Geohash Index Based Spatial Data Model for Corporate

Conference Paper · August 2015

DOI: 10.1109/ICEEI.2015.7352548

Iping Supriana
Bandung Institute of Technology
288 PUBLICATIONS **430** CITATIONS
SEE PROFILE

Dody Dharma
Bandung Institute of Technology
10 PUBLICATIONS **15** CITATIONS
SEE PROFILE

Dicky Prima Satya
Bandung Institute of Technology
5 PUBLICATIONS **14** CITATIONS
SEE PROFILE

Geohash Index Based Spatial Data Model for Corporate

Iping Supriana Suwardi, Dody Dharma, Dicky Prima Satya, Dessi Puji Lestari
School of Electrical Engineering & Informatics, ITB,
Bandung, Indonesia
{iping, dody, dicky, dessipuji}@informatika.org

Abstract— **Spatial data wrapped and processed by an application known as Geographic Information Systems (Geographical Information System / GIS). In desktop based GIS, the spatial information services only occurs when a variety of basic data has been loaded into the applications database. The development of information technology has provide ready to access, worldwide scale, web based mapping system, like Google Maps. The addition of the information content to the map can also be done easily by any user into the system. Until now, both the basic data provided by Google Maps or data that is added by users are loosely connected, meaning that there are minimum linkage between data. Thus, the data for greater benefit of a corporation or government where the data is closely related to each other still yet to be served. As spatial data being managed are voluminous, the scalability of querying performance will be a challenge. To anticipate this, we describe an improvement that built on top of our proposed spatial data model. We used a special data which derived by interleaving bits obtained from latitude-longitude pairs of a spatial data, the string called geohash. A geohash can be used as an index of every object in Spatial data table. The longer the Geohash string, the more precise the bounding box around the location it references. This approach will improve the performance of querying process of a single or even collection of spatial data in the data table of corporate GIS. The main study of this research is to provide information services along with the availability of a variety of basic spatial data owned by Google Maps. This paper highlights our recent effort in theoretical and applied research in spatial data management.**

Keywords—geohash, geographical information system, spatial data model, mapping system

I. Introduction

Modern technology now permits improved acquisition, distribution, and utilization of geographic or geospatial data (Craglia, 2006). There are many popular web based mapping technology. By using the technology, people are able to get any information based on earth coordinates. The technology are very sufficient for personal and daily use, but there is a need to utilize it for corporate or governmental data needs.

Geospatial information is critical to promote economic development and improve stewardship of natural resources. The use the mapping technology for corporate or governmental need will involve the preparation of the data and the relationship between geographic data in a very detailed level (e.g. to support the interests of plantation management, taxation, city layout, quality of roads, quality of rivers, etc.). Different corporate will have a different data needs, and different data representation.

The development of information technology has provide ready to access, worldwide scale, web based mapping system, like Google Maps. The addition of the information content to the map can also be done easily by any user into the system. Information added by these users may also be registered into Google Maps as information that can be read by any other users. So that a variety of important information that is needed by the public can be easily searched and known, for example the location of hospitals around the world, the location of government institutions, sports location or any location by contributing the data.

Voluminous geographic data have been, and continue to be, collected with modern tools or gadget like GPS or smartphone, location aware services, and crowdsourcing internet geospatial information. The scale of geospatial data is often too large and too unstructured with not enough relationship explanation.

Until now, both the basic data provided by Google Maps and data that is added by users are loosely connected, meaning that there are minimum linkage between data. Thus, the data for greater benefit of a corporation or government where the data is closely related to each other still yet to be served. For example, State Electricity Company (PLN) needs to manage spatial data like electricity substation location, the location of electricity poles, wires between poles and the location of the customer. In other words, web based mapping system like Google Maps can not be utilized effectively for the sake of analysis, especially for the benefit of corporate analysis. In the example above, when the electricity substation hit by a lightning strike, analysis is required to identify any customer who will be affected by a power outage; and then the data of affected customers will appear as a certain color in the Google Maps display

In order to provide facility for services model above, this research proposed. The main study of this research is to provide information services along with the availability of a variety of basic information owned by Google Maps. This paper highlights our recent effort in theoretical and applied research in spatial data management. We are proposing a new model of spatial data management that will accept spatial data structure and create relationship to basic spatial data.

