

USING GAZETTEERS TO ANNOTATE GEOGRAPHIC CATALOG ENTRIES

Daniela F. Brauner, Marco A. Casanova, Karin K. Breitman, Luiz André P. Leme
Informatics Department, PUC-Rio, Rua Marquês de S. Vicente, 225, Rio de Janeiro – RJ, Brazil

Keywords: Gazetteer, Geographic Metadata Catalog, GIS Integration, Thesauri Alignment.

Abstract: A gazetteer is a geographical dictionary containing a list of geographic names, together with their geographic locations and other descriptive information. A geographic metadata catalog holds metadata describing geographic information resources, stored in a wide variety of sources, ranging from the researchers' personal computers to large public databases. This paper argues that unique characteristics of geographic objects can be explored to address the problem of automating the generation of metadata for geographic information resources. The paper considers federations of gazetteers and geographic metadata catalogs and discusses in detail two problems, namely, how to use gazetteers to automate the description of geographic information resources and how to align thesauri used by gazetteers. The paper also argues why such problems are important in the context of a federated architecture.

1 INTRODUCTION

Scientific data are housed in a wide variety of resources, ranging from the researchers' personal computers to highly organized repositories maintained by organizations. To help users locate and access the available data, a common solution is based on metadata catalogs. Among the metadata stored in catalogs, one typically finds a classification scheme for the data, defined as a structured collection of terms, that is, as a thesaurus.

However, the process of manually generating metadata can be tedious, if not impossible, depending on the amount of new data generated. Therefore, a catalog should be equipped with a component that automates metadata generation as much as possible. Such component may derive metadata from the characteristics of the data acquisition platform, from a (quick) analysis of the data content and from related data and metadata stored in the catalog itself or elsewhere.

Frequently, metadata catalogs are not isolated, but they form a federated system. In this case, a second problem arises, namely, the question of aligning different classification schemes. This problem also requires semi-automated solutions to make the catalog federations viable.

In the geographic information systems domain, we find two valuable ways to address these problems. First, we have various geo-referencing

schemes that associate each geographic object with a description of its position on the Earth's surface, which acts as a universal identifier for the object, or at least an approximation thereof. Second, many gazetteers, or dictionaries of geographic names, have been developed in recent years and made available on the Web.

Based on these preliminary observations, we first propose an architecture that takes advantage of gazetteers to automatically generate meaningful metadata for geographic information resources. However, we also argue that gazetteers will have to be expanded to include information about scale, and catalog metadata schemes will have to be carefully designed to facilitate integration with gazetteers.

Next, we consider federations of gazetteers and geographic metadata catalogs. In this context, we argue that gazetteer thesauri alignment is the central problem, which we propose to solve by introducing a technique that takes advantage of geo-referencing to avoid the pitfalls of aligning the thesauri terms based solely on syntactical proximity.

As for related work, Fu et al. (2003) and Souza et al. (2005), for example, use gazetteers to help index Web resources in general. Klien and Lutz (2005) propose a method for automating the annotation process based on spatial relations.

Our approach uses the gazetteers to describe geographic information resources. As argued in Section 3, we do not limit ourselves to checking the

occurrence of geographic names in the data to be indexed or described, but use geo-referencing to relate gazetteer entries and the data to be catalogued. We also use geo-referencing to address the problem of gazetteer thesauri alignment, as discussed in Section 4.

This paper is organized as follows. Section 2 summarizes the major characteristics of gazetteers and geographic metadata catalogs. Section 3 discusses how to generate meaningful metadata from gazetteer entries. Section 4 expands the discussion to federations of gazetteers and geographic metadata catalogs. In special, it discusses how to consolidate thesauri from different gazetteers. Finally, Section 5 contains the conclusions.

2 GAZETTEERS AND CATALOGS

This section briefly summarizes basic concepts pertaining to gazetteers and geographic metadata catalogs, and lists additional references to related work.

A *feature* is an abstraction of a real world phenomenon and a *geographic feature* is a feature associated with a location relative to the Earth. In the familiar Computer Science jargon, a (geographic) feature is an object with a special attribute that describes the object's location on the Earth surface, using a given *coordinate (geo)reference system* (CRS).

