Automatic Generation of Geospatial Metadata for Web Resources*

Aneta J. Florczyk[1], Francisco J. Lopez-Pellicer[1],
Javier Nogueras-Iso[1], F. Javier Zarazaga-Soria[1]

[1]Department of Computer Science and Systems Engineering, Universidad Zaragoza, Spain
{florczyk, fjlopez, jnog, javy}@unizar.es

Abstract

Web resources that are not part of any Spatial Data Infrastructure can be an important source of information. However, the incorporation of Web resources within a Spatial Data Infrastructure requires a significant effort to create metadata. This work presents an extensible architecture for an automatic characterisation of Web resources and a strategy for assignation of their geographic scope. The implemented prototype generates automatically geospatial metadata for Web pages. The metadata model conforms to the Common Element Set, a set of core properties, which is encouraged by the OGC Catalogue Service Specification to permit the minimal implementation of a catalogue service independent of an application profile. The performed experiments consisted in the creation of metadata for Web pages of providers of Geospatial Web resources. The Web pages have been gathered by a Web crawler focused on OGC Web Services. The manual revision of the results has shown that the coverage estimation method applied produces acceptable results for more than 80% of tested Web resources.

Keywords: Metadata Generation, Geospatial Web, Web Resource, Geospatial Web Resource provider, Spatial Data Infrastructure

DOI: 10.2902/1725-0463.2012.07.art8

1. INTRODUCTION

Some Web resources that are not part of a Spatial Data Infrastructure (SDI) can be an important source of information, or at least an additional source. There are resources generated by experts in the field of geographic information that are published on the Web but not published for discovery in any SDI catalogue. Also, there is geographic information generated by communities of Web users, known as "neo geography" (Turner, 2006), "naïve geography" (Egenhofer and Mark, 1995) or "Volunteered Geographic Information (VGI) (Goodchild, 2007), whose importance as a source of information for SDI highlights recent research (Craglia et al, 2008; Keßler et al, 2009). Other Web resources that traditionally are not published in SDI catalogues are geoportals, although they are an important source of Geospatial Web resources, and a variety of Web resources that are of interest for geospatial tools and systems, such as the Location Based Systems (LBSs) systems that use spatially annotated Web pages.

Regardless of the source, the Web resources that are not within a SDI could enrich it. However, their incorporation requires a significant effort to create their metadata. This is caused by the different approach to the discovery of resources on the Web and within a SDI. The resource discovery in the context of SDI is based on the Digital Library paradigm, where the geospatial metadata of resources are gathered in catalogues for their search and retrieval (Béjar et al, 2009). The most common metadata profiles used by SDIs are based on Dublin Core (DC, Powell et al, 2011) and ISO 19115 (ISO/TC 211, 2007).

Standardized metadata help users to identify and select relevant digital resources. The DC metadata standard has been also adopted in the Web community. However, metadata creation is costly and Web content creators rely on the crawler-based resource discovery on the Web. The general search for resources on the Web is mainly based on the usage of search engines, whose content is being created through specialised agents, known as robots or crawlers, which automatically analyse the content of Web resources and relationships among them (Page and Brin, 1998). As a result, a lot of content without being described properly is created and added to the Web continuously. Therefore, a tool capable of the automatic generation of metadata to characterise those Web resources with conformance to the requirements of a SDI would facilitate the inclusion of such Web resources into the SDI. Such tool should emulate functionality of a Web crawler by the analysis of the content of a Web resource and relations between such resource and resources associated with it.

The aim of this work is to propose architecture of a system dedicated to the automatic creation of geospatial metadata of Web resources. Such process is similar to the common ETL (Extract, Transform and Load) process intensively investigated in the databases field. From the point of view of the geospatial community, the ETL process is one of the main elements of Business Intelligence infrastructure (Percivall and Singh, 2011). Such systems aim to support better business decision-making and apply techniques used for identifying, extracting and analysing business data, and ETL processes are used to extract data from unstructured formats and translate it into a usable format.

In the context of this work, such architecture should be able to generate standardised metadata from resources in variety of formats. It should be prepared to support various metadata models and different types of Web resources (i.e. easily extensible). A heuristic-based method for geographic scope estimation of Web pages has been proposed as well. A prototype capable of generating a geospatial metadata (in DC profile) of a HTML Web page has been developed and tested. The metadata model is the Common Element Set, a set of core properties, which is encouraged by the OGC Catalogue Service Specification to permit the minimal implementation of a catalogue service independent of an application profile (Nebert et al, 2007). In this way, such model can be easily tailored to the requirements of an SDI (for example, by using crosswalks (Nogueras-Iso et al, 2004)). The experiments have been run on a realistic corpus made of OGC Web Service (OWS) capabilities documents, which has been generated by a Web crawler focused on OWS (López-Pellicer et al, 2011).

The rest of the article is organised as follow. Section 2 presents the state of the art of approaches in automatic creation of metadata focused on Web pages. Section 3 presents a coverage estimation method, and then, Section 4 presents briefly the proposed architecture. Section 5 describes the implemented prototype, and results of performed experiments. In the end, the conclusions are given and future work is outlined.

2. AUTOMATIC GENERATION OF METADATA

In this Section, we revise well-known solutions for the automatic generation of metadata proposed by the Geospatial and Web communities. The proposals of the Geospatial community are appropriate for Geospatial Web resources; however, they cannot be applied successfully to Web resources such as geoportal Web pages. Therefore, the usage of metadata in Web pages has been examined and the main research works on this issue from the Web community has been studied.

2.1. Geospatial Community

In the context of SDI, metadata are usually created manually by experts or by data producers. This task is costly (in terms of time and effort) but ensures high quality (Kalantari et al, 2010). There are some works on automatic generation of geospatial metadata dedicated to SDI. For example, Kalantari et al (2010) propose a tagging-based framework to create, update and enrich spatial metadata. Many tools help to generate automatically metadata for a variety of geospatial resources. For example, CatMDEdit[1] is a metadata editor that can generate metadata from different geospatial data file formats, and from OWS instances via their *getCapabilities* response. Other example is a task synchronisation-based workflow proposed in Manso-Callejo et al (2010). This workflow can compile up to 83 metadata elements for geospatial datasets (i.e. images, vector data and DTM). The customisation of proprietary GIS software can also help in automatic metadata generation for geospatial datasets (Batcheller, 2008). This approach requires data preparation, management and documentation. Therefore, it cannot be applied efficiently in services-oriented environments where on-demand data products are generated by geo-services chained dynamically. Yue et al (2010) propose a solution based on metadata tracking in geospatial service chains.