II. State of The Art

The development of information technology enables organizations to be able to perform digital geographic data management. Digital geographic data has been widely used to help solve critical problems through geographic analysis in various industrial sectors. The use of geographic data, among others, can help the analysis of geographic data for marketing optimization through segmentation of customer data; assist urban management and transport; management of natural resources; simulation; and environmental conservation [1]. The use of geographic data today has penetrated into various mobile devices to assist a person in getting the services closest personal information

In general, geographic data wrapped and processed by an application known as Geographic Information Systems (Geographical Information System / GIS). GIS has the ability to perform spatial analysis, which is a set of analysis methods that require access to the attributes and location information of an object (Goodchild (1988) in [2]). Spatial data consists of graphic maps correlated with the attribute table, which enables users to visualize and perform queries quickly. In implementation, a number of GIS has a different way of reading and storing spatial data.

In the early stages, graphical information service is created as computer aided design applications, which provide the service of making a line on the computer to design different drawing interest. Developments of graphical information service is continue specific for mapping which provide the service of making a line on a computer for the purpose of making a map that runs on the local computer (desktop). Further development is providing services in creating line and curvature more easily again that not just runs on desktop platforms, but also runs on a web platform.

Openshaw (1987) commented that such systems are basically concerned with describing the Earth's surface rather than analyzing it (Openshaw(1987) in [3]). In such systems, the spatial information services only occurs when a variety of basic data has been loaded into the applications database. It is a real difficulty for a corporate to build such system, because they have to supply the basic spatial data, their specific spatial data, and define the relation.

In the Google Maps service, information can be directly accessed by the user without worrying the entry of basic spatial data. The basic data likes coordinates, roads, borders, and so on is already provided. In addition, Google Maps has facilitate user that wants to enter the data for the user's own interests. System development opportunities also become possible, because Google Maps has the basic functions that can be accessed by other applications (via the Google Maps API facility). So that developers can develop applications which is an extension of the use of a variety of basic information that Google Maps for corporate needs.

There seems to be widespread agreement in the GIS community on two simple propositions: that as technology, GIS has the potential to support many different type of analysis and that this potential has not yet been realized [5].

We now can obtain much more diverse, dynamic, and detailed data. Generally speaking, geography and spatial sciences have moved from a data-poor era to a data-rich era. The availability of vast and high resolution spatial and spatiotemporal data provides opportunities for gaining new knowledge and better understanding of complex geographic phenomena, such as human-environment interaction and social economic dynamics, and address urgent real world problems, such as global climate changes and pandemic flu spread [4].

To gain those knowledge, there is a need to analyze the spatial data. But in order to do that, the data must be presented in a structure with clear relations between data from different geographic information infrastructure.

To increase efficiency and interoperability of geographic information infrastructure, many regional and global initiatives work in the establishment of open standards and agreements. Content is managed by means of regulated and standardized service types. This imposes a distinct life cycle of geospatial content in distributed environments which can be be described in four steps illustrated in Figure 1. First, content must be made available in distributed system, i.e., content must be published in standard services like discovery and access services. Second, users need to discover content which will be finally accessed by using these services (third step). Finally, users process the content and generate new content, which should be integrated and published in the distributed system closing cycle [5].

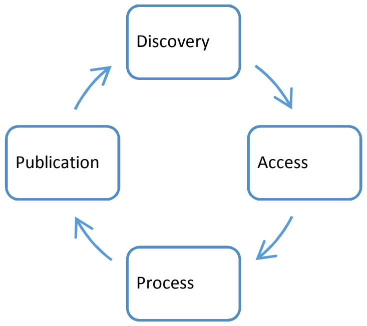

Figure 1. Content life cycle in Geographical Information Infrastructure[5]

Nowadays, because of many regional and global initiatives on mapping, there are so many standard, and not every government or corporate comply with those standard. They usually developed the in-house applications with their own standard because of different data needs.

III. The Research

The concept of graphical and non-graphical data management had developed in the Laboratory of Computer Graphics and Artificial intelligence ITB, a NGGIS (Next Generation Graphical Information System) [6], which provide services of desktop-based GIS system that can use the data in a web network.

This research aims to develop spatial data management technology for the corporate model based on the Google Maps platform. The results of this study can be used as:

1. The basis for the management of corporate data visually-based on Google Maps

   The model can be the basis of data management and general guidelines for various corporations, especially in Indonesia in visualizing its data based on Google Maps that can be useful for business development in Indonesia in the face of global competition, which in turn accelerate and expand economic development in Indonesia.

