Proceedings of the 6th AGILE
April 24th-26th, 2003 – Lyon, France

A GEOGRAPHIC SEARCH
WAYS TO STRUCTURE INFORMATION ON THE WEB

Angela Schwering (University of Münster, Fraunhofer ISST),
Roland M. Wagner (Fraunhofer ISST),
Bernd Schneider (University of Münster)

Institute of Information Systems, Muenster University
Leonardo Campus 3, D-48149 Münster, Germany
angela.schwering@uni-muenster.de, bernd.schneider@wi-ios.de

Fraunhofer Institute for Software- and System-Engineering
Emil-Figge-Str. 91, D-44227 Dortmund
roland.wagner@isst.fhg.de

1. INTRODUCTION AND MOTIVATION

The world wide web offers a great amount of information, which is expanding day by day. The variety of genre, content, quality and format and the explosive growth make it increasingly difficult to retrieve data. To make this mess of published information usable, an efficient strategy is needed to find the required information.

Search engines have been developed to make the knowledge on the web usable. Resource discovery has focused so far on text based searching: giving a document a title or using open-text keywords, the engine selects all resources, which correspond to the search query in any way and puts them in order according to a complex algorithm. Another way to find the required information is to navigate through categories. Category based search engines try to organize all the knowledge on the web in root- and subcategories or a network of categories.

But the existing search strategies are very poor to support the natural senses of humans to store and find information. People use their environment to orientate themselves. About 80% of all data and information that are daily retrieved, are associated with a particular place on the Earth's surface. To find ones way in a three dimensional world is one of the most developed natural tasks for human beings. For this reason many people use the geographic location sense to store information. This natural way of organizing and retrieving information can also be used by search engines on the Internet. A new opportunity of searching is given by supporting this spatial way of thinking. For specific search queries the spatial search is not only an alternative to a text based search - it can also disclose a more efficient or a totally new way of search to the user. It may be beneficial to facilitate geographic searching for web pages describing restaurants and retail stores, so that the consumer may select the closest suitable facility.

2. THE GEOGRAPHIC SEARCH

This chapter describes the steps that have to be taken to enable a geographic search. Fig. 1 shows the proposed structure needed to set up a geographic search engine. The cloud in the figure represents the web and the box below is the highly aggregated architecture of the search engine. The architecture itself is motivated by the architecture of knowledge based systems.

In the first step - the acquisition of geographic information - the spatial reference of documents from web sites are elicited. The author of a web site or a reader can be used to guide the information gathering process. Next, there has to be an adequate way to represent this information on the web site.

After the spatial reference is stored in a machine-readable way on the web, a search engine can start to search the web, to discover resources and to build a searchable database for answering user queries. This database is called information base. The results of a search query are shown by the presentation component.

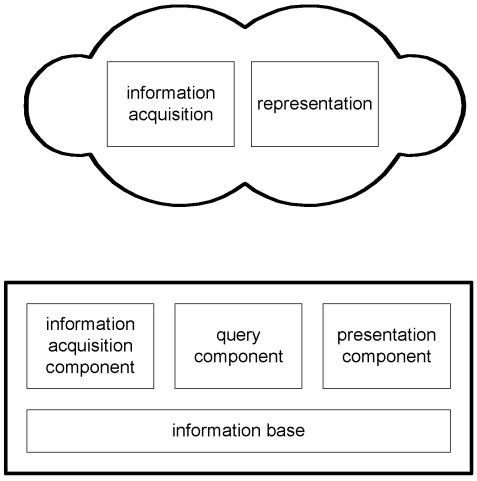

Fig. 1 Aggregated architecture of a geographic search engine

The question how to get and to represent spatial information on the web that it can be required by a search engine is focused in this article. Different methods are compared and assessed and finally the proposed approach is described in detail.

3. OVERVIEW OVER METHODS

The existing ways to gather the geographic information from the web can basically be divided in the manual and automatic acquisition. Fig. 2 shows a detailed overview over the various methods.

3.1 Manual acquisition - ISO 19115 registration

For the manual acquisition the web site author can register his web site via a registration form on the web. This form is based on the ISO 19115 standard for geographic meta data. So everyone can register his web site and insert the information about the geographic area in a standardized way. Because the information contained in the ISO standard is very large, this method is suited for users with professional knowledge. Optionally an editorial staff can check the information before it is stored in the database to ensure the correctness of the data.

