Dutch historical toponyms in the Semantic Web

Preprint version of the paper to be presented at the
Workshop "Population reconstruction", 19-21 February 2014, IISH - Amsterdam

Ivo Zandhuis, Ivo Zandhuis Research & Consultancy
Menno den Engelse, Islands of Meaning
Edward Mac Gillavry, Webmapper

In early 2013, the website gemeentegeschiedenis.nl *was launched. The website presents a uniquely identifiable web page for every municipality in the Netherlands since 1812. Each of these web pages provides internal relations to official spelling alternatives of their names, to toponyms of settlements within these municipalities, between former and current municipalities, and presents maps of all the geographical changes to the official boundaries of these municipalities. Furthermore, these web pages provide external relations to Wikipedia entries.*

Before the launch of this comprehensive, online resource, there have been various lists, "gazetteers", available for standardising Dutch historical toponyms, but these were not available online and were scattered across various cultural heritage institutes. Thus, an important aspect of this initiative is its ability to interrelate these diverse, de facto standards:

- *CBS code: assigned by Statistics Netherlands, the Dutch national statistical office*
- *Kloeke code: assigned by Meertens Institute for research and documentation of Dutch language and culture [kloeke1926].*
- *Amsterdam code: assigned by Van der Meer and Boonstra as part of the Historical Geographic Information System project [meer2006].*
- *GeoNames code: an online, crowd-sourced database of over 10 million toponyms*

Being able to interrelate these standards we introduce the historical dimension into the international standardisation of contemporary geographical names, combining the best of both worlds. On the one hand, the website aims to present information about the historical development of municipalities for a broader public. On the other hand, the website provides geographical services for cultural heritage institutes and research in the humanities that deal with historical toponyms. In this paper, we describe these online geographical services in the context of a few use cases.

In the first use case, we show how the website is able to assist historians who aim to standardise the historical toponyms in their data set. In the second use case, we show how the website can play a role in more elaborate data entry projects and is able to standardise historical toponyms semi-automatically. In the final use case we show how one coding standard for historical toponyms can be converted into another and how data related to municipalities or settlements can be mapped geographically using thematic mapping techniques in order to reveal spatio-temporal patterns in data sets.

Prior work

The importance of the standardisation of historical toponyms is broadly recognised. There are various online resources about historical placenames. The community-built gazetteer of ancient places Pleiades (http://pleiades.stoa.org/) gives scholars, students and enthusiasts worldwide the ability to use, create, share, and map historical geographic information primarily about the ancient world.

The Getty Thesaurus of Geographic Names (http://www.getty.edu/research/tools/vocabularies/tgn/), generally referred to as “TGN”, is a structured vocabulary currently containing over 2 million names and other information about places. These places include not only administrative political entities (e.g., cities, nations), but also physical features (e.g., mountains, rivers). Both current and historical places are included. The temporal coverage of the TGN ranges from prehistory to the present and the scope is global. While many records in TGN include geographic coordinates, these coordinates are approximate and are intended for reference only. These geographic coordinates in TGN typically represent a single point, corresponding to a point in or near the center of the inhabited place, political entity, or physical feature.

A widely-adopted open dataset for geographical entities is the Geonames Geographical Database (http://www.geonames.org). The data set has been created through crowdsourcing and contains all kinds of geographical entities, each with a unique identifier. One category of entities is the “administrative division”. The second order administrative division is equivalent to the municipality in the Amsterdam Code. However, there are no historical municipalities available in the data set.

Closely related to Geonames is Wikipedia (http://www.wikipedia.org). There are numerous web pages for geographical entities, amongst others municipalities. Due to its nature, Wikipedia does not provide any indication about the completeness and accuracy of the contents. Most data in Wikipedia is also available as Linked Data via the Semantic Web portal Dbpedia (http://www.dbpedia.org). For a broader discussion of resources for historical toponyms, see [southall2011].

Gemeentegeschiedenis.nl has a distinctive character compared to the various online resources of historical placenames described here. First of all, the primary geographic entity is the second order administrative area, i.e. “municipality” in the Netherlands, instead of settlements as is the case for the majority of examples provided in this paragraph. Hence, the geographic representation of these administrative areas used in our service platform is a two-dimensional polygon instead of a one-dimensional point location.