One of the interesting proposals from the geospatial community is the GeoKettle[2], a metadata-driven Spatial ETL tool, which is dedicated to the integration of different spatial data sources for building and updating geospatial data warehouses. It support a variety of formats, however, Web pages are not supported.

2.2. Web Community

There is much work done in the field of research on the development and maintenance of digital Web resources metadata (Ossenbruggen et al, 2004; Nack et al, 2003; Foulonneau et al, 2008). Greenberg et al (2001) shows that non-professionals equal professionals in the creation of metadata for Web resources. However, Web content publishers do not pay attention to ensure appropriate description of the resources or deliberately distort it (Golliher, 2008). For example, content publishers keep on using metadata trying to gain visibility within search engines because the metadata contained within HTML Web pages were used to rank. Today, this belief is useless because search engines rank resources mainly with graph-based algorithms (Brin and Page 1998).

The geospatial-based solutions that use Web resources (e.g. HTML Web pages) as part of searchable content are mainly LBS systems, which are popular in mobile environments. They require a previous description and the indexing of the resources for their further retrieval, and in these terms, they are similar to Web search engines. Although the metadata published by a Web resource may not be reliable, it may be still be used as the base for the automatic creation of

[1] http://catmdedit.sourceforge.net/
[2] http://www. geokettle.org/

metadata. The header section of HTML document compliant with the HTTP W3C Recommendation[3] may contain metadata via the *meta* elements (<META>), which contain a property-value pair, i.e. *name* (property name) or *http-equiv* (value of header of the HTTP response) and *content* (property value). A schema attribute may be added to specify how to interpret the property value. These values can be eventually described via a metadata profile declared in *head* element via a uniform resource identifier (URI). For example, the Dublin Core Metadata Initiative[4] (DCMI) recommends a DC metadata profile[5] that can be encoded using HTML elements and attributes. However, there is no specification that enumerates legal values of a *name* attribute. The mapping of the metadata used popularly in the Web to a required metadata model might be developed by analysing the W3C and WHATWG (Hickson, 2011; Hickson, 2011a) recommendations and lists of the *meta* elements frequently used in Web pages gathered by initiatives such as metadata.org[6]. There are a variety of *meta* elements apart from *coverage* of DCMI for describing the geographic scope of a Web resource. The geographic scope might be represented as a disambiguated textual description of a location, a spatial object (e.g. a point or a bounding box) or both (see Table 1).

The Named-entity recognition (NER) tools apply natural language processing techniques to identify toponyms in a text. The results may contain false positives, i.e. words or phrases that are not toponyms or are not toponyms in the used context within the analysed text. A geocoder georeferences a toponym and returns a ranked list of matching locations (Goldberg, 2008). Research on toponym resolution focuses on georeferencing toponyms in the text (Zong et al, 2005, Jones and Purves, 2008). The effectiveness of this task depends on the reference dataset and the used algorithm. A place may have several names (e.g., endonyms and exonyms) that may change over time. Its footprint may also change over time. These changes can result in an incomplete datasets. The algorithm must take into account the ambiguity: (1) common words should be distinguished from proper names (geo ambiguity / non-geo) (Amitay et al, 2004), and (2) the mapping between toponyms and locations is ambiguous (e.g. there are about 40 inhabited places named "London" in the world). There is a variety of approaches, such as using other toponyms in the text to improve the toponym disambiguation (Overell and Rüger, 2008), and the usage of simple taxonomies based on gazetteers (Amitay et al, 2004) or more complex ontologies (Jones et al, 2001), which might be transformed to a graph for the “importance” score

[3] http://www.w3.org/TR/html401/struct/global.html
[4] http://dublincore.org/
[5] http://dublincore.org/documents/2008/08/04/dc-html/
[6] http://www.metatags.org/all_metatags

computation (Silva et al, 2006). The next Section presents a heuristic-based method for geographic scope estimation of Web pages.

Table 1: Geospatial *Meta* Elements used in Web Pages.

Meta Element	***Format***	***Note***
ICBM[7]	latitude, longitude	WGS 84, (e.g., "51.66,6.88")
geo.position (geotags[8])	latitude, longitude	WGS 84, (e.g., "51.66;6.88")
geo.placename (geotags)	free text placename	Placename (e.g., "Steinbergweg, 46514 Schermbeck, Germany")
geo.region (geotags)	ISO 3166-2 code (ISO, 2007a)	Code of country subdivision (e.g., "DE-Nordrhein-Westfalen")
DC.coverage [.x/y/z/ placeName/ longitude/ latitude]	x/y/height/ placename/ longitud/ latitude	The coordinate system must be defined by the additional scheme attribute when x or y is used. The WGS 84 is default system for latitude and longitude (e.g., "World", "51.66, 6.88")
geographic-coverage	place-class, lower-case / code	Region definition according to WHATWG (e.g., "city, Sao Paulo, Sao Paulo, Brazil")

3. COVERAGE ESTIMATON

The goal of the coverage estimation method is assigning the minimum bounding box (*mbox*) to a Web page. First, the used heuristics are described, and then, some details are presented. Finally, some disadvantages of the method developed are discussed.

The coverage estimation method consists in two heuristics: a content-based heuristic (*H3*) and a host-based heuristic (*Hhip*). The first one estimates the coverage by analysing geographic information found within different elements of Web pages (mainly the geocoded toponyms). The heuristic named *Hhip* heuristic is used when the *H3* has not been successful. It infers a country code (ISO 3166-1 alpha-2 codes (ISO, 2007)) from the host (host name or IP), which then is geocoded to its *mbox*. Finally, the coverage estimation method (*H3*+*Hhip*) returns: a *mbox*, a *code*, a textual representation of the geographic scope, and some provenance information (see Section 5.1).

[7] http://geourl.org/add.html
[8] http://geotags.com

Figure 1: Overview of the Coverage Estimation Method.

