS. Será el CRS el que especifique las direcciones positivas de cada eje de coordenadas, el orden de las coordenadas y las unidades. En caso de no indicar ningún CRS, se considera que la posición del punto adquiere un mayor grado de incertidumbre y que la tupla consta de la latitud, antecediendo a la longitud, expresándolas en grados decimales, e indicando a continuación la altitud (o profundidad) si se trata de una localización tridimensional. El respetar este orden se considera crítico para el caso de emergencias en marina y navegación, ya que tradicionalmente éste ha sido el orden de estas

García Balboa, J. L., Ariza López, F. J., Ureña Cámara, M. A. y Ureña-López, A. (2012): "Análisis de métodos de etiquetado geográfico de documentos html en internet y su aplicación en España", GeoFocus (Artículos), nº 12, p. 122-146. ISSN: 1578-5157

coordenadas; utilizar otro orden podría generar situaciones de riesgo. Esto es importante para los dos estilos, ICBM y GT, basados en la inclusión de una tupla bajo una misma etiqueta. En ambos casos se siguen el orden que indica la Norma ISO 6709. Solamente la versión 0 del estilo GT (Daviel, 1999) contemplaba un orden distinto, pero fue modificado en pocos meses en la versión 1.

En cuanto al CRS, ya se ha comentado que en todos los estilos se ha convenido utilizar WGS84, aunque no se explicita en ninguna etiqueta. Esto es consistente con el RFC 5870 (Mayrhofer y Spanring, 2010), en relación a la estructura del URI[8] para localizaciones geográficas, denominado 'geo' URI. Este URI permite incluir el parámetro opcional *crs*, pero si no se incluye, por defecto se ha de entender que el CRS es el WGS84. Si en el futuro se viera conveniente abrir los estilos de geoetiquetado a otros CRS, sería conveniente incluir una etiqueta a tal efecto, de forma similar al parámetro *crs* del RFC 5870, antes mencionado. En este sentido, y en la línea de lo adoptado en las Infraestructuras de Datos Espaciales, parece lógico que se utilizara el sistema de códigos EPSG (*European Petroleum Survey Group*) que barre todas las opciones posibles a nivel mundial, tanto en datums, elipsoides como sistemas de proyección. Por ejemplo, el sistema WGS84 se codificaría como EPSG 4326 (sólo componente horizontal: latitud y longitud) o 4979 (componentes horizontal y vertical: latitud, longitud y altitud elipsoidal).

3.5. Incertidumbre de las coordenadas geográficas

Todo conjunto de coordenadas geográficas (latitud, longitud, altitud) que represente la posición de un punto relativo a la ubicación de un fenómeno del mundo real, lleva asociado un error, que viene a ser la diferencia entre estas coordenadas y las coordenadas verdaderas o de referencia. Este error es teóricamente desconocido, al no poder conocer las coordenadas verdaderas.

Dado lo anterior, se recurre a la denominada incertidumbre de medida, que sí puede ser evaluada. La incertidumbre de medida caracteriza la dispersión de los valores, y viene dada por un parámetro, habitualmente la desviación típica (pasando a denominarse incertidumbre típica de medida). Existen otros términos relacionados, como exactitud de medida o precisión de medida, que presentan matices diferentes y por tanto no deben confundirse entre sí (para profundizar, véase Ruiz y otros, 2010).

En el caso del geoetiquetado, ningún estilo considera necesario el incluir información sobre la incertidumbre de las coordenadas proporcionadas. Por tanto el usuario no tiene a su disposición ningún parámetro, cuantitativo o cualitativo, sobre la calidad de estas coordenadas y en consecuencia las limitaciones de uso que conlleva. En realidad esto sólo supone un problema cuando se desea proporcionar la localización de un objeto del mundo real de poco tamaño o extensión (un pequeño edificio por ejemplo). Esta situación es cada vez más frecuente, dada la difusión del uso de navegadores GNSS integrados en teléfonos inteligentes, y su uso para localizar objetos de interés. A modo de ejemplo, para que los usuarios localicen un restaurante, se necesita proporcionar las coordenadas geográficas con una incertidumbre reducida, del orden de varios metros. En este caso sería conveniente contar con el parámetro de incertidumbre antes mencionado, para verificar que es adecuado para el objetivo de localizar el objeto de interés. También conlleva,

García Balboa, J. L., Ariza López, F. J., Ureña Cámara, M. A. y Ureña-López, A. (2012): "Análisis de métodos de etiquetado geográfico de documentos html en internet y su aplicación en España", GeoFocus (Artículos), nº 12, p. 122-146. ISSN: 1578-5157

como se ha dicho anteriormente, que no exista duda acerca del sistema de referencia en el que se ofrecen las coordenadas (en principio WGS84).