3.2 Manual acquisition - meta data

Alternatively to the registration form the web site author can implement the geographic information within meta data. He can use the resource description framework (RDF) to describe the spatial area and store it in an extra XML-file, called geo.xml, or integrate it in the existing HTML-file. To encode geographic objects like points and polygons the Graphical Markup Language (GML) can be used as a namespace for RDF. Besides HTML itself

provides META elements integrated in the title part of the web page do encode meta information.

The location can be encoded by coordinates or by an address that a geo-coder translates in coordinates later. A robot searches for web sites with spatial information and stores them in the information base.

3.3 Automatic acquisition

The automatic acquisition can be done by searching for already existing geographic information. On many web sites contact information are given like the flag, which often correlates with the spatial reference of the web site. In Google's first annual Google Programming Contest Daniel Egnor designed an interface to include a geographic search option based on contact addresses in plain text1. HTML provides the <address> tag to store contact information. But the information inside this tag is written just as plain text with no further structure. As well the URL country code, which gives information on the national origin, can be used as geographic reference. The search engine HOTBOT of Terra Lycos Network offers this search option2.

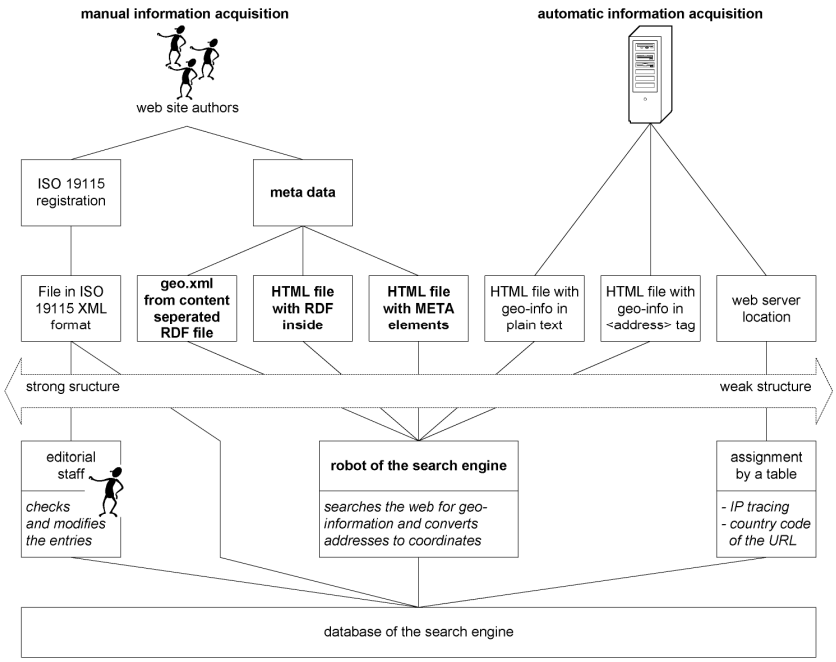

Fig. 2 Different ways of information acquisition

3.4 Requirements concerning the information acquisition

Fig. 2 shows several methods to implement the information acquisition. To be applicable for our purpose the methods have to satisfy some requirements. First of all the method has to be scalable for the web - this means they have to be scalable regarding the information acquisition on the web and regarding the information acquisition of the search engine. Furthermore an uniform definition on spatial reference is needed, e.g. the location of the company, the catchments area of customers etc.

1 For more information see http://www.google.com/programming-contest/winner.html.
2 There exist a lot of search engines that provide geographical search on base of land codes. Hotbot Lycos is only one example. For more information about Hotbot see http://hotbot.lycos.com/.

To reach a sufficient acceptance the way of encoding the information has to be very easy, user-friendly and expeditious. On the one hand these attributes depend on the way of getting the information on the web, but also on the representation (s. fig. 1).

The methods also have to meet the requirements of the information acquisition of the search engine. It is very important that the search engine can get the data with an acceptable speed (performance), independent of the amount of data (scalability). Only in this way the actuality of the data can be ensured. The encoding mechanism has to be able to encode with a sufficient spatial exactness and should not have any problems with changing formats or missing data. Maintainability and reusability are critical factors as well.

3.5 Benchmark of the different methods

A spatial reference of a web site can be stored by a ISO 19115 registration form on the web. Because of its complexity and lots of information that has to be filled in, a web site author has to spend a lot of time on this. An editorial team checking each entry leads to a decreasing performance and therefore to a limited scalability. As well the information is only available for the search machine - this means that the data is not reusable for other applications in the web.

Encoding of meta data in RDF format by using an extra or integrated file are very similar concepts. Tools can support the encoding of the data in a user-friendly way for people that are unfamiliar with the terminology of geographic information systems (GIS). A robot scanning the web constantly and filtering the necessary data leads to high scalability and appropriate actuality as well. Working with missing data is treated as a web site without an associated geographic position and changing formats just need some extensions in the presentation standard.