A *gazetteer* is a list of geographic names, together with their geographic locations and other descriptive information. A *geographic name* is a proper name for a geographic place or feature, such as the City of Rio de Janeiro. We are interested in gazetteers that are available over the Web, such as the GNS and the ADL Gazetteer.

The GEOnet Names Server (GNS) (GNIS, 2005) provides access to the National Geospatial-Intelligence Agency (NGA) and the U.S. BGN database of foreign geographic names, containing about 4 million features with 5.5 million names.

The ADL Gazetteer (Hill et al., 1999) has approximately 5.9 million geographic entries classified according to the ADL Feature Type Thesaurus (FTT), a classification scheme that combines the vocabularies of the GNIS and the GNPS. Indeed, gazetteers often organize feature types as a thesaurus, which we will generically call a *feature type thesaurus*, by analogy with the ADL FTT.

A thesaurus (ISO-2788, 1986) is "the vocabulary of a controlled indexing language, formally organized so that a priori relationships between

concepts (for example as "broader" and "narrower") are made explicit." A thesaurus usually provides: a *preferred term*, defined as the term used consistently to represent a given concept; a *non-preferred term*, defined as the synonym or quasi-synonym of a preferred term; relationships between the terms, such as *narrower term*, indicating that a term *T* – the *narrower term* – refers to a concept which has a more specific meaning than another term *U* – the *broader term*.

An *information resource* is a (logical) entity that can be managed by a catalog service (Senkler et al., 2004). We will be particularly interested in geographic datasets in what follows. We will assume that geographic datasets will have at least a scale and a description of the area of the Earth's surface that the dataset covers. This description is often a rectangle or a parallelogram, whose vertices are defined by coordinates in a given coordinate reference system (CRS). The scale, the description of the area covered and the CRS used are treated as metadata of the dataset.

Several standards for metadata appear in the literature. The Content Standard for Digital Geospatial Metadata (CSDGM) is mandatory since 1994 for all geographical datasets in the US (FGDC, 2002). The CSDGM and European initiatives have been unified as the ISO 19115 metadata standard (ISO-19115, 2002).

A *metadata catalog* holds metadata describing information resources stored in data sources (Nebert, 2002). A catalog offers services to query and manage metadata, as well brokering services to retrieve resources, which is relayed to the data sources. Typically, a catalog does not store or manage the information resources themselves.

The OpenGIS Catalog Services (OCS) (Nebert, 2002) specification defines a collection of services, a minimal query language, and a core metadata schema, based on the ISO19115 - Geographic Information Metadata. The specification includes services to update the catalog, invoked by an application or by the catalog itself.

3 CENTRALIZED ARCHITECTURE

We first consider a centralized architecture that combines a single gazetteer and a single geographic metadata catalog. We show how to use the gazetteer thesaurus to classify geographic datasets and how to use gazetteer entries to describe geographic datasets.

The *Centralized Enhanced Metadata Catalog* has two major components. The *Catalog Manager* provides query and management services to

maintain a *Metadata Catalog*, describing information resources, and a set of relationships between catalog and gazetteer entries. The *Gazetteer Manager* provides query and management services to maintain a *Gazetteer*, describing geographic features, and a *Gazetteer Thesaurus*, with classification terms for geographic features.

Let GA be the gazetteer and assume that each entry E in GA , representing a geographic feature F , has a geo-referenced representation $geo(E)$ of (an approximation of) the location of F , and a $type(E)$ for E , whose value is a term taken out of a thesaurus $T[GA]$.

Let MC be a geographic metadata catalog and assume that each entry C in MC , representing a geographic dataset R , has a geo-referenced representation $geo(C)$ of (an approximation of) the area (on the Earth's surface) that R covers, a description $scale(C)$ of the scale of $geo(C)$. We argue that:

Q1. $T[GA]$ can be extended to also provide a classification for catalog entries, that is, to provide a feature type $type(C)$ for C .

Q2. A description $desc(C)$ for C can be generated by relating C to gazetteer entries in relevant ways.

Intuitively, assuming that a geographic dataset R covers an area of the Earth's surface, we may describe R by the collection of features that occur in the area. However, we filter out features whose type is inconsistent with the intended interpretation of R or whose extent is smaller than the scale of R . In the latter case, the type of the feature should give sufficient indication if the feature is compatible with the scale of R .