Parece interesante la propuesta de Daviel y Kaegi (2007) en relación al estilo GT, donde se sugiere que si se está geoetiquetando un país o una parte de él (como un estado o una provincia), no se indiquen las coordenadas geográficas, presciendiendo de la etiqueta *geo.position*, y sólo se cumplimenten las etiquetas *geo.country* y *geo.a1*.

Por otro lado, también en Daviel y Kaegi (2007) se sugiere cuidar la ubicación del punto que va a representar al objeto del mundo real, cuando se está tratando con pequeños objetos y una incertidumbre reducida (del orden de pocos metros), como es el caso del restaurante antes mencionado. Puede ser más interesante capturar las coordenadas de la puerta de entrada que el centro geométrico del objeto. Si se captura el centro en lugar de la puerta, el usuario no tendría la certeza de por dónde se entra al edificio, a pesar de contar con unas coordenadas de incertidumbre reducida. Esto último conlleva, en cierto sentido, un aumento de la incertidumbre de la información que se quiere proporcionar al usuario, que ante todo debe ser útil. Evidentemente, toda regla del geoetiquetado, como la anterior respecto a la entrada a edificios, debería quedar reflejada en unas especificaciones del estilo de geoetiquetado, que actualmente se centran en el formato de las geoetiquetas, pero comentan poco o nada sobre su uso.

En relación a la referida ausencia de información expresa sobre la incertidumbre, es interesante observar la propuesta de Daviel, Kaegi y Kofahl (2007), relativa a la inclusión de una extensión de tipo geográfico para la transacciones http (es decir, las peticiones que se realizan desde un navegador web a un servidor para que le devuelva el contenido de una página web). En esta propuesta se sugiere un identificador denominado *geo-position*, que incluye la posibilidad de informar sobre la incertidumbre de la posición mediante la clave *epu* (*estimated position uncertainty*). Se propone que esta incertidumbre se corresponda con el radio de un círculo (o esfera) con una probabilidad del 95%. Aunque no se indica expresamente, esto se corresponde con la medida CE95 definida en la norma ISO 19138 (ISO, 2006b). También hay que citar en este sentido el RFC 5870 (Mayrhofer y Spanring, 2010), que permite incorporar un parámetro *u* opcional sobre la incertidumbre. Se trata de un parámetro único, ya se trate de una localización bidimensional o tridimensional. Podría ser interesante incorporar algo similar en los estilos de geoetiquetado.

Existe la posibilidad de relacionar el número de cifras decimales de las coordenadas proporcionadas con su incertidumbre, tal y como se sugiere en Daviel y Kaegi (2007) y en ISO (2008). Una menor incertidumbre, se correspondería con un mayor número de cifras decimales. A modo de ejemplo, si se han capturado con un navegador GNSS las coordenadas de la entrada de un edificio, éstas deberían darse con 4 decimales (asumiendo, que 1 grado son unos 111 km, 1 diezmilésima de grado supone unos 10 metros) y en WGS84. En cambio si sólo se sabe que el edificio está en una gran ciudad, se podrían tomar las coordenadas del centro de la ciudad a partir de cualquier fuente cartográfica y ofrecer estás coordenadas para el edificio, pero sólo con dos decimales (la centésima de grado equivaldría a 1 km aproximadamente) y por supuesto sin necesidad de que sean WGS84.

García Balboa, J. L., Ariza López, F. J., Ureña Cámara, M. A. y Ureña-López, A. (2012): "Análisis de métodos de etiquetado geográfico de documentos html en internet y su aplicación en España", GeoFocus (Artículos), nº 12, p. 122-146. ISSN: 1578-5157

Obsérvese, por otro lado, que el RFC 5870 indica que el número de cifras decimales no debe ser asociado a la incertidumbre, recomendando el anteriormente referido parámetro *u* para informar sobre esta última. Esto saca a la luz que el tratamiento de la incertidumbre es un asunto pendiente en la búsqueda de la interoperabilidad de la información geográfica en internet. La introducción de un parámetro específico para la incertidumbre sería lo más oportuno.

4. Análisis de la situación actual del uso de geoetiquetas en España

Como se ha ido indicando son diversas las opciones de geoetiquetado, sin embargo no existe un estudio ni análisis del grado de aplicación de estas geoetiquetas, ni a nivel mundial, ni a nivel nacional, con la excepción del referido trabajo de Fink *et al*. (2009) sobre blogs. El único dato que se ha podido encontrar procede de *Whois DataBase* (http://reviews.gcoupon.com), donde se indica que se ha explorado el millón de sitios web más relevantes del mundo y se presenta el resultado del conteo de las etiquetas ICBM en dichos sitios. El dato que ofrece este sitio es la aparición de 3267 casos en ese millón de sitios analizados, es decir, sólo un 0,0033% (un tres por mil). Desgraciadamente, no se sabe la fecha del trabajo, los criterios de selección, ni se dispone de ninguna información adicional sobre estos datos.