Because most web site authors are familiar with HTML syntax META tags are easy to use. But expressing complex geographical objects in META tags becomes difficult and error-prone, so the same geographic exactness as with RDF encoding with the GML namespace can not be reached.

	info acquisition + representation on the web				information acquisition of the search engine						
	scalable	uniform terms	user-friendly	expeditious	performance	scalable	actual	geographic exactness	missing data	reusable data concept	benchmark*
ISO 19115 registration	+	+	–	–	+	–	+	+	+	–	2
geo.xml - from content separated RDF file	+	+	+	+	+	+	+	+	+	+	10
HTML file with RDF inside	+	+	+	+	+	+	+	+	+	+	10
HTML file with META elements	+	+	+	+	+	+	+	–	+	+	6
HTML file with geo-info in plain text	+	–	+	+	+	+	+	–	+	–	4
HTML file with geo-info in <address> tag	+	–	+	+	+	+	+	–	+	–	4
web server location	+	–	+	+	+	+	+	–		–	3

* + = plus 1 point

– = minus 1 point

Table 1 Benchmark of the different information acquisition methods

Using already existing information on web pages (e.g. contact address in plain text or within the <address> tag) leads to problems, because the information is not encoded

explicitly. The contact address must not be the same as the spatial reference of a web site. For this reason the uniqueness of the structure of terms is not given automatically. Furthermore the robot can not differentiate between the "real" address and another string using the same lexicographic structure. To rise complexity the kind of noting addresses differ not only from country to country but also from region to region.

To localize web sites by their country code is a very easy way, but for our purpose it is much too imprecise. The country code as well does not necessarily be the spatial reference, too. This concept is only usable on the web.

Table 1 gives an overview of the properties of the different methods. As you can see from the comparison, encoding of meta data in RDF meets the requirements best. As well the concept of using META tags of HTML is a pretty good way. Therefore a technical proposal of these three methods is presented in the next chapter.

4. TECHNICAL PROPOSAL

The technical approach proposed here uses meta data to encode the spatial reference. The content of the meta data represents the geographic reference of the web site.

4.1 Meta data in META tags

The first possibility to encode meta data is through META tags embedded in HTML3. This construct is familiar to many HTML authors and it is concise, clear and very simple to use. There are several identifiers you can choose from depending on giving the location as coordinates or as real world addresses.

The most precise way to encode the spatial reference is to use the identifier geo.position. Generally any coordinate of the system can be chosen - here the coordinates should be expressed in degrees North of Latitude and degrees East of Longitude as signed decimal numbers. The META tag looks like:

```
<META NAME="geo.position" CONTENT="latitude; longitude">
```

The following example is taken from the web site London-MTB Wallpaper, which associates with the position 51,7 degrees North, 0,4 degrees West, which is located near-by London, Great Britain.

```
<HTML>
  <HEAD profile="http://geotags.com/geo">
    <TITLE> ... </TITLE>
    <META name="geo.position" content="51.7;0.4">
  </HEAD>
  <BODY>
    ...
  </BODY>
</HTML>4
```

You can encode the spatial reference of the web site as well in an address. There are five geographic identifiers such as geo.streetaddress to fill in the street name, geo.no to encode the street number, geo.postal for the zip code, geo.city for the town and geo.region for the country and its division. The geo.region identifier can be taken from the international standard on the Representation of Names of Countries and their subdivision ISO 3166.

3 For more information see also the HTML Specification 4.01, Section 7.4.4.

4 This HTML code is taken from the web site of the London-MTB Wallpaper http://www.london-mtb.com/wallpaper/. The webmaster encoded the geographic location using the schema of A. Daviel, which is very close to the one here proposed. For more information about the search engine geotags of A. Daviel see http://geotags.com/.

```
<HTML>
 <HEAD profile="http://myprofile.com/address">
  <TITLE> ... </TITLE>
  <META name="geo.address"Philipp Street">
  <META name="geo.no"584">
  <META name="geo.postal"E14J8G">
  <META name="geo.city"London">
  <META name="geo.region"GB-LND">
 </HEAD>
 <BODY>
  ...
 </BODY>
</HTML>
```

The representation using META tags is very strong in its simplicity, but it is as well limited in power, e.g. it cannot represent a area as the geographic reference. Besides the META tags there exist many standards to encode meta data, which are strong, but also complex. Though only a very small part of these standards is needed and leads to the advantage of well-known, well-defined standards.