For example, let R be a dataset. If R should be interpreted as a political map, then R should represent cities and the political division of a given area. Depending on the scale of the map, only cities above a certain population should in fact be included. If R should be interpreted as a hydrographical map, we maintain rivers and creeks, but suppress cities. Moreover, if the map has a large scale, we maintain only rivers, and also suppress creeks. As a third example, if R is a satellite image of a given area, then R potentially represents all features occurring in the area. However, features whose extent is smaller than the resolution of R should be filtered out.

In general, we first suggest: (i) to classify geographic datasets also using $T[GA]$; (ii) to indicate the scale that is *compatible* with each term in $T[GA]$.

This is formalized as a function $s: T[GA] \rightarrow \mathbb{R}^*$ that maps each term of $T[GA]$ into a non-negative real number. For each $t \in T[GA]$, if $s(t) > 0$, we interpret $s(t) = n$ as indicating that all features of type

t are *compatible* with a scale $1:n$, or smaller (in the sense that they can be represented in that scale). If $s(t) = 0$, we interpret $s(t)$ as indicating that t can be used to classify geographic datasets.

As an example, consider the fragment of the ADL Feature Type Thesaurus shown in Figure 1(a), at the end of the paper. For example, we may then define:

$s(\text{"creek"}) = 10,000$ to indicate that features of type "creek" should only be represented in a scale $1:10,000$, or smaller;

$s(\text{"hydrographic features"}) = 0$ to indicate that the term "hydrographic features" can be used to classify geographic datasets.

In certain situations, the function $s: T[GA] \rightarrow \mathbb{R}^*$ may be partly computed from other attributes of the gazetteer thesaurus terms. For example, if each term t under "streams" has an attribute w indicating the width of the streams that are classified as t , then the value of w for t may be used to define $s(t)$.

As for question Q2, we first define that a gazetteer entry E , representing a geographic feature F , is *relevant* to a catalog entry C , representing an information resource R , iff:

$geo(E)$ and $geo(C)$ are related by one of the usual topological relationships – *touch*, *in*, *cross*, *overlap* (Clementine et al., 1993) – as well as others, such as *within*;

$type(E)$ is compatible with $scale(C)$.

We then define $desc(C)$ as the set of pairs (E, r) such that E is relevant to C and $geo(E)$ and $geo(C)$ are related by the topological relationship r .

Some gazetteers also include the concept of a *famous place*, such as a tourist attraction, an important city, etc. (GNIS, 2005). If the gazetteer adopted implements this concept, we may expand the notion of relevance previously defined to include famous places as a significant piece of information. For example, consider a satellite image of the City of Friburgo, which lies northwest of the City of Rio de Janeiro. Then, instead of just associating the image with the City of Friburgo, we may also indicate that the image covers an area northwest of the City of Rio de Janeiro, a famous place. Note that, when relating famous places and information resources, we may adopt directional relationships – *north-of*, *south-of*, *east-of* and *west-of* – or qualitative relationships, such as *near*.

As an example of how to generate $desc(C)$, suppose that we adopt the ADL Gazetteer and the ADL Feature Type Thesaurus. Consider the image fragment of the City of Rio de Janeiro, taken out of the Website "Brazil seen from Space" (Embrapa, 2004), shown in Figure 1(b).

This image will be processed as follows:

Extract the georeferencing parameters from the

information resource. In this case, the image fragment is consistent with a scale of 1:25,000 and has a bounding rectangle defined by the pair of coordinates ((43°15'W, 22° 52' 30"S), (43° 07' 30"W, 23°S)).

Assume that the user chooses to relate the image fragment with “hydrographic features”, a term of the ADL FTT that, in our running example, can be used to classify geographic datasets.

Since the ADL Gazetteer entries have no associated scale information, ignore it.

Access the ADL Gazetteer, using the parameters extracted in Step 1 and the ADL FTT terms under “hydrographic features”, the term selected in Step 2. The query returns 9 entries, among which the first 3 are:

- a. *Feature*(“Rodrigo de Freitas, Lagoa - Brazil”, lakes, within)
- b. *Feature*(“Comprido, Rio - Brazil”, streams, within)
- c. *Feature*(“Maracana, Rio - Brazil, streams, within)

Store the result of the query as a description of the information resource, that is, as a list of pairs (N,r) , where N is a geographic feature returned in Step 4 and r is the topological relationship between the image and N (in this case, r is “within”).