Un objetivo complementario del presente trabajo es realizar una primera aproximación a la situación que se da en España en cuanto al uso del geoetiquetado. Para ello se ha desarrollado el siguiente proceso, que se detalla en cada uno de los subapartados posteriores:

- Selección de una muestra de sitios web.
- Descarga automatizada de los sitios por medio de una araña o robot.
- Búsqueda de estilos de geoetiquetado en las páginas descargadas.
- Análisis de los resultados.

4.1. Selección de la muestra

El aspecto que más condiciona el estudio es la selección de la muestra. En este caso no se puede realizar un muestreo estadístico y de tipo aleatorio, por la imposibilidad de conocer el marco o población de manera previa. Para solventar esta situación se han desarrollado un muestreo dirigido en dos líneas de actuación:

¬ Selección según el liderazgo, entendido como el ranking de visitas. Para este caso se han usado dos fuentes de información independientes, que proporcionan listados con los sitios más visitados por país:

- Alexa: Empresa del grupo Amazon conocida internacionalmente por publicar listados de sitios más visitados tanto a nivel mundial como por país y por categoría. Estos listados se basan en usuarios que utilizan Internet Explorer, Firefox o Chrome y tienen instalado el *plug in Alexa Toolbar*. En este trabajo se ha utilizado el listado correspondiente a España

García Balboa, J. L., Ariza López, F. J., Ureña Cámara, M. A. y Ureña-López, A. (2012): "Análisis de métodos de etiquetado geográfico de documentos html en internet y su aplicación en España", GeoFocus (Artículos), nº 12, p. 122-146. ISSN: 1578-5157

para la fecha de julio de 2011, que proporciona un listado con 500 sitios. La dirección de acceso es: http://www.alexa.com/topsites/countries/ES.

- Google: Google dispone de herramientas (*Ad Planner*, *Google Analytics*, *Google ToolBar*) que le permiten conocer con gran detalle los flujos de la red. Desde mayo de 2010 libera una lista con los mil sitios más visitados a nivel mundial y con los cien sitios más visitados para una selección de países. Se ha utilizado el listado correspondiente a España para la fecha de enero de 2011 (último disponible al comienzo del estudio), que proporciona un listado con 100 sitios. La dirección de acceso es: http://www.google.com/adplanner.

¬ Selección guiada, basada en los resultados de búsquedas de Google a partir de un conjunto de palabras clave. Con esta selección guiada se pretende incluir un conjunto de sitios representativos de diversas realidades sectoriales (p.e. educación superior, banca, administración) que se consideran relevantes por las iniciativas que toman en cuanto al desarrollo de sus sitios web. Las palabras clave utilizadas han sido las siguientes: ayuntamiento, diputación, ministerio, universidad, banco, camping, cadena de hotel, agencia de viajes, transportes, empresa, puerto, aeropuerto, seguros, instituto cartográfico, colegio. Es importante señalar que las soluciones que proporciona Google no pueden ser utilizadas directamente por la gran cantidad de ruido que pueden incluir. Por este motivo, sobre dichas soluciones se realizó un proceso de selección manual de los sitios realmente relacionados con la búsqueda. Finalmente se concretó un listado con 600 sitios.

Según lo indicado, se disponía de un total de 1200 direcciones de sitios web procedentes de fuentes distintas, lo que puede ocasionar la presencia de direcciones repetidas. Por este motivo se hizo necesaria la eliminación de reiteraciones y sitios externos, llegándose a disponer de una lista depurada con 1027 sitios, conformados por 72 sitios procedentes del ranking de Google, 442 sitios procedentes del ranking de Alexa y 513 procedentes de la selección guiada.

4.2. Captura de la información

El proceso de visita y descarga automática se realizó mediante la herramienta HTTrack (versión 3.44-1). Se trata de un robot o araña bajo licencia GPL. Este programa es configurable y permite la descarga de sitios web completos a un directorio local, bajando todos los directorios, ficheros html, imágenes, etc., siguiendo todos los enlaces (internos y externos) de cada una de las páginas. HTTrack es robusta frente a interrupciones de la red, seguir los cambios de direcciones, y permite sacar un informe de errores.

El proceso de visita se realizó a mediados de noviembre de 2011. La herramienta se configuró de tal manera que sólo se visitaran las páginas principales (ni enlaces internos ni externos). Los 1027 sitios se suministraron mediante un fichero de texto plano. Como resultado final del proceso se pudieron descargar un total de 951 páginas válidas, correspondientes a 951 sitios distintos, que son la base del análisis posterior. La diferencia entre el total seleccionado (1027) y lo conseguido en la descarga (951) se debe fundamentalmente a problemas de desaparición de las páginas, páginas fuera de línea (p.e. servidores en mantenimiento o fuera de línea en ese momento) y a posibles fallos en las conexiones de Internet durante el periodo de trabajo de la araña.