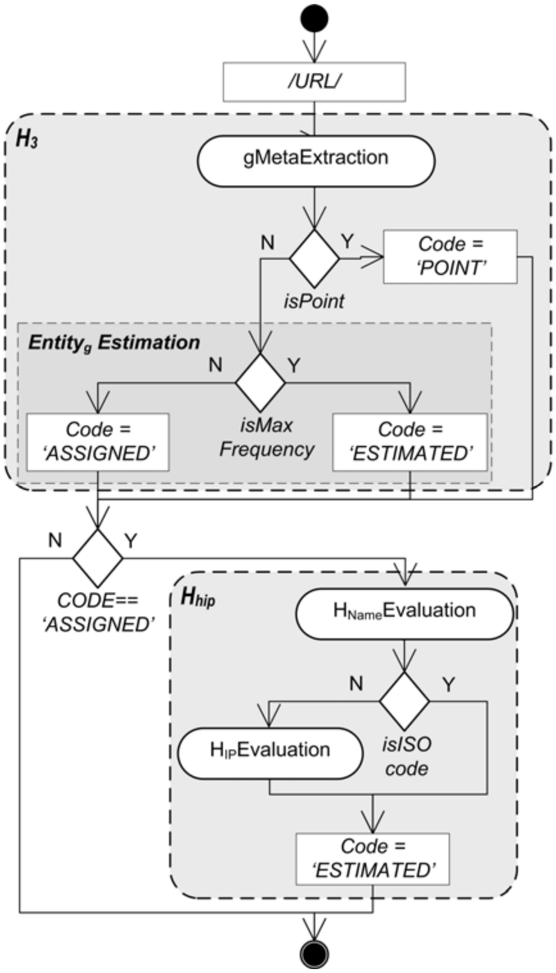


Figure 1 shows an overview of the coverage estimation method giving emphasis to the *code* attribute. It can be observed that the final value of the *code* might be POINT or ESTIMATED. First, the content-based heuristic tries to identify the geo-metadata (*gMeta*) within header metadata that provide latitude and latitude. If this information is not provided (see Section 5.3.), the content-based heuristic focuses on place names found within the Web page. The task of coverage estimation from a text comprises of three general steps:

1. *Toponym recognition.* This step produces a candidate place names list (*nerList*).

2. *Toponym resolution*. This step identifies the geographic entity (*entityg*, an element of a simple territorial ontology) to which refer each place name in the *nerList*, and it produces a set of *entitiesg* (*geList*).
3. *Geo-scope estimation*. This step tries to estimate the *mbox* that the best represent the *geList*.

In this work, the task of the estimation of the representative geographic entity from a set of toponyms found in a Web page is called *EntitygEstimation*. Two external tools are used: a NER tool and a geocoder. According to the Web page element, the NER tool creates one of the following *nerLists:* (1) *gMetaPn* that is a *nerList* of *gMeta* identified within header metadata of Web page; (2) *metaPn* that is a *nerList* extracted from header metadata (other than *gMeta*) and *title* element of Web page; and (3) *bodyPn* that is a *nerList* of Web page body (i.e. the visible text and invisible tags of images).

The geocoder is used to create the *geList* from a *nerList*. A ranked list of geographic entity proposals is created for each item of the *nerList*. A simple territorial ontology has been used in this work, which is result of the analysis of three existing standard models: the FIPS 10–4 standard for countries, dependencies, areas of special sovereignty and their principal administrative divisions developed by the United States Federal Government (NIST, 1995); the ISO 3166 Codes for representation of names of countries and their subdivisions (ISO, 2007a); and the Nomenclature of Territorial Units for Statistics (NUTS) developed by the EU (EC, 2003). In this simple ontology, geographic entities are the concepts, and the only relationships of interest are the spatial aggregations, i.e., “has–part” or “part–of”. It is a modification of the Administrative Unit domain ontology proposed in López-Pellicer et al (2008). Additionally, natural phenomena and towns have been considered as well. The resultant ontology gathers geographic entities of the following types: *feature* (FT) that represents a natural phenomena (e.g. rivers, continents), *earth region* (ERT) that defines international organisations (e.g. “European Union”), *country* (CT), *region* (RT) that represents the top level administrative divisions of a country, *sub-region* (SRT) that represents the rest of administrative divisions, and *town* (TT) that refers to cities. For example, in case of “Barcelona” toponym, the expected *entityg* is “Barcelona” (TT) in “province of Barcelona” (SRT) in “Catalonia” (RT) of “Spain” (CT). The ERT entities are related to countries they gather (“has-part”), and FT entities are related to countries they belong to (“part-of”). The *geList* is created by assigning to each item in the *nerList* the first *entityg* from the ranked list. The *geo-scope estimation procedure* uses a *geList* to calculate frequencies of the *entitiesg* for different levels of accuracy (in the following order: TT, SRT, RT, CT, FT, ERT and EARTH), i.e. each *geList* items is represented via the *entityg* to which the item is related at the given accuracy level (e.g. “Barcelona” TT will be represented by “Catalonia” at RT level of accuracy). The method returns the *entityg* of maximum frequency and the ESTIMATED *code*. If the procedure could not have estimated coverage (e.g. it fails when the *geList* is empty), the “Global” *entityg* and the ASSIGNED code are returned.

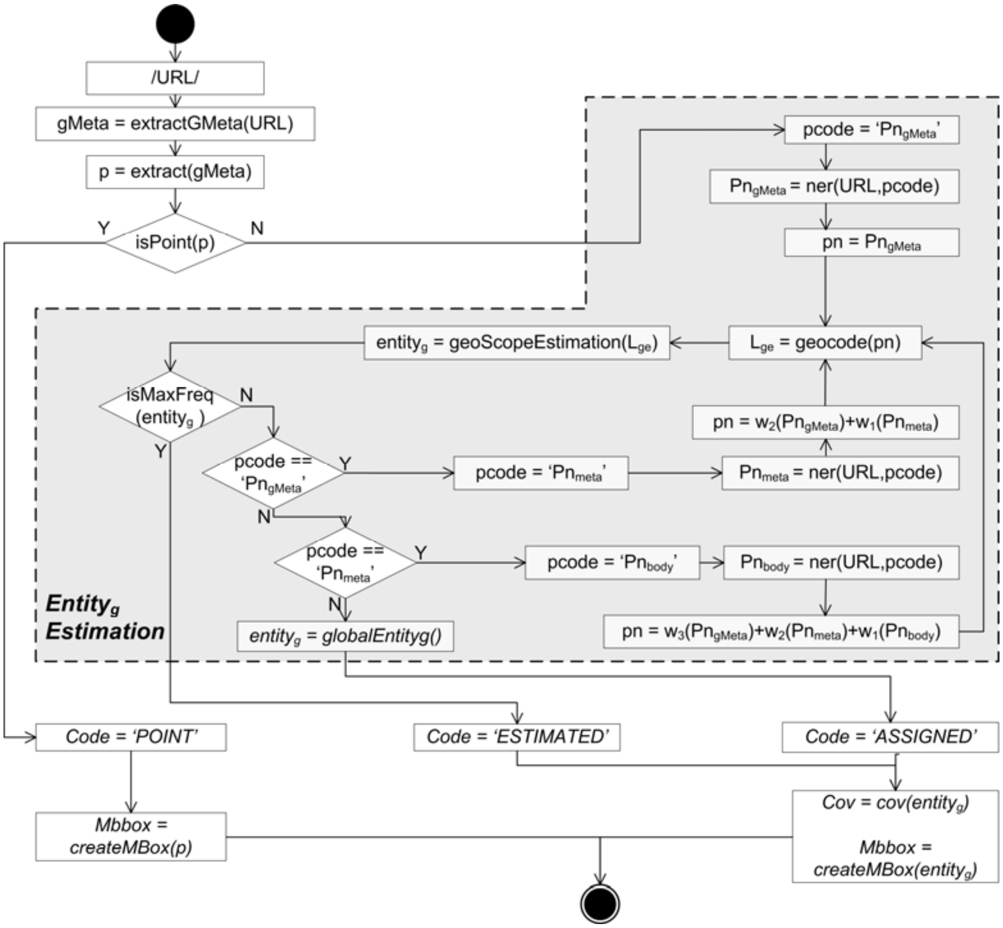
Table 2: Example of the Hhip Heuristic Results.