This brief example illustrates some of the basic ideas of the paper. First, the use of the gazetteer thesaurus to also classify geographic datasets precludes the adoption of a second classification scheme, such as the ISO19115 Topic Categories (ISO-19115, 2002). This approach simplifies mediating access to multiple catalogs and gazetteers, as discussed in Section 4. However, it requires defining the compatibility function $s: T[GA] \rightarrow R^*$.

Second, a useful description of a geographic information resource R can be created as a list of pairs (N,r) , where N is a geographic feature and r is the topological relationship between R and N , obtained by querying the gazetteer. In addition, the list contains only features whose type is compatible with the scale of R .

4 FEDERATED ARCHITECTURE

The discussion in Section 3 assumed a single gazetteer, with a geographic feature type thesaurus, and a single catalog. However, as pointed out in the introduction, a federation of gazetteers and geographic metadata catalogs, supported by mediator, is a more realistic architecture. Such mediators will need a tool to align different gazetteer thesauri that, according to the discussion in Section

3, are used to classify both gazetteer and catalog entries.

More precisely, let G and H be two gazetteers. Assume that they classify features using two thesauri, T and U , respectively. Suppose that we adopt the schema of G as the mediated schema, but we allow changing T to accommodate terms in U that have no counterpart in T .

This means defining a function $reclass:U \rightarrow V$ that maps terms in U into terms of a new thesaurus V , created from T and U . If $reclass(t)=u$ then we say that t is the *reclassification* of u . In the rest of this section, we only analyze two cases of this sub-problem, for reasons of brevity.

Suppose first that G and H contain entries that represent disjoint sets of features, and that T and U represent disjoint sets of concepts. Albeit simple, this is a common scenario.

We first *graft* U into T , using a term p of T as *pivot*, that is, we add the root u of U as a new narrow term of a p . This operation creates a new thesaurus, denoted $T[p, U]$. Now, when the mediator accesses entries in H , it will not change their type, that is, $reclass:U \rightarrow T[p, U]$ is the identity function. However, note that the grafting operation requires user intervention, since there is no failsafe way to automatically identify p by observing just the terms in T and U , and their definitions.

For example, let H be the list of real state assets of a large company, classified according to a thesaurus U . Assume that the company operates in Brazil, Venezuela and Argentine. Suppose that G is a copy of the ADL Gazetteer, restricted to these three countries. Then, to access the company’s assets in H , using the ADL Gazetteer schema, we first add the root of U as a new narrow term of “manmade features”, a term of the ADL Feature Type Thesaurus (FTT), on the grounds that the company’s assets are neither listed in G , nor they can be classified with the terms found in the ADL FTT. The result of the alignment process will be a thesaurus that contains the ADL Feature Type Thesaurus entries plus the entries in U .

Suppose now that G and H represent non-disjoint sets of features, and that they have thesauri that represent non-disjoint sets of concepts. This is a complex, but not uncommon scenario, which occurs when the mediator wants to access both G and H .

For brevity, we consider only the case where T , the thesaurus of G , will remain unchanged, which means that the range of $reclass$ is T . We discuss how to use a gazetteer sampling technique that takes advantage of geo-referencing to avoid the pitfalls of syntactical alignment.

We first define a relationship $Ident \subseteq G \times H$ such that, for any $(E,F) \in G \times H$, we have that $(E,F) \in Ident$ iff E and F denote the same (real-world) feature.

Note that *Ident* is not total in G or H , since G may have entries that do not correspond to features represented in H , and vice-versa.

There are several possibilities to compute *Ident* from the attributes of the gazetteers entries. For example, suppose that both G and H associate each entry with the centroid of the location of the feature the entry represents. Then, we may define that $(E,F) \in \text{Ident}$ iff E and F have the same centroid (after conversion to a common coordinate reference system), under a given margin of error. This approach is not failsafe since two different features may have the same centroid, by coincidence, or because the scale is not precise enough. Note that we could have used any other representation or approximation of the features' locations, such as bounding boxes, for that matter.

We access both gazetteers to create a *partial alignment* relation $P \subseteq G \times T \times H \times U$ such that $(E,t,F,u) \in P$ iff $(E,F) \in \text{Ident}$ and t and u are the types of E and F , respectively. Note that t is a term of T and u is a term of U . That is, P contains all pairs of entries from G and H that denote the same feature, together with their classifications from T and U .