García Balboa, J. L., Ariza López, F. J., Ureña Cámara, M. A. y Ureña-López, A. (2012): "Análisis de métodos de etiquetado geográfico de documentos html en internet y su aplicación en España", GeoFocus (Artículos), nº 12, p. 122-146. ISSN: 1578-5157

4.3. Búsqueda de estilos de geoetiquetado

El proceso de búsqueda de las geoetiquetas consiste en la localización del patrón formado por las palabras clave de cada tipología de etiquetas. Este proceso se puede realizar de manera elemental con cualquier editor de textos. Para este trabajo se ha utilizado el programa Notepad++ v5.9.2 que permite el análisis recursivo en directorios.

Se han realizado búsquedas de cortas cadenas de texto (ver el parámetro "patrón buscado" en la tabla 2, relativas al código html necesario para incluir geoetiquetas según los diferentes estilos. Se ha realizado la búsqueda para cada uno de los estilos descritos en el apartado 2 (ICBN, GT, MFG y RDFaG). En esta búsqueda también se han incorporado:

- La cadena "georss" para explorar el uso del geoetiquetado basado en este modelo de datos.
- La cadena "maps.google.com" como patrón de inclusión del servicio de Google Maps. Si bien Google Maps no es un mecanismo de geoetiquetado, se ha considerado de interés su inclusión por ser una opción, a priori, presumiblemente bastante difundida que, además, presenta la ventaja de ser reconocida por los usuarios dado su carácter visual. Ello permitirá tener una estimación cuantitativa de un recurso que resulta común y plantear con ello una comparación más ajustada del grado de adopción de las geoetiquetas.

4.4. Análisis de resultados

Los resultados obtenidos se resumen en la tabla 2. El recuento se ha desglosando por sitios web y etiquetas. La primera opción contabiliza el número de sitios que incluyen una o más etiquetas de un mismo estilo, y la segunda, el número de etiquetas de ese estilo en los sitios anteriores. Ambos casos se expresan tanto en cantidad contabilizada como en porcentaje respecto al total (951 casos). Como se puede observar, se trata de unos resultados pobres en cuanto a la presencia de geoetiquetas, y por tanto las posibilidades de análisis quedan limitadas. No obstante, se pueden realizar las siguientes consideraciones de interés respecto a la muestra analizada:

- Los resultados apuntan a un uso limitado del geoetiquetado.
- Las escasas páginas geoetiquetadas están relacionadas, principalmente, con sitios pertenecientes a actividades turísticas.
- El estilo GT es el más difundido (0,94%), seguido del ICBM (0,63%) y del MFG (0,21%).
- El estilo RDFaG y las propuestas basadas en el modelo de datos GeoRSS no tienen presencia.
- El número de sitios geoetiquetados con Google Maps es superior al de sitios etiquetados con el estilo GT, pero con una diferencia porcentual limitada (1,26% frente al 0,94%).
- Aquellas páginas que disponen de geoetiquetado ICBM también adoptan el GT.

García Balboa, J. L., Ariza López, F. J., Ureña Cámara, M. A. y Ureña-López, A. (2012): "Análisis de métodos de etiquetado geográfico de documentos html en internet y su aplicación en España", GeoFocus (Artículos), nº 12, p. 122-146. ISSN: 1578-5157

- Los sitios con presencia de geoetiquetado suelen incluir más de una etiqueta. El caso más significativo es el de las etiquetas GT, donde en 9 sitios aparecen un total de 23 etiquetas, lo que da un ratio de 2,55 etiquetas por sitio.
- Por el contrario, si el recuento se realiza por número de etiquetas, la situación se invierte. La llamada al servicio de Google suele ser única, posiblemente porque como resultado proporciona una imagen a incrustar en el sitio web (recordemos que no se trata de un estilo de geoetiquetado). Aquí ya aflorarían criterios de diseño web, en los que tendría poco sentido reiterar en un mismo sitio el rico mensaje visual proporcionado para la zona de interés. No obstante, no es objetivo de este trabajo el analizar el servicio de Google Maps y por ello no se entra en mayor detalle.
- En relación a las tres fuentes utilizadas para la creación de la muestra, los ratios de etiquetas por número de elementos en la muestra son los siguientes: 0,1667 en Google, 0,0294 en Alexa y 0,0078 en la selección guiada. Estos datos indican grandes diferencias según el origen. Tomando como base el menor de ellos (selección manual), en Alexa hay casi cuatro veces más mayor número de etiquetas y en Google esto se dispara hasta 21 veces más. Dado que la pertenencia a uno de estos índices se basa en el liderazgo de visitas consideramos que, de alguna manera, este liderazgo también debe tener una contrapartida tecnológica en el geoetiquetado.
- Respecto a la información procedente de la referida *Whois DataBase*, que indicaba un uso a nivel mundial de etiquetas ICBM del 0,33%, los resultados apuntan a un ratio de casi el doble entre nuestro trabajo y esa fuente, valor de ratio que, además, ha de ser entendido con las debidas precauciones por las diferencias entre los dos métodos. Por otra parte, dado que en Whois DataBase no se indica nada sobre otras tipologías, no se pueden comparar los resultados obtenidos para los demás casos. Cabe aquí también mencionar los resultados de Fink *et al*. (2009), centrados en el estudio de blogs, arrojan un valor del 0,11% de la suma de etiquetas ICBM y GT. La suma de estos dos estilos en la tabla 2 daría lugar a un valor bastante superior, del 1,57%.