URL	***Manual estimation***	***Code of H3***	***Hhip HOST***	***Hhip IP***	***Hhip***	***Code of H3+Hhip***
bnhelp.cz	CZ	ASSIGNED	CZ	-	CZ	ESTIMATED
b5m.gipuzkoa.net	Gipuzkoa, Basque Country, ES	ASSIGNED	-	ES	ES	ESTIMATED

The final heuristic (*H3+Hhip*) is performed as follows. First, the content-based heuristic is run (see Figure 2.). The *gMeta* are checked and if a point is provided, it is used to create the *mbox* and the *code* has POINT value. If no spatial object has been distinguished, the text values are analysed to create *gMetaPn* and then corresponding *geList*. If the *geo-scope estimation procedure* fails (the *code* has value ASSIGNED), a weighted list is created by joining the *gMetaPn* are the *metaPn* (*w(gMetaPn)=2*, *w(metaPn)=1*). If the *geo-scope estimation procedure* fails again, the *bodyPn* is added, new weights are assigned (*w(gMetaPn)=3, w(metaPn)=2, w(bodyPn)=1*), and then the *geo-scope estimation procedure* is run again. The host-based heuristic is used only when the heuristic H3 fails to estimate the coverage (i.e. it returns the ASSIGNED *code*), which happens usually due to the lack of metadata and poor NER results. The heuristic *Hhip* tries to extract the ISO country code from host name of the analysed Web page and if it is not successful, its IP is georeferenced to an ISO country code (Table 2 shows some examples). Then, the ISO code is geocoded to its *mbox* and the ESTIMATED *code* is returned.

The developed content-based heuristic is simple and has several problems. First, the candidate place names are trimmed from their context. For example, the geocoder does not consider other place names from the same *nerList*, which have been identified by the NER tool near the searched place name. Also, the ranking is delegated to the geocoder, when the *geList* is created. There has been tested various approaches to define the algorithm that creates *geList* (the location of the other place names within the text, the re-ranking of geocoding list according to other items within the *geList*, etc.). In general, the results of the tests of the other analysed algorithms have not shown a significant difference in performance. This simple approach has been chosen as a compromise between effectiveness and performance. As showed later in Section 5.3, even such simplistic approach can provide satisfactory results in the context of this work.

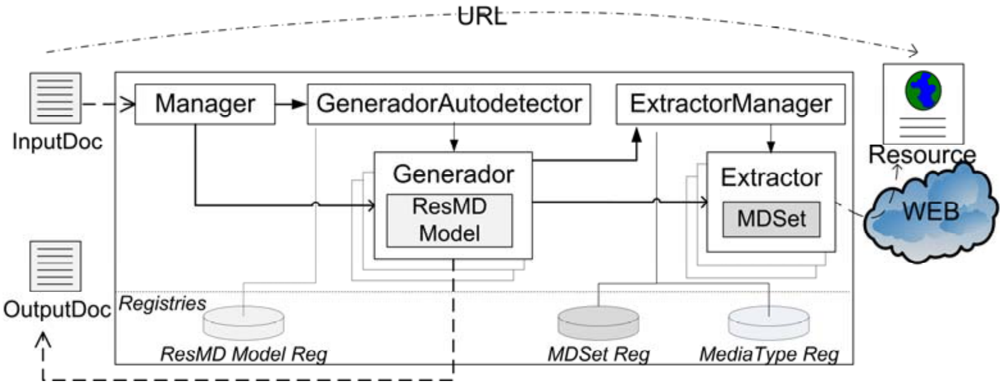
Figure 2: Overview of the Content-based Heuristic.



4. ARCHITECTURE

The proposed architecture for metadata generation presents Figure 3. It receives an XML metadata document (*InputDoc*). The document determines the metadata model that should be generated via the declared schema. The schema (*ResMD Model*) and a module capable to manage it (*Generator*) should be registered in the system a priori (*ResMD Model Reg* and *GeneratorAutodetector* respectively). The *Manager* calls the *GeneratorAutodetector,* which returns proper *Generator* according to the *InputDoc* schema. The *InputDoc* has to contain an URL of the Web resource that should be analysed. The *Generator* looks for it in the model, and then it calls the *ExtractorManager,* which initialises the required *Extractors* according to Web resource Media Type (*Media Type Registry*) and desired functionality. An *Extractor* specialises in extracting of pieces of information (*MD Set*) from the Web resource, for example a list of links within *body*, metadata from the *header* or the content of the *title* in case of Web page. Some extractors (i.e. compound *Extractors*) may apply the existing ones. For example, an *Extractor,* which implements the coverage estimation method introduced in Section 3, may use existing *Extractors* to get the *nerLists*. The *Generator* can implement a variety of logics and can be used to fulfil, validate or improve the input metadata. Next, the result metadata are returned (*OutputDoc*).

Figure 3: The Architecture



5. IMPLEMENTATION AND EXPERIMENT

This Section presents the prototype implementation and describes the performed experiments and their results. First, the supported metadata model is introduced with details on applying mappings. Then, the main characteristics of the implemented prototype are detailed. Next, the used corpus and performed experiment are described, and then the results are discussed.