For each term $u \in P[U]$, the projection of P onto U , we define $\text{reclass}(u) = v$ iff $v \in T$ is the least common ancestor of all terms $t \in T$ such that there is $(E,t,F,u) \in P$, for some $E \in G$ and $F \in H$. Intuitively, T might have a richer classification for the features that U classifies as u .

Since the reclassification has to be automatic, we can only choose one term v of T to map u into. We then choose the least general term that is consistent with the current entries in G and H . An interesting degenerated case is when, for any $(E,t,F,u) \in P$ and $(E',t',F',u) \in P$, we have $t=t'$, in which case $\text{reclass}(u) = t$. Intuitively, as currently observed, G and H consistently classify entries as u and t , respectively.

For each term $u \in U$ that does not occur in $P[U]$, we define $\text{reclass}(u) = \text{undefined}$. User intervention is then required to redefine $\text{ident}(u)$, if undefined is unacceptable.

For example, suppose that G is the ADL Gazetteer and H is the GEOnet Names Server (GNIS, 2005) and that we query both gazetteers for "Rio de Janeiro". The GEOnet Names Server will return two entries, representing cities in Colombia, not represented in the ADL Gazetteer. By contrast, the ADL Gazetteer will return two entries, representing streams in Brazil, not represented in the GEOnet Names Server.

Since both gazetteers return the centroid, we define $(E,F) \in \text{Ident}$ iff E and F have the same centroid. Based on the data returned on the definition of *Ident*, Table 1 shows the partial alignment relation and Table 2 exhibits the partial

definition of the function *reclass*.

Based on this fragment of *reclass*, the mediator may then access entries in the GEOnet Names Server classified as "SMTI", "PPLA", "PPL", "MT", "ADMI" and "HLLS" and reclassify them according to the ADL FTT.

5 CONCLUSIONS

We described in this paper an approach to partly automate metadata generation and argued that gazetteers will have to be expanded to include information about scale, and that catalog metadata schemes will have to be carefully designed to facilitate integration with gazetteers.

In particular, to address the problem of gazetteer thesauri alignment, we introduced a gazetteer sampling technique that takes advantage of geo-referencing to avoid the pitfalls of aligning the thesauri terms based on syntactical proximity.

We are currently working on the implementation of the Gazetteer Aligning Tool, which includes a component that semi-automates thesauri alignment, using the gazetteer sampling technique.

ACKNOWLEDGEMENTS

This work is partially supported by CNPq under grant 551241/05-5.

REFERENCES

- Clementini, E.; di Felice, P.; van Oosterom, P., 1993. A Small Set of Formal Topological Relationships Suitable for End-User Interaction. In: Abel, D.; Ooi, B. C., eds., SSD '93: Lecture Notes in Computer Science, v. 692: New York, NY, Springer-Verlag, p. 277-295.
- Embrapa, 2004. Embrapa Monitoramento por Satélite. *Brazil seen from Space*. <http://www.cdbrasil.cnpm.embrapa.br/>
- FGDC, 2002. Federal Geographic Data Committee <http://www.fgdc.gov>
- Fu, G.; Abdelmoty, A. I.; Jones, C. B., 2003. Design of a Geographical Ontology, D5 3101, Public Deliverable SPIRIT Project. Available at: http://www.geospirit.org/publications/SPIRIT_WP3_D5.pdf
- GNIS, 2005. Geographic Names Information System, U.S. Department of the Interior, U.S. Geological Survey, Reston, USA. Available at: <http://geonames.usgs.gov/index.html>
- Hill, L.; Frew, J.; Zheng, Q., 1999. Geographic Names:

The Implementation of a Gazetteer in a Georeferenced Digital Library. In: D-Lib, January-1999. Available at: <http://www.dlib.org/dlib/january99/hill/01hill.html>.

ISO-2788, 1986. *Documentation -- Guidelines for the development of monolingual thesauri*, International Standard ISO-2788, Second edition -- 1986-11-15

ISO-19115, 2002. ISO 19115 Geographic information – Metadata. Draft International Standard.

Klien, E.; Lutz, M., 2005. The Role of Spatial Relations in Automating the Semantic Annotation of Geodata. In Conf. on Spatial Information Theory.