Como complemento a los resultados obtenidos en el muestreo, sería interesante conocer su distribución estadística. La técnica de remuestreo conocida como *bootstrap* (Efron, 1979) permite establecer un valor medio de estimación, y la incertidumbre (en forma de desviación típica) asociada, mediante el uso de la simulación. En este trabajo se ha aplicado el *bootstrap* por medio del paquete estadístico R. Como parámetro fundamental del proceso se han considerado 10000 simulaciones, lo que permite obtener los resultados presentados en la tabla 3. Tomando como ejemplo el caso de las etiquetas de Google Maps, al establecer 10000 iteraciones se están considerando 10000 poblaciones de sitios web con una estructura poblacional equivalente a la que se ha conformado para este trabajo (951 sitios seleccionados), lo que lleva a esperar, en término promedio, la presencia de 12,02 ± 3,43 sitios web (1,26 ± 0,36 %) con ese tipo de etiquetas, en esta tipología de poblaciones.

García Balboa, J. L., Ariza López, F. J., Ureña Cámara, M. A. y Ureña-López, A. (2012): "Análisis de métodos de etiquetado geográfico de documentos html en internet y su aplicación en España", GeoFocus (Artículos), nº 12, p. 122-146. ISSN: 1578-5157

5. Conclusiones

Este trabajo se centra en el análisis comparativo de los estilos más difundidos para incorporar información geográfica en los archivos html, lo que viene a denominarse etiquetado geográfico o geoetiquetado. La incorporación de este geoetiquetado es importante para el desarrollo de servicios basados en la posición o LBS (*location-based services*). Aunque el propósito del geoetiquetado es simple, existen múltiples opciones, muchas de ellas con características muy similares. Existe pues una situación de evidente confusión que, además, limita la consolidación de esta tecnología e impide conformar un estándar. Sería muy adecuada una actividad de normalización que abarcara todas las comunidades de interés que han desarrollado este tipo de etiquetas al objeto de proponer un modelo único lo más robusto posible (p.e. con posibilidades de etiquetar líneas y superficies, uso de CRS, etc.).

La información sobre los estilos suele estar publicada, de forma muy heterogénea en cuanto al nivel de detalle, en los respectivos sitios webs de las diferentes autoridades que los proponen, y sin ningún tipo de vínculo entre ellas, ya que se trata de propuestas aisladas. Además en la red existen diferentes publicaciones en blogs, páginas personales, etc. poco rigurosas sobre la temática, en muchas ocasiones fundamentadas simplemente en experiencias de uso de distintos usuarios. Este trabajo ha pretendido suplir la carencia de análisis que permiten conocer la situación actual, resumiendo las principales características de los diferentes estilos, y explicitando sus ventajas e inconvenientes. Se ha analizado la existencia de documentación y herramientas, la información contenida en el estilo, ubicación de las etiquetas en el archivo html y formato de coordenadas geográficas. Además se ha prestado especial atención al análisis de la utilización de los sistemas de referencia y de la incertidumbre de la información geográfica proporcionada, aspectos básicos en el ámbito de la cartografía y SIG, y que se observa no han sido tratados, o muy ligeramente, en los distintos estilos. La información obtenida puede orientar al gestor de una web a la hora de elegir los estilos a utilizar, de forma aislada o combinada, y servir de punto de partida para futuros trabajos sobre propuestas generales de uso de los diferentes estilos.

Este trabajo ha incluido una parte experimental centrada en el estudio de la adopción de las geoetiquetas en España. El análisis se ha realizado sobre un total de 951 sitios web, entre los que se encuentran los más visitados por los ciudadanos. Para la selección de sitios se han utilizado los rankings de Alexa y de Google, más una selección guiada basada en búsquedas en Google a partir de un conjunto de palabras clave. Los resultados indican una adopción muy limitada de las geoetiquetas, pero en general superior a la publicada por *Whois DataBase* para el caso de las etiquetas ICBM a nivel mundial (0,33%). El estilo de geoetiquetado más utilizado es el GT (0,94%), seguido por el ICBM (0,63%) y el MFG (0,21%), y por último el RDFaG del cual no se ha hallado ningún caso de uso. Además, en esta parte experimental también se han considerado los códigos de llamada a Google Maps para poder establecer una comparación frente a este servicio de Google. El número de sitios en los que se han encontrado llamadas a este servicio es escasamente superior (1,26%) al número de sitios en que se ha utilizado el estilo GT. Además, gracias a las técnicas de bootstrap, se ha podido estimar la distribución estadística de los valores obtenidos.