5.1. Prototype

The implemented prototype is dedicated to the generation of geospatial metadata of Web pages of Geospatial Web resource providers[9]. The generated metadata will be deployed in an OGC catalogue service. Therefore, the metadata model consists of the core returnable properties supported by the OGC catalogue services (Nebert et al, 2007). The Table 3 presents a metadata model and the

[9] More information about the project can be found here: http://dl.dropbox.com/u/79742992/MetGen_project.txt

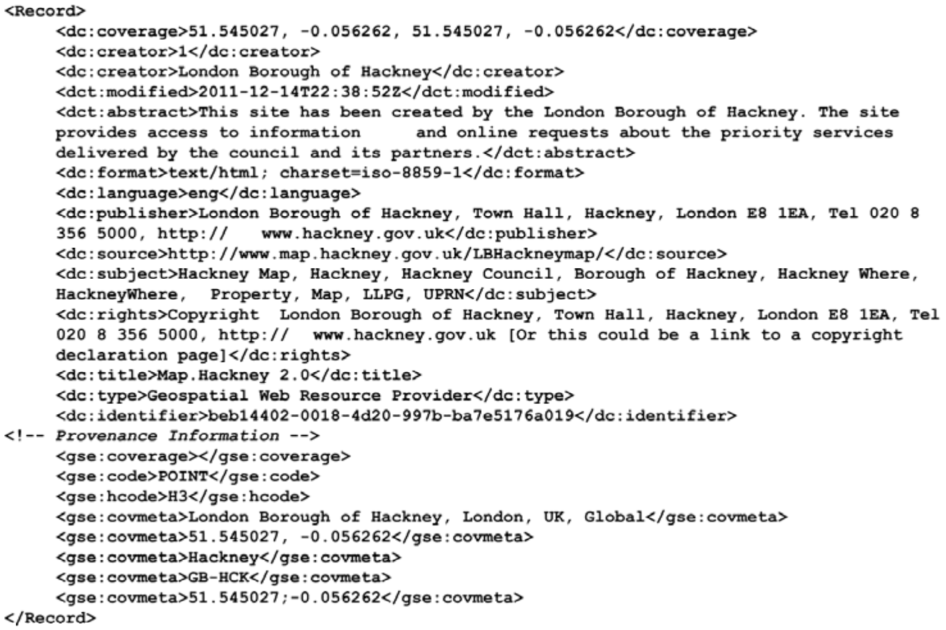
main mappings applied by the system. The system receives an XML document which schema conforms to this metadata model. The *source* element contains the URL of the Web page for which the metadata should be generated. The *modified* (the metadata creation date) *identifier* (a unique identifier of the metadata) and *type* ("Web page of a Geospatial Web resource provider" by default) are filled first. Additionally, the HTML *title* element is mapped to *title* and a procedure to extract copyright information from body element is implemented as well.

Table 3: Metadata Model and Mapping to HTML *Meta* Elements.

CSW Record Element	***Min..Max***	***HTML Element***
contributor	0..*	DC.contributor
coverage	1..*	DC.coverage.*, geographic-coverage ICBM, geo.position, geo.placename geo.region
creator	0..*	DC.creator, author, webauthor
modified	1	-
description	0..*	DC.description, description
format	0..*	DC.format, content-type (http-equiv)
identifier	1	-
language	0..*	DC.language, language
publisher	0..*	DC.publisher, publisher
relation	0..*	DC.relation
rights	0..*	DC.rights, rights, copyrights
source	1	-
subject	0..*	DC.subject, keywords
title	0..*	DC.title, application-name (http-equiv)
type	1	-

The prototype has been implemented in Java and supports redirection (HTTP redirections and simple JavaScript declaration to follow a link). The TIKA project has been selected as the base for the implementation of the architecture. A generator that works with the above metadata model uses several extractors (the geo-metadata extractor and metadata extractor which work on the header element, the title element extractor, coverage estimator and several *nerList* extractors). The extractors that extract the *nerLists* use the Stanford NER (Finkel et al, 2005). The NER tool is configured in function of the Web page language. Various classifiers have been tested (Manning and Klein, 2003, Faruqui and Pado 2010) and, in this prototype, the NER tool only handles properly English or German text (that covers 60% of the cases used in the experiment as showed later). By default, the classifier trained with an English language corpus is selected. The coverage estimator implements the method presented in Section 3. A geocoding module has been implemented to support required functionality. It uses the Google Maps API as an external geocoder, which has been chosen due to its global coverage, a rich data model (it includes administrative divisions up to CT), multilanguage support and relevance ranking. It was necessary to provide a geographic ontology with parent-child relations ("part-of" and "has-part") for the other types of geographic features, i.e. *feature* and *earth region*. Additionally, an extension for the ISO codes of countries and their subdivisions (ISO, 2007a) has been added because the external geocoder does not support them adequately.

Figure 4: Generated Metadata after Applying the Coverage Estimator.



```
<Record>
     <dc:coverage>51.545027, -0.056262, 51.545027, -0.056262</dc:coverage>
     <dc:creator>1</dc:creator>
     <dc:creator>London Borough of Hackney</dc:creator>
     <dct:modified>2011-12-14T22:38:52Z</dct:modified>
     <dct:abstract>This site has been created by the London Borough of Hackney. The site
     provides access to information      and online requests about the priority services
     delivered by the council and its partners.</dct:abstract>
     <dc:format>text/html; charset=iso-8859-1</dc:format>
     <dc:language>eng</dc:language>
     <dc:publisher>London Borough of Hackney, Town Hall, Hackney, London E8 1EA, Tel 020 8
     356 5000, http://   www.hackney.gov.uk</dc:publisher>
     <dc:source>http://www.map.hackney.gov.uk/LBHackneymap/</dc:source>
     <dc:subject>Hackney Map, Hackney, Hackney Council, Borough of Hackney, Hackney Where,
     HackneyWhere,  Property, Map, LLPG, UPRN</dc:subject>
     <dc:rights>Copyright  London Borough of Hackney, Town Hall, Hackney, London E8 1EA, Tel
     020 8 356 5000, http://  www.hackney.gov.uk [Or this could be a link to a copyright
     declaration page]</dc:rights>
     <dc:title>Map.Hackney 2.0</dc:title>
     <dc:type>Geospatial Web Resource Provider</dc:type>
     <dc:identifier>beb14402-0018-4d20-997b-ba7e5176a019</dc:identifier>
<!-- Provenance Information -->
     <gse:coverage></gse:coverage>
     <gse:code>POINT</gse:code>
     <gse:hcode>H3</gse:hcode>
     <gse:covmeta>London Borough of Hackney, London, UK, Global</gse:covmeta>
     <gse:covmeta>51.545027, -0.056262</gse:covmeta>
     <gse:covmeta>Hackney</gse:covmeta>
     <gse:covmeta>GB-HCK</gse:covmeta>
     <gse:covmeta>51.545027;-0.056262</gse:covmeta>
</Record>
```

Figure 4 shows an example of the metadata generated by the prototype. It might be observed that apart from the *dc:coverage* element some provenance information is offered as well, i.e. coverage *code* (*gse:code*), coverage in textual format (*gse:coverage*), the extracted geo-metadata (*gse:covmeta*) and the *heuristic code* (*gse:hcode*) that informs which heuristic produced the coverage.