2. Liaison between corporate information systems with visualization of information.

   With the resulting technology, the information will be presented in a more intuitive and more attractive. Information can be presented visually and is positioned as a layer on Google Maps. With this approach, the analysis of visual information can be done so that the benefit of corporate interests

Based on the purpose of the research and the result of the previous research, we modify the cycles proposed by Trilled et.al. (2013) to explain NGGIS data life cycle.

First, the geospatial data must be made available to NGGIS, but the content must be published in a simple standard that easy to comply by any corporate. NGGIS will learn the data structure and define the relation between the data and the other data that already exist on the system.

Second, users need to discover content with by using a bridging application that doing spatial and textual queries. Third, the query result finally displayed as an information layer on top of popular web based mapping technology. The technology is chosen because of its simplicity and familiarity to the users. Finally, users process the content and generate new content, which should be integrated and published in the closing cycle using the same standard and mechanism.

The system is expected to be able to manage the spatial data needs generally required by various forms of corporations, using popular web based mapping technology as the basis for the presentation of geographical information

Our research will be focused on three main aspects: (1) defining data models; (2) defining information search model for textual queries and spatial queries; and (3) model and simulation analysis.

The first research question that must be solved and will be explained in this paper is to define the spatial data model which is simple enough to comply with. From studies conducted, there are some needs of the data model:

(1) The data model must be flexible in addressing the data requirements of different business entities.

(2) The data model must be adaptive to changing requirements / data structure.

IV. THE PROPOSED DATA MODEL

We build the data model based on relational model. The relational model successfully couples a precise mathematical definition with a useful representation based on tables. Relational model is the basis of relational database, but in our model we change some of characteristics specific to the relational database.

A spatial information is represented as a collection of spatial data tables. Each table is assigned a unique name. A row in table represents a relationship among sets of values. Column headings contains distinct names, and for each column there is a set of possible values, called the domain. The structure is illustrated in Figure 2.

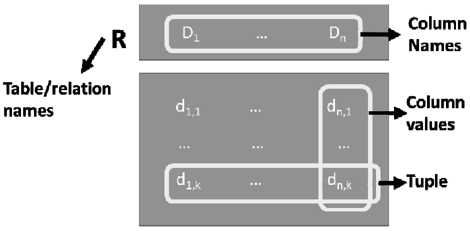

Figure 2. Structure of geospatial information

Each row is an ordered n-tuple of values $<d_1, d_2,.., d_n>$ such that each value d_j is in the domain of j-th column, for $j=1,2,3,..n$. In a database, a sequence of tuples $<d_1, d_2,.., d_n>$ must be distinct, but in our data model we proposed that the sequence is not necessarily distinct, just like in the mathematical set theory. This is because there is probabilities in voluminous geographic data, the data is not distinct and come from many sources.

Relation is defined as the formal description of a correspondence among elements of sets, whereas a table is only a possible representation of the relation itself [7]. By observing a relation represented by means of a table, we can state the following in the context of our geospatial data model:

(1) The values of each column are homogenous

(2) The rows are not necessarily different with respect to one another

(3) The order of the column is irrelevant, since they are always identified by name and not by position

(4) The order of the row may be relevant, since they are identified by content and also by position.

When there are columns that contains unique value that can be used as a key, than the order of the row becomes irrelevant. Identification by position is relevant when there is no column that contains unique value for each row. In this case, the system will create a column that contains row numbers.

The relation between tables is created when there are same column names between those tables. Table that has column with unique values will become the master table for the other table.

Let $r_1(YX)$ and $r_2(XZ)$ be two relations are represented by two tables A and B consecutively where Y and X is the column of A and X and Z is the column of B, such that $YX \cap XZ = X$. Relation between A and B is the join of r_1 and r_2 and is a relation on Y X Z consisting all of tuples resulting from concatenation of tuples in A with tuples in B that have identical values for the attributes X. The relation is illustrated in Figure 3.

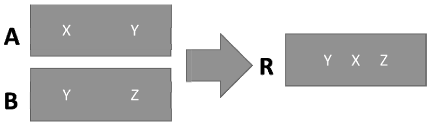

Figure 3. Relation of two spatial tables with common column

If *A* and *B* have no common column then their relation is a Cartesian product. Every table can create a relation with other table by creating the Cartesian product or subset of the Cartesian product between tables. The relation is illustrated in Figure 4.