García Balboa, J. L., Ariza López, F. J., Ureña Cámara, M. A. y Ureña-López, A. (2012): "Análisis de métodos de etiquetado geográfico de documentos html en internet y su aplicación en España", GeoFocus (Artículos), nº 12, p. 122-146. ISSN: 1578-5157

5. Agradecimientos

Este trabajo ha sido cofinanciado por el Fondo Europeo de Desarrollo Regional (FEDER) y por el proyecto P08-TIC-4199 de Excelencia de la Junta de Andalucía. Asimismo, agradecemos a la Junta de Andalucía la financiación económica del Grupo de Investigación Ingeniería Cartográfica (PAIDI-TEP-164) desde 1997 hasta la fecha.

6. Referencias

Amitay, E., Har'El, N., Sivan, R. y Soffer, A. (2004): "Web-a-Where: Geotagging Web Content", en: *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval* (SIGIR'04), Sheffield, Reino Unido, Julio 2004. [consulta: 1-03-2012]. Disponible en: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.107.7621&rep=rep1&type=pdf

Berners-Lee, T., Fielding, R. y Masinter, L. (2005): *Uniform Resource Identifier (URI): Generic Syntax*. *The Internet Engineering Task Force* (IETF). [consulta: 1-03-2012]. Disponible en: https://datatracker.ietf.org/doc/rfc3986/

Berners-Lee, T., Hendler, J. y Lassila, O. (2001): "The Semantic Web: A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities", *Scientific American*, 284, 34-43. [consulta: 1-03-2012]. Disponible en: http://www.scientificamerican.com/article.cfm?id=the-semantic-web

Crockford, D. (2006): *The application/json Media Type for JavaScript Object Notation (JSON)*. Informational. The Internet Engineering Task Force (IETF). [consulta: 8-05-2012]. Disponible en: http://tools.ietf.org/html/rfc4627

Efron, B. (1979): "Bootstrap Methods: Another Look at the Jackknife", *The Annals of Statistics*, 7, 1, pp. 1-26.

Daviel, A. (1999): *Geographic registration of HTML documents*. *Version 0*. Internet-draft. The Internet Engineering Task Force (IETF). [consulta: 23-03-2011]. Disponible en: http://tools.ietf.org/html/draft-daviel-html-geo-tag-00.

Daviel, A. y Kaegi, F. A. (2007): *Geographic registration of HTML documents. Version 8.* Internet-draft. The Internet Engineering Task Force (IETF). [consulta: 23-03-2011]. Disponible en: http://tools.ietf.org/html/draft-daviel-html-geo-tag-08.

Daviel, A., Kaegi, F. A. y Kofahl, M. (2007): *Geographic extensions for HTTP transactions. Version 5.* The Internet Engineering Task Force (IETF). [consulta: 23-03-2011]. Disponible en: http://tools.ietf.org/html/draft-daviel-http-geo-header-05.

Delboni, T. M., Borges, K. A. V., Laender, A. H. F. y Davis, C. A. (2007): "Semantic Expansion of Geographic Web Queries Based on Natural Language Positioning Expressions", *Transactions in GIS*, 11, 3, pp. 377-397.

García Balboa, J. L., Ariza López, F. J., Ureña Cámara, M. A. y Ureña-López, A. (2012): "Análisis de métodos de etiquetado geográfico de documentos html en internet y su aplicación en España", GeoFocus (Artículos), nº 12, p. 122-146. ISSN: 1578-5157

Efron, B. (1979): "Bootstrap Methods: Another Look at the Jackknife", *Annals of Statistics*, 7, 1, pp. 1-26.

Fink, C., Piatko, C., Mayfield, J., Finin, T. y Martineau, J. (2009): "Geolocating Blogs From Their Textual Content", en: *Proceedings of the AAAI 2009 Spring Symposium*, Standford, CA, EE.UU., marzo 2009.

GeoJSON (2008): *The GeoJSON Format Specification*. [consulta: 08-05-2012]. Disponible en: http://www.geojson.org/geojson-spec.html

Graupmann, J. y Schenkel, R. (2006): "GeoSphereSearch: Context-Aware Geographic Web Search", *SIGIR '06. Workshop on Geographic Information Retrieval*.

Hu, Y. y Ge, L. (2007): "GeoTagMapper: An Online Map-based Geographic Information Retrieval System for Geo-Tagged Web Content", *International Perspectives on Maps and The Internet. Lecture Notes in Geoinformation and Cartography*, Parte B. Berlin, Springer, pp. 153-164.

ISO (2003): *ISO 19112:2003. Geographic information -- Spatial referencing by geographic identifiers.* International Organization for Standardization.

ISO (2006a): *ISO 3166-1:2006. Codes for the representation of names of countries and their subdivisions -- Part 1: Country codes*. International Organization for Standardization.