The provenance information on metadata generation process is not required by a catalogue; however it may help in evaluation of the accuracy of the generated metadata.

5.2. Corpus

The corpus is a realistic set of resources retrieved from geospatial Web resources publishers. A set of OWS URLs returned by an OWS crawler (López-Pellicer et al, 2011) was used to identify the publishers (each URL was trimmed to its host). In this work, we assume that it is highly probably that the Web pages reached via this list are pages published by the service publishers. In other words, we assume that OWSs are related in some way with the Web pages served by the same host.

The OWS host list (1122 elements) was analysed manually (October 2011). More than half of them (51.07%) were not considered due to some errors (e.g. duplication, connection and page loading errors) or did not provide information that might be processed (e.g. "Under construction", an empty page, a server test page). The rest of the OWS hosts (549 elements) were analysed for identifying the geographic scope and the language. The Web pages were classified as follows: 48% are geoportal main pages (Gp), 13.45% are portal main pages (P), 2.36% are pages of a logical part of a portal dedicated to geographic information (P/Gp), 12.73% are map visor based Web pages or even geoportals (AV) and the rest (S) are mainly companies, research group, community or personal pages. The manual estimation of coverage treats a Web site as a whole and considers its published geospatial resources in the estimation. According to the estimated coverage the pages have been classified as follows: 49% are *local* (L), i.e. the coverage refers a part of a country, 30.97% are *national* (N), 2.91% are *regional* (R), i.e. the coverage crosses the country boundaries, and 10.38% are *global* (G). There are 2 examples of Web pages whose coverage has been estimated as *other* (O) because they do not refer to the Earth. It was impossible to define a scope manually (NN) for some pages classified as the rest (S). Most pages are in English (43.90%), German (14.39%), Spanish (12.02%), Polish and Italian (about 4% each), and Czech, French and Catalan (about 3% each). The rest of the examined pages were mainly in one of the official languages of Europe but also in languages of Asia (e.g. Thai or Chinese). Additionally, there are six pages which content is in two languages (e.g. atlastenerife.es).

5.3. Experiments and Results

The experiment consisted in generating the geospatial metadata for each Web page from the corpus. Due to the dynamic characteristics of the Web (temporal unavailability of the Web resources), several test runs have been performed during the period of November and December of 2011. In general, 3.1% of the elements of corpus were not processed due to errors (data format errors, frequently repeated connexion problem, etc.). The applied mapping extracted the following metadata 3.21% of coverage, 97.54% of title, 43.67% of subject, 42.72% of description, 25.71% of creator, 1.32% of contributor, 7.56% of publisher, 10.96% of copyrights, 80.91% of format and 34.22% of language. The generated metadata will be exploited in a catalogue. Therefore, we filter out faulty metadata that is metadata without *title*, *description* or *subject* fields. After removing metadata that do not conform this restriction (2.37%), all remaining elements do have at least *title, description* and *subject* fields. It can be observed that the information about the geographic scope (i.e., coverage) is rare in the examined corpus and it varies in format (for example, textual information in different format, lat/long point). Therefore, the next experiment consisted in applying the coverage estimator to obtain a geographic scope (i.e. minimum bounding box). The Web pages for which it was not possible to estimate coverage manually (6.38%) have been removed from the corpus as well. Table 4 summarises the percentage of removed elements of corpus.

Table 4: Trimmed Corpus (Lang – language).

Lang	*Total processed*	*Process error*	*Not valuable metadata*	*Manual coverage not estimated*
EN	241	11	6	29
DE	79	3	2	1
Other	229	3	5	5
Total	*549*	*17 (3.10%)*	*13 (2.37%)*	*35 (6.38%)*

The results of the coverage estimation experiment (see Table 5) shows that *H3* is able to estimate the coverage in 56% of cases. In the case of these Web pages, It produces *equal* results to the manual estimation of the geographic scope of Web pages in almost half of cases (48.3%). 73.4% of results are correct with the *country* accuracy or better (i.e., the country of the computed *entityg* is equal to the country of the manually estimated *entityg*). In other words, the *H3* procedure yields *acceptable* results in 73.4%, and the erroneous results (i.e., *error*) in 26.6% of cases.

After applying the *Hhip* for the Web pages with assigned “Global” coverage, the coverage estimator produces *acceptable* results in 78.9% of cases. The result is poor due to several problems. The NER tool is not properly configured for almost half of the corpus. Therefore, *H3* produces poor results for them. Surprisingly, *H3* produces a similar percentage of errors for the English Web pages (i.e., the EN corpus). Closer analysis of the corpus and the *H3* results has shown that *H3* behaves worse for pages classified via coverage as *global*, regional or other than for pages classified as *national* or *local*. Since 87.7% pages of those classified as *global*, *regional* or *other* consist of Web pages in English, therefore, the results are worse than expected for the EN corpus. An improvement of the *equal* results after applying the *Hhip* heuristic should not be expected because most of elements of the corpus are classified as *local* and the *Hhip* handles only the *country* level coverage or higher. Nevertheless, the result shows in fact an improvement, i.e. the number of *errors* decreases and the percentage of the *acceptable* results increases. This tendency is not shown in the EN corpus. After meticulous analysis of the results, it has been observed that this effect is produced by the fact that 72.3% of the elements evaluated by the *Hhip* do not permit the estimation of the ISO code from the host name. In such a case, the *Hhip* georeferences IP and it introduces an error. The coverage estimator is quite good for the DE corpus. Therefore, it might be suspected that if the NER tool is properly configured, the coverage estimator is efficient in 91.8% of cases (for Web pages classified as *local* or *national* at least). Nevertheless, the estimator should be improved for the Web pages classified as *global*, *regional* and *other*.

Table 5: Results of the Experiment on Coverage Estimation.