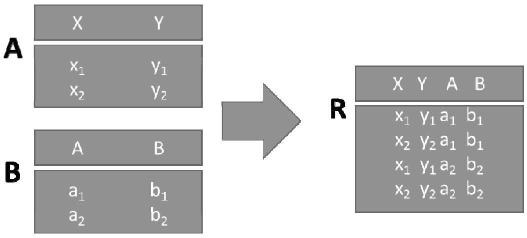

Figure 4. Relation of two spatial tables with no common column

V. GeoHash Spatial Index

In spatial data, the basic data and the most probable common column is the earth coordinates, represented by latitude and longitude for two dimensional, and including height or deep in three dimensional representation. However, rather than storing the coordinate data in multiple column in the data table, we propose to store the hash string of the coordinate data.

In order to map the coordinate data into single shorter string, we use the geocoding algorithm called Geohash [10]. The geohash can be used to partition the information in our spatial data table. Geohash provides a grid-based, hierarchical model of the earth where locations are represented by Base32 strings.

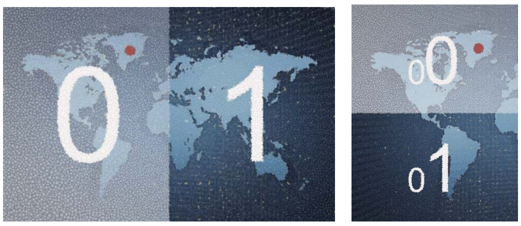

Figure 5. Alternating divisions of the global longitude-latitude rectangle [12]

Each character in in a binary string of geohash indicates alternating divisions of the global longitude-latitude rectangle [−180,180] × [−90,90] (See Figure 5). The first division splits the rectangle into two squares ([−180,0] × [−90,90] and [0, 180] × [−90, 90]). Points (or more generally geometries) which are to the left of the vertical division have a geohash beginning with a '0' and the ones in the right half have geohashes beginning with a '1'. In each of the squares, the next split is horizontal; points below the line receive a '0' and the ones above a '1'. This splitting continues until the desired resolution is achieved. The longer the Geohash string, the more precise the bounding box around the location it references. A Geohash is derived by interleaving bits obtained from latitude-longitude pairs; for example, the decimal coordinates is -45.995 -41.728, we can sub-dividing the space until we get the more detail level. The Geohash will be longer and represented as:

0010110101011100011000110001101111000111

This binary can be represented as alphanumeric characters (32 bit encoded). Each 5 bits is converted to one character:

00101 10101 01110 00110 00110 00110 11110 00111

Which comes out as:

5 p f 6 6 6 y 7

The hash string is *"5pf666y7"*, representing 40 bits of precision (eight characters, five bits per character).

In 2008, Gustavo Niemeyer invented geohashes with the purpose of geocoding specific points as a short string to be used in web URLs (http://www.geohash.org/). He entered the system into the public domain by publishing a Wikipedia page on February 26, 2008 [11]. In order to make geohashes more useful for the web, the inventor assigned a plain text, base-32 encoding for his web service. As binary strings, geohashes can be of any non-negative length, but for web use, geohashes are typically seen in lengths which are multiples of five.

Geohashes implicitly define a recursive quadtree over the world-wide longitude-latitude rectangle. Geohashes provide some notable properties [13].

- Each geohash represent a longitude latitude rectangle.
- Containment. Adding characters to the end of a geohash specifies a smaller rectangle contained inside the initial one. t ⊂ tt ⊂ ttuv. (Figure 7)
- Geohashes provide a z-order traversal of rectangles covering the Earth at each resolution. (Figure 6)
- Locality. Shared prefixes imply closeness. ("dp" is close to "dr"). But adjacent nodes don't have common prefix, like 9z and dp (Figure 7)

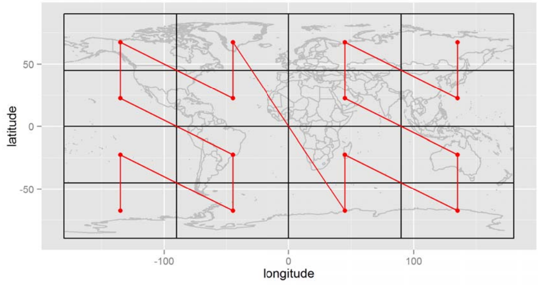

Figure 6. Relation of two spatial tables with common column [13]

Encoding the position takes only (lat, lon) pairs neglecting the third coordinate (altitude) and as such it is unaware of real 3D position of the object in space.

As a consequence of the gradual precision degradation, nearby places will often (not in some edge cases, equator or a meridian) present similar prefixes. The longer a shared prefix is, the closer the two places are.