ISO (2006b): *ISO 19138:2006. Geographic information -- Data quality measures*. International Organization for Standardization.

ISO (2007): ISO 3166-2:2007. Codes for the representation of names of countries and their subdivisions -- Part 1: Country subdivision code. International Organization for Standardization.

ISO (2008): ISO 6709:2008. Standard representation of geographic point location by coordinates. International Organization for Standardization.

Jones, C. B., Abdelmoty, A. I., Finch, D., Fu, G. y Vaid, S. (2004): "The spirit spatial search engine: Architecture, ontologies and spatial indexing", *Proceedings of the third international conference geographic information science* (GIScience 2004), Adelphi, MD, EE.UU, octubre 2004.

Mayrhofer, A. y Spanring, C. (2010): *A Uniform Resource Identifier for Geographic Locations ('geo' URI).* Standards Track. The Internet Engineering Task Force (IETF). [consulta: 20-05-2011]. Disponible en: https://datatracker.ietf.org/doc/rfc5870

McCurley, K. S. (2001): "Geospatial Mapping and Navigation of the Web", en: *Proceedings of Tenth International World Wide Web Conference* (WWW10), Hong Kong, mayo 2001. [consulta 1-03-2012]. Disponible en: http://www10.org/cdrom/papers/278

Microformats (2011a): *Microformats wiki*. [consulta: 23-03-2011]. Disponible en: http://microformats.org/wiki/Main_Page.

Microformats (2011b): *Geo*. Microformats wiki. [consulta: 23-03-2011]. Disponible en: http://microformats.org/wiki/geo.

Microformats (2011c): *hCard 1.0*. Microformats wiki. [consulta: 23-03-2011]. Disponible en: http://microformats.org/wiki/hcard.

Pyalling, A., Maslov, M. y Braslavski, P. (2006): "Automatic Geotagging of Russian Web Sites", *Proceedings of the 15th international conference on World Wide Web* (WWW2006), Edimburgo, Escocia, mayo 2006. [consulta: 1-3-2012]. Disponible en: http://wwwconference.org/www2006/programme/p72.html.

Ruiz, A.M., García, J.L. y Mesa, J.L. (2010): "Error, incertidumbre, precisión y exactitud, términos asociados a la calidad espacial del dato geográfico", en Alcázar *et al*. (Eds.): *Catastro: formación, investigación y empresa. Selección de ponencias del I Congreso Internacional de Catastro Unificado y Multipropósito*. Servicio de Publicaciones, Universidad de Jaén, pp. 95-102. [consulta: 20-05-2011]. Disponible en: http://coello.ujaen.es/congresos/cicum/ponencias/Cicum2010.2.02_Ruiz_y_otros_Error_incertidumbre_precision.pdf.

Raymond, E. S. (1996): *The new hacker's dictionary.* The MIT Press. Massachusetts Institute of Technology.

Scharl, A. (2007): "Towards the Geospatial Web: Media Platforms for Managing Geotagged Knowledge Repositories", en Scharl, A. y Tochtermann, K. (Eds.): *The Geospatial Web. How Geobrowsers, Social Software and the Web 2.0 ahre Shaping the Network Society*. London, Springer, pp 3-14.

Schiller, J. y Voisard, A. (2004): *Location-Based Services*. San Francisco, Morgan Kaufmann, Elsevier.

Studer, R., Benjamins, V. R. y Fensel, D. (1998): "Knowledge Engineering: Principles and methods", *Data & Knowledge Engineering*, 25, pp. 161-197.

Torniai, C., Battle, S. y Cayzer, S. (2007): "Sharing, Discovering and Browsing Geotagged Pictures on the Web", en Scharl, A. y K. Tochtermann (Eds.): *The Geospatial Web. How Geobrowsers, Social Software and the Web 2.0 ahre Shaping the Network Society*. London, Springer, pp 159-170.

Tsai, F. S. (2011): "Web-based geographic search engine for location-aware search in Singapore", *Expert Systems with Applications*, 38, 1, pp. 1011-1016.

Viana, W., Bringel-Filho, J., Gensel, J., Villanova-Oliver, M. y Martin, H. (2007): "PhotoMap – Automatic Spatiotemporal Annotation for Mobile Photos", W2GIS2007, *Lecture Notes in Computer Science*, 4857, Berlin, Springer, pp. 187-201.

W3C (1999): *HTML 4.01 Specification*. W3C Recommendation 24 December 1999. [consulta: 23-03-2011]. Disponible en: http://www.w3.org/TR/1999/REC-html401-19991224/html40.pdf.gz.

W3C (2006): *Basic Geo (WGS84 lat/long) Vocabulary*. Revision 1.21. W3C Semantic Web Interest Group. [consulta: 29-05-2011]. Disponible en: http://www.w3.org/2003/01/geo.

W3C (2007): *W3C Geospatial Vocabulary*. W3C Incubator Group Report 23 October 2007. [consulta: 29-05-2011]. Disponible en: http://www.w3.org/2005/Incubator/geo/XGR-geo-20071023.