	EN	**DE**	**Other**	***Total***
Total	195	73	216	*484*
H3 (% of Total)	137 (70.26%)	45 (61.64%)	89 (41.20%)	*55.99%*
H3 Acceptable (% of H3)	97 (**70.80%**)	41 (**91.11%**)	61 (**68.54%**)	***73.43%***
H3 Equal (% of H3)	*66 (48.18%)*	*32 (71.11%)*	*33 (37.08%)*	*48.34%*
H3 Error (% of H3)	40 (**29.20%**)	4 (**8.89%**)	28 (**31.46%**)	***26.57%***
H3+Hhip (% of Total)	195 (100%)	73 (100%)	216 (100%)	*100%*
H3+Hhip Acceptable (% of Total)	134 (**68.72%**)	67 (**91.78%**)	181 (**83.80%**)	***78.93%***
H3+Hhip Equal (% of Total)	*81 (41.54%)*	*39 (53.43%)*	*81 (37.50%)*	*41.53%*
H3+Hhip Error (% of Total)	61 (**31.28%**)	6 (**8.22%**)	35 (**16.20%**)	***21.07%***

6. CONCLUSIONS AND FUTURE WORKS

This work presents the automatic generation of geospatial metadata for Web pages. This is a first step to create a tool capable to characterise geospatially Web resources based on the analysis of contextual information provided by related Web pages (for example, KML (Wilson, 2008)). This context-based approach may be also useful to improve existing SDI resource metadata, for example, the metadata offered by an OGC Web Service *getCapabilities* response.

A prototype has been tested on Web pages that belong to the Web sites that publish Geospatial Web resources. The test corpus is composed of Web pages published on hosts that also publish OWS services. An OWS crawler gathered this corpus automatically. On one other hand, the analysed corpus offers a brief overview of Geospatial Web publishers. Geospatial Web resources can be found mainly in geoportals and general portals (almost 64% of the evaluated Web pages) but also in the Web sites of companies, research group, communities and personal Web pages (frequently as demo resources). Most of Web sites have been classified as local (49%) or national (30.97%). The Web sites of regional and global publishers (almost 90%) usually use the English language. The performed experiments show current practices in the Geospatial community in using metadata in Web pages (e.g. lack of geospatial metadata). An important issue is the generation of the geographic scope of retrieved Web resources, which is an obligatory element of the geospatial metadata. The developed prototype shows how simple heuristics can supply automatically this information when a publisher does not provide it.

On of the future work will be the generation of metadata which model is in conformance with the INSPIRE Metadata Implementing Rules (EC, 2007). The model used by the presented prototype will be used as a starting point. As for the coverage estimation method, the future work will focus on providing better support to more languages. The improvement of the coverage estimation method will be investigated, especially in the terms of the identification of a regional or global scope of a Web page. The related resources found in the analysed Web pages will be considered to provide additional information to improve functionality of the system. For example, the extraction of the geographic extent from the capabilities of an OWS linked from a page. As the next step, the work will focus on using the prototype to develop a tool for the validation and the improvement of OWS capabilities by using the contextual information extracted from the Web pages on which links with the OWS request appears. Additionally, this work gives bases for developing a framework for automatic classification of Web sites that publish Geospatial Web resources.

ACKNOWLEDGEMENTS

This work has been partially supported by Spanish Government by the projects TIN2009-10971 and “España Virtual” ref CENIT 2008-1030 (through contracts with the National Center of Geographic Information (CNIG) and GeoSpatiumLab) and Zeta Amaltea S.L. The work of Aneta J. Florczyk has been partially supported by a grant (ref. AP2007-03275) from the Spanish Government.

REFERENCES

Amitay, E., Har'El, N., Sivan, R. and A. Soffer (2004). “Web-a-Where: Geotagging Web Content”, *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 273-280, ACM.

Batcheller, J. K. (2008). Automating geospatial metadata generation - An integrated data management and documentation approach, *Computers & Geosciences*, 34(4):387-398.

Béjar, R., Nogueras-Iso, J., Latre, M.A., Muro-Medrano, P. R. and F. J. Zarazaga-Soria (2009). “Digital Libraries as a Foundation of Spatial Data Infrastructures”, *Handbook of Research on Digital Libraries: Design, Development, and Impact*. IGI Global, pp. 382-389.

Brin S. and L. Page (1998). The anatomy of a large-scale hypertextual Web search engine, *Computer Networks and ISDN Systems*, 30(1-7):107-117.

Campelo, C.E. and C. Souza Baptista (2009). “A Model for Geographic Knowledge Extraction on Web Documents”, *Proceedings of the ER 2009 Workshops on Advances in Conceptual Modeling - Challenging Perspectives*, , pp. 317-326, Springer-Verlag.

Craglia, M., Goodchild, M.F., Annoni, A., Camara, G., Gould, M., Kuhn, W., Mark, D., Masser, I., Maguire, D., Liang, S. and E. Parsons (2008). Next-Generation Digital Earth: A position paper from the Vespucci Initiative for the Advancement of Geographic Information Science, *International Journal of Spatial Data Infrastructures Research*, 3(1):146-167.

EC (2003). Regulation 2003/1059/EC of the European Parliament and of the Council of 26 May 2003 on the establishment of a common classification of territorial units for statistics (NUTS). Official Journal of the European Union.

EC (2007). INSPIRE Metadata Implementing Rules: Technical Guidelines based on EN ISO 19115 and EN ISO 19119. Version 1.2. Guidance Document. Drafting Team Metadata and European Commission Joint Research Centre.

Egenhofer, M.J. and D.M. Mark (1995). “Naive Geography, Spatial Information Theory: A Theoretical Basis for GIS”, *Spatial Information Theory: A Theoretical Basis for GIS, International Conference COSIT '95, Semmering, Austria, September 21-23, 1995, Proceedings*, pp. 1-15.

Faruqui M. and S. Pado (2010). “Training and Evaluating a German Named Entity Recognizer with Semantic Generalization”, *Proceedings of KONVENS 2010*.

Finkel J.R., Grenager T. and C. Manning (2005). “Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling”, *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*, pp. 363-370.

Foulonneau, M. and J. Riley (2008). “Metadata for digital resources: implementation, systems design and interoperability”, *Chandos Information Professional Series*, Oxford.

Goldberg, D.W. (2008). *A Geocoding Best Practices Guide*, North American Association of Central Cancer Registries (NAACCR).

Golliher, S.A (2008). Search Engine Ranking Variables and Algorithms, *semj.org*, 1, Supplemental Issue.

Goodchild M.F. (2007). Citizens as voluntary sensors: spatial data infrastructure in the world of Web 2.0, *Journal of Spatial Data Infrastructures Research*, 2(1): 24-32.