4.2 Meta data in RDF

The resource description framework (RDF) is the base to process meta data and emphasizes facilities to enable automated processing of web resources. It can be used in resource discovery to provide better search engine capabilities. Generally RDF is syntax independent but the syntax used in this paper is the Extensible Markup Language (XML). The geographic description of the web site - the RDF code - is saved in an extra file, called geo.xml.

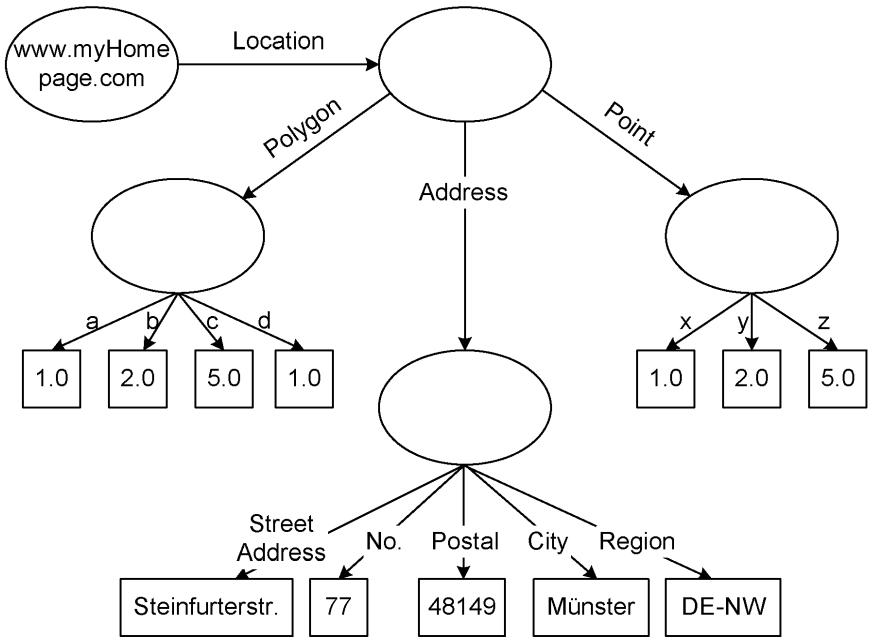

Fig. 3 RDF Model

RDF provides only a framework to process spatial information. For specific purposes integrated namespaces can be used. Here we use the Geographic Markup Language (GML) as the namespace to realize the geographic components.

The RDF data model distinguishes between resources, their properties, and the property values. A web site is represented as a resource with a property location5. If there are several characteristics of the property, it is represented as a structured entity.

Fig. 3 shows a RDF data model describing a web site www.myHomepage.com with its geographic reference, represented as the property location. The location can be encoded as geographic objects, e.g. a point or a polygon, or as a real world address. The example below shows the RDF implementation of the point-location in XML syntax.

```
<?xml version="1.0"?>
<rdf:RDF
 xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
 xmlns:gml="http://www.opengis.net/gml#">
<rdf:Description about="http://www.london-mtb.com/">
  <gml:location rdf:parseType="Resource">
    <gml:Point rdf:parseType="Resource">
      <gml:coord rdf:parseType="Resource">
        <gml:X>51.7</gml:X>
        <gml:Y>0.4</gml:Y>
      </gml:coord>
    </gml:Point>
  </gml:location>
</rdf:Description>
</rdf:RDF>
```

Besides the RDF namespace the GML namespace is imported. GML is a very powerful standard to represent geographical objects. To represent a polygon as a geographic reference just exchange the location with the GML specification for a polygon. 6

The same structure can be used to encode an address. The following example shows the address of the example above with a self defined namespace adr:

```
<?xml version="1.0"?>
<rdf:RDF
 xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
 xmlns:adr="http://www.mySchema.net/adr#">
<rdf:Description about="http://www.schwering.de/angela.html">
  <adr:address rdf:parseType="Resource">
    <adr:StreetAddress>Steinfurterstr.</adr:StreetAddress>
    <adr:No>77</adr:No>
    <adr:Postal>48149</adr:Postal>
    <adr:City>Muenster</adr:City>
    <adr:Region>DE-NW</adr:Region>
  </adr:address>
</rdf:Description>
</rdf:RDF>
```

5 Nodes drawn as ovals represent resources and arcs represent named properties. The value of the properties are described by rectangles no matter of the type (literal, decimal). For more information about the RDF data model and is notation see the Resource Description Framework (RDF) Model and Syntax Specification, Chapter 2.

6 For more information please see the OpenGIS Geography Markup Language (GML) Implementation Specification, version 2.1.2, chapter 4.3.