To retrieve the latitude and longitude bits from an initial pair of coordinates representing a target point in space, the algorithm is applied recursively across successive and more precise geographical regions bounding the coordinates. The remaining geographical area is reduced by selecting a halfway pivot point that alternates between longitude and latitude at each step. If the target coordinate value is greater than the pivot, a 1 bit is appended to the overall set of bits; otherwise, a 0 bit is appended. The remaining geographic area that contains the original point is then used in the next iteration of the algorithm. Successive iterations increase the accuracy of the final Geohash string. An appealing property of the Geohash algorithm is that nearby points will generally share similar Geohash strings. The longer the sequence of matching bits is, the closer two points are.[9].

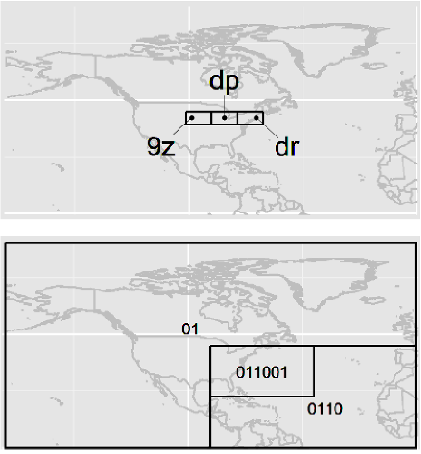

Figure 7. Geohash Locality and Nesting

We can improve the indexing in spatial data table by using a geocoding algorithm because it determines the ranges of information that must be stored in each instance of the index. The prefix of Geohash string of a spatial location are used to determine the group of nodes responsible for storing the data. This has benefit, let's say, by specifying the first two characters (10 bits) of a Geohash can significantly reduce the search space for spatial queries without additional indexing.

VI. Data Model Instatiation

As we explained before, in spatial data, the basic data and the most probable common column is the earth coordinates, but we aware that common column can exists not just in the form of coordinates. Spatial data are often large and complex collections of point. Several point can create a feature called line. Closed loop line can create a feature called region. By using the geohash to represent each point could even simplify the case. For instance, let say we wan to to store collection of point in to table:

{(6° 50′ 53.08″ S, 107° 36′ 35.32″ E), (6° 51′ 41.00″ S, 107° 21′ 22.21″ E), (6° 52′ 77.32″ S, 107° 24′ 13.41″ E), (6° 53′ 53.08″ S, 107° 36′ 35.32″ E)},

Geohash function compact these data into:

{qqu909r543dz, qqu2fyyms03s, qqu2u4118hvu, qqu8b8ppf34c}.

We can see that the strings have common prefix "*qqu*", since they are located in the same region (nearby each other). So they are grouped automatically. This is very powerful to save the space of storage and time to search the data based on location. Since string is a native data type in database system, so it is very easy to be found and filtered using *string-searching* algorithm.

Rather than using desktop geographic information services, Google Maps already created these basic meaningful feature such as place coordinates (points), roads (lines), or city borders (regions). This is the basic geospatial information. Corporates can create information based on this ready to use basic geospatial information. They can focus to their primary and specific geospatial information and combine it with Google Maps basic geospatial information as illustrated in Figure 8.

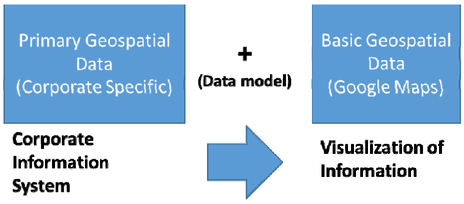
Primary Geospatial Data (Corporate Specific)
+
(Data model)
Basic Geospatial Data (Google Maps)
Corporate Information System
Visualization of Information

Figure 8. Corporate GIS on Google Maps Platform

In order to relate primary geospatial data into basic geospatial data and to relate one primary geospatial data to other primary/secondary geospatial data, we use the proposed data model.

For example, State Electricity Company (PLN) needs to manage spatial data like electricity substation location. They can define the location based on earth coordinates. The location is represented in Table 1.

TABLE I. Substation Location

Substation	Latitude	Longitude
ST001	6° 54′ 53.08″ S	107° 36′ 35.32″ E
ST002	6° 40′ 44.03″ S	107° 25′ 15.13″ E

By using geohash, we can save the data in more compact form, represented in Table II

TABLE II. Substation Location With GeoHash

Substation	Geohash
ST001	qqu909r543dz
ST002	qqu6h2jkxem7

When this table is loaded into NGGIS, system will automatically related it to the basic spatial data, decode the hash string and create a point of ST001 and ST002 as a new layer on top of Google Maps display.