Greenberg, J., Pattuelli. M.C., Parsia. B. and W. Davenport Robertson (2001). Author-generated Dublin Core Metadata for Web Resources: A Baseline Study in an Organization. *Journal of Digital Information*. 2:38-46.

Hickson, I. (2011). HTML5 A vocabulary and associated APIs for HTML and XHTML. Editor's Draft 19 February 2011. W3C, http://dev.w3.org/html5/spec/Overview.html#meta, [accessed 14 May 2011].

Hickson, I. (2011a). HTML Living Standard (accessed 4 January 2012), WHATWG Web Applications 1.0 specification.

Humphreys, J.B.K. (2002). PhraseRate: An HTML Keyphrase Extractor. Technical report, University of California, Riverside.

ISO (2007) 3166-1:2006/Cor 1:2007Codes for the representation of names of countries and their subdivisions -- Part 1: Country codes, http://www.iso.org/iso/country_codes/iso-3166-1_decoding_table.htm, [accessed 20 October 2011]

ISO (2007a). ISO 3166-2:2007 Codes for the representation of names of countries and their subdivisions - Part 2: Country subdivision code, http://www.iso.org/iso/country_codes/iso-3166-1_decoding_table.htm, [accessed 20 October 2011]

ISO/TC 211 (2003). ISO 19115:2003 Geographic information – Metadata, *International Organization for Standardization*, Geneva.

Jones, C.B., Alani, H., and D. Tudhope (2001). "Geographical Information Retrieval with Ontologies of Place", *Spatial Information Theory: Foundations of Geographic Information Science, International Conference, COSIT 2001, Morro Bay,CA, USA, September 19-23, 2001, Proceedings*, pp. 322-335, Springer-Verlag.

Jones, C.B. and R.S. Purves (2008). Geographical Information Retrieval (editorial article), *International Journal of Geographical Information Science*, 22(3): 219-228.

Keßler, C., Janowicz, K. and M. Bishr (2009). "An agenda for the next generation gazetteer: Geographic information contribution and retrieval", *Proceedings of the International Conference on Advances in Geographic Information Systems 2009*, ACM.

Leider, J.L. (2007). *Toponym Resolution in Text*, PhD dissertation, University of Edinburgh, http://hdl.handle.net/1842/1849 , [accessed 20 May 2011].

López-Pellicer, F., Florczyk, A.J., Lacasta, J., Zarazaga-Soria, F. and P. Muro-Medrano (2008). Administrative Units, an Ontological Perspective. In: Song, I.-Y. e. a. (Ed.), Advances in Conceptual Modeling – Challenges and Opportunities. Vol. 5232 of Lecture Notes in Computer Science. Springer Berlin / Heidelberg, pp. 354–363. DOI: 10.1007/978-3-540-87991-6_42

López-Pellicer, F.J., Florczyk, A.J., Béjar, R., Muro-Medrano, P.R. and F.J. Zarazaga-Soria (2011), Discovering geographic web services in search engines, *Online Information Review*, 35(6):909-927.

Manning C. and D. Klein (2003). Optimization, Maxent Models, and Conditional Estimation without Magic. Tutorial at HLT-NAACL 2003 and ACL 2003.

Manso-Callejo, M., Wachowicz, M., Bernabé-Poveda, M., Sánchez-Alonso, S. and I.N. Athanasiadis (eds.) (2010). "The Design of an Automated Workflow for Metadata Generation", *Metadata and Semantic Research*, Springer Berlin Heidelberg, 108, pp. 275-287.

Nack, F., Ossenbruggen, J. and L. Hardman (2003). That Obscure Object of Desire: Multimedia Metadata on the Web, Part II, *IEEE Multimedia*, 12(1):54-63.

Nebert, D., Whiteside, A. and P. Vretanos (eds.) (2007). OpenGIS Catalogue Services Specification. OpenGIS Implementation Specification. Version

2.0.2, Corrigendum 2 Release. OGC 07-006r1. Open Geospatial Consortium Inc.

NIST (1995). FIPS PUB 10–4: Standard for Countries, Dependencies, Areas of Special Sovereignty and Their Principal Administrative Divisions. National Institute of Standards and Technology, Gaithersburg, MD, USA.

Nogueras-Iso, J., Zarazaga-Soria, F., Lacasta, J., Béjar, R. and P. Muro-Medrano (2004). Metadatastandard interoperability: application in the geographic information domain. Computers, Environment and Urban Systems 28 (6), 611 – 634. DOI:10.1016/j.compenvurbsys.2003.12.004

Ossenbruggen, J., Nack, F. and L. Hardman (2004). That Obscure Object of Desire: Multimedia Metadata on the Web, Part I, *IEEE Multimedia*, 11(4):38-48.

Overell, S. and S. Rüger, (2008). Using co-occurrence models for placename disambiguation, *International Journal of Geographical Information Science*, Taylor & Francis, Inc., 22(3):265-287.

Page, L. and S. Brin (1998). The anatomy of a large-scale hypertextual Web search engine, *Computer Networks and ISDN Systems*, 30(1-7).

Percivall, G. and R. Singh (2011). Geospatial Business Intelligence (GeoBI). OGC White Paper. OGC 09-044r2. Open Geospatial Consortium Inc.

Polfreman, M. and S. Rajbhandari (2008). MetaTools - Investigating Metadata Generation Tools Final Report, London, at http://www.jisc.ac.uk/media/documents/programmes/reppres/metatoolsfinalreport.pdf [accessed 14 May 2011].

Powell, A., Nilsson, M., Naeve, A. and P. Johnston (2005). Dublin Core Metadata Initiative - Abstract Model (White Paper), *DCMI*, at http://www.dublincore.org/documents/abstract-model/, [accessed 14 May 2011].

Silva, M.J., Martins, B., Chaves, M., Afonso, A.P. and N. Cardoso (2006). Adding Geographic Scopes to Web Resources, *Computers, Environment and Urban Systems*, 30(4):378-399

Turner, A. (2006). Introduction to Neogeography. *O'Reilly Media*.

Wilson, T. (2008). OGC KML Version 2.2.0 (OGC 07-147r2), Open Geospatial Consortium, Inc.

Yue, P., Gong, J. and L. Di (2010). Augmenting geospatial data provenance through metadata tracking in geospatial service chaining. *Computers & Geosciences*,36(3):270-281.

Zong, W., Wu, D., Sun, A., Lim, E.P. and D.H. Lian Goh (2005). “On Assigning Place Names to Geography Related Web Pages”, *Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries, June 7-11 2005, Denver, CO, USA*, pp. 354-362.