Merging Local and Global Gazetteers

Øyvind Vestavik and Ingeborg T. Sølvberg

Dept. of Computer and Information Science
Norwegian University of Science and Technology
{oyvindve,ingeborg}@idi.ntnu.no

Abstract. Most gazetteers with a global scope contain few local names, and gazetteers with a local scope mostly do not contain foreign names. However, people often use both local and foreign place names in a discourse. We describe some of the challenges in mapping local and global gazetteers that serve different needs and hence may have different structure, granularity and coverage. We pay special attention to the problem of identifying duplicate place descriptions in such registries.

Keywords: Geographic Information Retrieval, Gazetteers, Gazetteer Merging, Standards, Controlled Vocabularies.

People tend to use both local and foreign names in a discourse and in documents. In order to index such documents based on their geographic *aboutness* we need gazetteers with both detailed local coverage and foreign names. We describe a schema level and instance level mapping of gazetteer data from the Alexandria Digital Library Gazetteer (ADL GAZ) and Sentralt Stedsnavnregister (SSR). The findings of this project provide insights into some of the general problems of mapping gazetteers and the resulting gazetteer will be used for indexing Norwegian newspaper articles.

SSR [1] is the official registry of Norwegian place names. It contains information about approx. 600 000 current place names in Norway and is built up around 4 conceptual objects/entities (*Locations*, *Place Names*, *Spelling Variations* and *Occurences* (on maps, road signs etc)). Each entity has a set of attributes and the entities are organized in a tree structure (see figure 1, left side). Each tree describes one place. The ADL gazetteer [2] contains descriptions of approx. 4.4 mill. places and approx. 5.9 mill names used as labels for these places. ADL Standard Reports (See fig. 1, right side) describe a place using a set of key attributes of the place, including *Names*, *Footprints* (location(s)), *Relationships* (to other places) and *Classes* (type of place).

Schema level mapping consists of identifying attributes in the two models that contain the same or comparable information. Schema level mappings are shown in figure 1. Instance level mapping deals with identifying pairs of ADL entries and SSR records that describe the same place (duplicate detection). 4 indicators are used to calculate a score for the probability that a given ADL Standard Report and a given SSR record describe the same place: 1) Co-location or near co-location. 2) Number of shared place names. 3) Shared or similar feature types

D.H.-L. Goh et al. (Eds.): ICADL 2007, LNCS 4822, pp. 495–496, 2007.

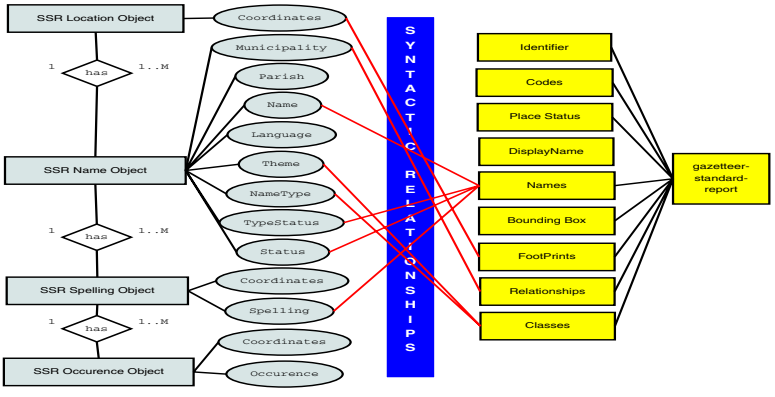

Fig. 1. Schema level mappings between the ADL and SSR data models

(kind of place eg. city, school, mountain etc). The similarity is measured as graph distance in a weighted graph implemented over the two vocabularies describing the feature type of the places described. 4) Part of same larger place. Valid mappings are selected based on best scores and mappings with low scores are discarded.

Initial investigation has revealed some challenges to our approach: 1) Comparing names is difficult for place names containing language specific characters because these names have been transcribed in ADL GAZ. The names *Røa* and *Roa* are for instance both represented as *Roa*. 2) Often, the two gazetteers agrees on only a subset of names for a place. 3) Often, ADL GAZ and SSR do not agree on which names should be considered primary and what names are current. 4) The vocabularies used to describe the feature type of a place are to some extent culturally biased, the set of concepts only partially overlaps, the granularity of similar concepts often differ and there is often only a partial overlap in semantics between similar concepts. 5) ADL GAZ describes a place as being a part of a county whereas SSR describe a place as being part of a municipality. 6) Coordinates describing the location of a place might vary considerably, even for identical places.

References

1. The Norwegian Gazetteer / Sentralt Stedsnavnregister (norwegian only), `http://www.statkart.no/standard/sosi/html_34/navn/navn.htm`
2. Alexandria Digital Library Gazetteer.: Map and Imagery Lab, Davidson Library, Santa Barbara CA, University of California, Santa Barbara (1999), `http://www.alexandria.ucsb.edu/gazetteer`