And if we define the region that serve by that substation ST001 as illustrated in Table 3, the system will create relation automatically between Table 2 and Table 3 based on substation column, it was done because they have same column name.

On the other side, system will also relate Table 2 with basic spatial data and create a region, because it has collection of geohash with the same substation value. Region is a collection of earth coordinates with a specific properties and preferably in a closed loop. The region will be drawn as a new layer on top of Google Maps display.

TABLE III. SUBSTATION REGION

Substation	Geohash
ST001	*qqu909r543dz*
ST001	*qqu2fyyms03s*
ST001	*qqu2u4118hvu*
ST001	*qqu8b8ppf34c*
ST002	…
ST002	…

Another instance to show the advantage of using the geohash string to store the location in a relational database system, for example we want to zoom in and out on a map and show a summary of how many points there are on grid square, we can use a geohash prefix length that is relative to the zoom resolution in the relational query syntax:

SELECT SUBSTR (geohash, 0, 2), COUNT (*) FROM locations GROUP BY SUBSTR (geohash, 0, 2);

The query execution will bring the result as:

```
"aa", 67
"ab", 456
…
"qq", 128
...
"zy", 8
```

In a popular RDBMS, MYSQL, we can do query using geohash function, for example:

SELECT ST_AsText (ST_PointFromGeoHash(@gh,0));

This query will give a result:

POINT (45 -20)

A simple index on the geohash column of the above location table allows us to focus any queries on specific areas on the map.

VII. CONCLUSION

Our research is focused on three main aspects: (1) defining data models; (2) defining information search model for textual queries and spatial queries; and (3) model and simulation analysis. The first aspect is already answered by defining the spatial data model which is simple enough to comply with.

The second aspect is partially answered by using our proposed data model which enhanced by the geohash string to represent the coordinate field of spatial data. We can take advantage of existing RDBMS feature which has been natively developed to support string data operation like searching or manipulating string. Even in popular RDBMS like MySQL has a specific Spatial Geohash Functions to encode or decode a hash string. This fact will enhance the overall performance of GIS System for corporate especially in data filtering. To fully answer the second aspect of our research we need explore in more detail relating the information search model.

ACKNOWLEDGMENT

We thank Direktorat Jenderal Pendidikan Tinggi (DIKTI), Department of Education and Culture, Republic of Indonesia, which has funded this research through Penelitian Prioritas Nasional Masterplan Percepatan dan Perluasan Pembangunan Ekonomi Indonesia 2011-2025 (PENPRINAS MP3EI 2011-2025).

REFERENCES

[1] ESRI, “Who use GIS?”, http://www.esri.com/what-is-gis/who-uses-gis, accessed on March 14, 2013.

[2] A. C. Gatrell, Concept of Space and Geographical Data. pp. 119 – 134, 1991.

[3] M. Goodchild, R. Haining, and S. Wise, “Integrating GIS and spatial data analysis: problem and possibilities,” International Journal of Geographical Information System, vol. 6 no. 5, pp. 407–423, 1992.

[4] D. Guo, J. Mennis, “Spatial data mining and geographic knowledge discovery – an introduction,” Computers, Environment and Urban Systems, vol. 33, pp. 403-308, 2009.

[5] S. Trilles, L. Diaz, J. Gill, and J. Huerta, “Assissted generation and publication of geospatial data and metada,” under review for the International Journal of Spatial Data Infrastructures Research, submitted 2013-02-27.

[6] I. Supriana, P. R. Aryan, “Next Generation Graphical Information System engine (NGGIS): tourism application case study,” International Joint Conference TSSA & WSSA, 2006.

[7] P. Atzeni & V. de Antonellis, Relational Database Theory, The Benjamin/Cummings Publishing Company, Inc., 1993.

[8] M Malensek, SPallickara, SPallickara, “Polygon-Based Query Evaluation over Geospatial Data Using Distributed Hash Tables”, 2013

[9] Z Balkić, D Šoštarić and G Horvat, “GeoHash and UUID identifier for Multi agent systems”, 2012

[10] http://geohash.org/site/tips.html

[11] Geohash. http://en.wikipedia.org/wiki/Geohash. [Online; accessed 20-June-2013].

[12] http://www.bigfastblog.com/geohash-intro

[13] A Fox, C Eichelberger, J Hughes, “SkylarSpatio-temporal Indexing in Non-relational Distributed Databases”