Compared with the META tag proposal the RDF solution is more complicated. You can encode different geographic objects like points, line strings and polygons and the author can decide and the number of locations a page refers to can be variable.

The resource description framework is a well known and accepted standard. Other progresses are founded on RDF like the ontologic languages OIL and DAML OIL. The integration of features of the semantic web is possible as well.

RDF can be extended by any existing or self defined namespace. To meet special requirements different namespaces can be combined. Because of this flexible concept the resource description framework is open to new developments or extensions.

By the description tag <rdf:Description about="..."> the granularity can be chosen variable - a geographic location for a web site, a web page or an abstract on a page.

4.3 Meta data in RDF integrated in HTML

Because RDF is well-formed XML it can be included directly in the head part of a HTML document. A browser following the HTML recommendations for error handling in invalid documents however will render any exposed string content, which is anything that appears between the end tag and the start tag of the next expression. The RDF code written above within the head of an HTML document with the simple content "Content of the document" would be shown like this:

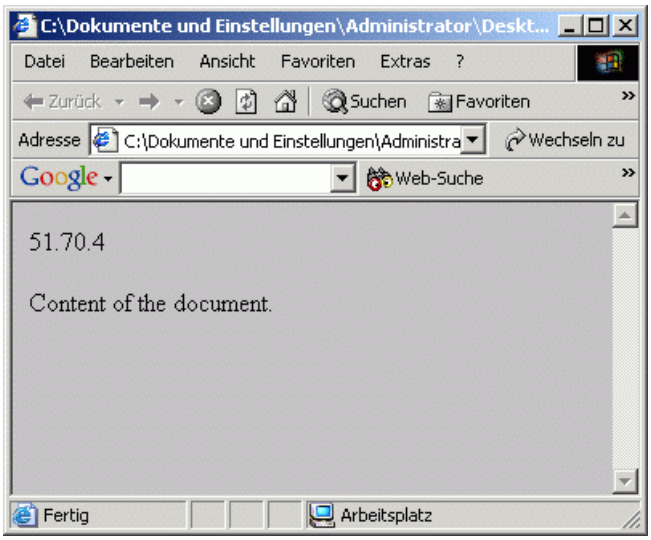

Fig. 4 Interpretation of a RDF syntax in a browser

RDF provides an abbreviated syntax that has no exposed strings. For using the GML namespace, the GML syntax also has to be written in an abbreviated way. The following code gives an example how this abbreviated GML could look like on the base of the container description of RDF:

```
<html>
<head>
 <rdf:RDF
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:gml="http://www.opengis.net/gml#">
  <rdf:Description about="http://www.london-mtb.com/">
   <gml:location rdf:parseType="Resource">
    <gml:Point rdf:parseType="Resource">
```

```
          <gml:coord
            gml:X="51.7"
            gml:Y="0.4"/>
        </gml:Point>
      </gml:location>
    </rdf:Description>
  </rdf:RDF>
</head>
<body>
<P>Content of the document.</P>
</body>
</html>
```

All browsers compliant with HTML 3.2 should only render the content of this HTML document but not the GML coordinates.

5. CONCLUSION

The paper discusses different approaches to enrich "traditional" web sites with geographic information as a base for new search or navigation methods as well as for the so called location based services. The approaches are rated according to the also worked out requirements and the best approach is shown in more detail.

The Meta tag approach benefit from the simplicity of the encoding which can be done very fast and at low costs but this is exceeded by the disadvantage that meta tags can only be used within HTML documents. So other kinds of documents as plain text, ordinary office documents, PDF or images can not be enriched this way.

Because of its open and standardized architecture encoding meta data with RDF in an extra file or within HTML is even more powerful and flexible. Namespaces can be used to implement complex geographic objects.

6. REFERENCES

[1] Daviel, A.; Kaegi, F.: Geographic registration of HTML documents. April 2001, http://geotags.com/geo/draft-daviel-html-geo-tag-05.html, last visited 2002-11-19.

[2] Kurbel, K.: Entwicklung und Einsatz von Expertensystemen. Springer, 1992.

[3] Lassila, O.; Swick, R.: Resource Description Framework, (RDF) Model and Syntax Specification - W3C Recommendation 22 February 1999, http://www.w3.org/TR/REC-rdf-syntax, last visited 2002-11-20.

[4] Ragget, D.; Hors, A.; Jacobs, I.: HTML Specification 4.01 - W3C Recommendation 24. December 1999, http://www.w3.org/TR/1999/REC-html401-19991224, last visited 2002-11-20.