Nebert, D., 2002. *Catalog Services Specification*, Version 1.1.1, OpenGIS® Implementation Specification, Open GIS Consortium, Inc.

Senkler, K.; Voges, U.; Remke, A., 2004. An ISO 19115/19119 Profile for OGC Catalogue Services CSW 2.0. In 10th EC GI & GIS Workshop, ESDI State of the Art, Warsaw, Poland, 23-25 June 2004.

Souza, L.A. et al., 2005. The Role of Gazetteers in Geographic Knowledge Discovery on the Web. In Proc. 3rd Latin American Web Congress. Buenos Aires, Argentina (Oct. 2005).

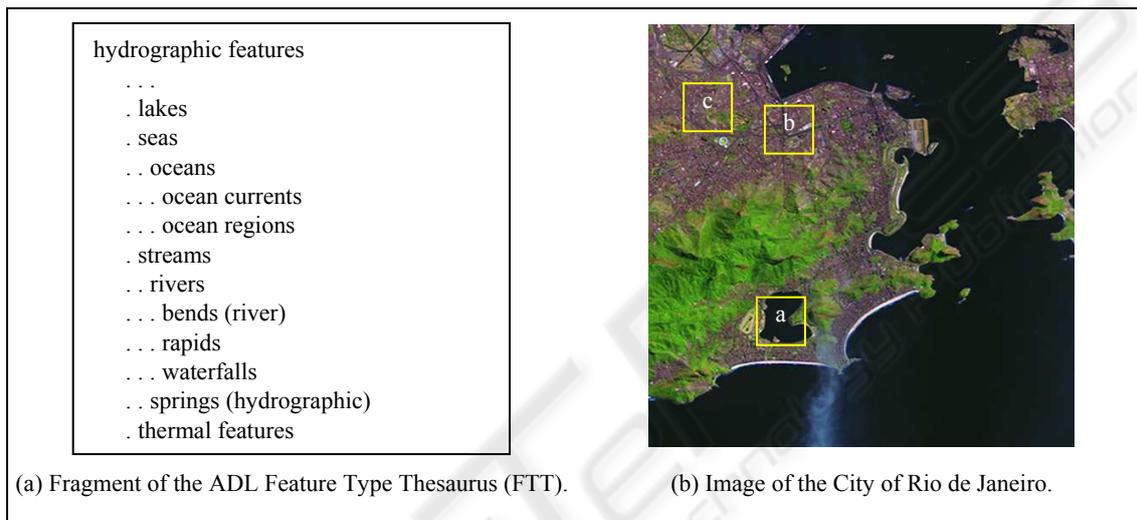


Figure 1: A fragment of the ADL Feature Type Thesaurus and an image of the City of Rio de Janeiro.

Table 1: Partial alignment of the ADL Gazetteer and the GEONet Names Server based on the LAT/LONG returned.

| ADL Gazetteer | | | GEONet Names Server | | |
|---------------|--|----------------------|---------------------|---------------------------|-------|
| # | Name | Class | # | Name | Class |
| 1 | Rio de Janeiro, Igarape - Acre – Brazil | streams | 6 | Rio de Janeiro, Igarapé | STMI |
| 2 | Rio de Janeiro – Brazil | populated places | 4 | Rio de Janeiro | PPLA |
| 4 | Rio de Janeiro - Loreto - Peru | populated places | 5 | Río de Janeiro | PPL |
| 5 | Rio de Janeiro, Serra - Paraíba – Brazil | mountains | 2 | Rio de Janeiro, Serra | MT |
| 6 | Rio de Janeiro, Estado do - Brazil | administrative areas | 3 | Rio de Janeiro, Estado do | ADM1 |
| 8 | Rio de Janeiro, Serra do - Brazil | mountains | 1 | Rio de Janeiro, Serra do | HLLS |

Table 2: Partial definition of the function *reclass*.

| GEONet Names Server Class | ADL Gazetteer Class |
|---------------------------|----------------------|
| <i>reclass</i> (STMI) | streams |
| <i>reclass</i> (PPLA) | populated places |
| <i>reclass</i> (PPL) | populated places |
| <i>reclass</i> (MT) | mountains |
| <i>reclass</i> (ADM1) | administrative areas |
| <i>reclass</i> (HLLS) | mountains |