W3C (2008): *RDFa Primer. Bridging the Human and Data Webs*. W3C Working Group Note 14 October 2008. [consulta: 23-03-2011]. Disponible en: http://www.w3.org/TR/xhtml-rdfa-primer.

García Balboa, J. L., Ariza López, F. J., Ureña Cámara, M. A. y Ureña-López, A. (2012): "Análisis de métodos de etiquetado geográfico de documentos html en internet y su aplicación en España", GeoFocus (Artículos), nº 12, p. 122-146. ISSN: 1578-5157

W3C (2011): *GeoRDF*. W3C Wiki, 15 March 2011. [consulta: 8-05-2012]. Disponible en: http://www.w3.org/wiki/GeoRDF.

TABLAS

Tabla 1. Principales características de los estilos de geoetiquetado analizados

Característica	ICBM	GT	MFG	RDFaG
Organismo que lo formaliza	ninguno	IETF	Microformats	SWIG, W3C
Documentación sobre el estilo	no	sí	sí	sí
Herramientas en línea	frecuentes	frecuentes	escasas	no
Tipo de localización	puntual	puntual	puntual	puntual[3]
Posicionamiento indirecto	no	sí	no	no
Datos obligatorios	latitud, longitud	latitud, longitud[1]	latitud, longitud	latitud, longitud
Datos opcionales	-	altitud, país, región, dirección postal, nombre común	-[2]	altitud
Posición en el documento html	cabecera	cabecera	cuerpo	cuerpo
Formato de las coordenadas geográficas	grados decimales	grados decimales	grados decimales	grados decimales
CRS horizontal	WGS84	WGS84	WGS84	WGS84
CRS vertical	-	WGS84	-	"elipsoide de referencia local"
Otros CRS	no	no	no	no
Incertidumbre	no	número de decimales	no	no
Recomendaciones de uso	no	algunas directrices	no	no

[1] Si las coordenadas hacen referencia a un país o región, el estilo GT permite obviar las coordenadas latitud, longitud e incluir las etiqueta referente al país y región.
[2] El estilo MFG es una parte del microformato *hCard*, que permite incluir mayor cantidad de datos.
[3] Aquí se ha considerado el vocabulario BGV, no la propuesta del GeoXG.

Fte. Elaboración propia.

García Balboa, J. L., Ariza López, F. J., Ureña Cámara, M. A. y Ureña-López, A. (2012): "Análisis de métodos de etiquetado geográfico de documentos html en internet y su aplicación en España", GeoFocus (Artículos), nº 12, p. 122-146. ISSN: 1578-5157

Tabla 2. Resultados de uso de las geoetiquetas en la muestra de sitios seleccionada

	ICBM	GT	MFG	RDFaG	GeoRSS	Google Maps
Patrón buscado	name="ICBM"	"geo.	class="geo"	"geo:	"georss"	maps.google.com
Nº Sitios encontrados	6	9	2	0	0	12
% de Sitios encontrados	0.63 %	0.94 %	0.21%	0.00 %	0.00%	1.26 %
Nº de etiquetas encontradas	6	23	7	0	0	15
% de etiquetas encontradas	0.63 %	2.41 %	0.72 %	0.00 %	0.00%	1.57 %
Total de sitios analizados: 951.						

Fte. Elaboración propia.

Tabla 3. Valores estimados por *bootstrap* para el parámetro "nº de sitios encontrados" de la tabla 2

Etiqueta	Media	Desviación típica
ICBM	6.04 (0.64%)	2.48 (0.26%)
GT	8.97 (0.94%)	2.97 (0.31%)
MFG	1.99 (0.21%)	1.40 (0.15%)
Google Maps	12.02 (1.26%)	3.43 (0.36%)

Fte. Elaboración propia.

[1] Con anterioridad a la versión 8 del borrador del estilo GT, las etiquetas definidas eran: *geo.position*, *geo.placename* y *geo.region*. La primera ha permanecido, la segunda ha sido sustituida por la etiqueta *geo.lmk* y la tercera ha sido sustituida por las etiquetas *geo.country* y *geo.a1*. Aún puede encontrarse un gran número de documentos html que usan esta estructura más antigua.

[2] Una ontología es una especificación formal (procesable por un ordenador) y explícita de una conceptualización compartida (por un grupo) (Studer, Benjamins y Fensel, 1998).

[3] http://www.geo-tag.de/generator/en.html

[4] http://www.mygeoposition.com

[5] http://www.willamowius.de/geo-tags.html

[6] http://www.addressfix.com

[7] http://www.geo-tag.de/validator/en.php

[8] Un URI (Uniform Resource Identifier, o Identificador Uniforme de Recurso) es una secuencia compacta de caracteres que identifica un recurso abstracto o físico (RFC 3986, Berners-Lee *et al.*, 2005).