Secondly, the scope of geographic entities within our service platform is strictly defined. Municipalities were instituted by law and geographic changes to their boundaries, mergers and separations all require changes in the law. Thus, the contents are more or less complete already and have been compiled based on datasets that have been released as open data or that have been provided by various parties to this project. Changes only happen annually, sometimes twice a year in some cases. Other resources such as Pleiades, Wikipedia, and Geonames all require large-scale community involvement to grow their contents, because settlements are much more loosely defined and the rise and fall often goes unrecorded. Nevertheless, it is already obvious from the preliminary feedback on Gemeentegeschiedenis.nl that this service platform will also benefit from community contributions, for example to improve the geographic accuracy of the delineation of the municipalities.

Finally, Gemeentegeschiedenis.nl holds all consecutive stages of the second order administrative subdivisions, i.e. the Dutch municipalities, from 1812 until now without any interruption. In contrast, the history of settlements is not governed by law and their rise and fall often goes unrecorded. Therefore the temporal continuity of the contents of other online resources for historical toponyms is patchy in comparison to Gemeentegeschiedenis.nl.

Geographical standardisation with a historical dimension

Administrative areas change over time. The name “Abcoude” in use between 1812 and 1818 refers to another municipality with different boundaries than the name “Abcoude” in use 1941 and 2011. The current municipality named “Almelo” is a merger of the town (Stad_Almelo) and its surroundings (Ambt_Almelo) that used to be two municipalities during the most part of the 19th century. This geographic situation can be observed in five other municipalities located in the eastern part of the Netherlands[1].

The comparison of historical data related to municipalities over time and space is a complex matter due to the heterogeneity of historical toponyms. Historical documents that describe people, economic activity or political votes contain references to the contemporary administrative areas. In many instances, the place of birth (actually, the municipality where the birth was registered) recorded in historical documents presents a peculiar anachronism as it reflects the administrative subdivision at the time of birth, not at the time at which the place of birth was recorded.

Digitising these historical sources, historians not only have to copy the geographical description as precisely as possible as it was recorded in the original source, but also have to interpret the original and provide a standardised spelling of this geographical description. At a later stage this facilitates statistical analysis or spatial analysis using a Geographic Information System (GIS). Within the context of the dataset, this standardised form must be unambiguous and unique. Drawing this standardised spelling from a widely adopted thesaurus, historians are then able to combine and contrast their own datasets with other datasets that use the same thesaurus.

The main source for Gemeentegeschiedenis.nl is the thesaurus of Dutch municipalities, generally referred to as the Amsterdam Code [meer2006]. Although previously published in a book and available under an open license as a downloadable PDF file, there was no online resource in a machine-readable format. In the Amsterdam Code, identifiers are reused over time: a municipality that only changed name therefor kept the same identifier. When two municipalities were combined into one, the Amsterdam Code of the municipality with the largest number of inhabitants was inherited by the newly formed municipality. This implies that the municipalities are comparable through time, but the codes are not unique. They are only unique in combination with a timestamp indicating the period of the existence of a municipality.

Each of the municipalities on Gemeentegeschiedenis.nl are further identified by their CBS code. This code is assigned and maintained by Statistics Netherlands, the Dutch national statistical office. The historical dimension is limited, but the code is used in most of the currently available datasets, which enables the temporal comparison between the current situation and the past. Since the CBS codes do not cover municipalities until 1830, the Amsterdam Code encompasses a broader, temporal range as it identifies municipalities from as early as 1812.

The Kloeke Code for settlements, assigned by Meertens Institute, is used in the research and documentation of geographical patterns in Dutch language and culture [kloeke1926]. Since these settlements are no official administrative areas, disambiguation is ensured by combining the Kloeke code identifier with the coordinates of the settlements. Furthermore, Kloeke codes are not available for all settlements in The Netherlands.

1. Gemeentegeschiedenis.nl

On Gemeentegeschiedenis.nl, each of the Dutch municipalities has its own web page. Each of these web pages contains a set of maps that do not only reflect the changes in the delineation of the

[1] A sixth situation (Ambt_Montfort) is in the southern part of the Netherlands and is a modern use of this old-fashioned name.

municipalities, but also the mergers of municipalities over time [boonstra1992]. Furthermore, all settlements that are located within the boundaries of a municipality are listed together with links to other online resources that describe that particular municipality or settlement. Visitors to the website can learn about the history of the Dutch municipalities and researchers can obtain information to assist them in standardising the administrative areas in their datasets.

Each of these web pages is available at a short and readable URL that also acts as a Unique Resource Identifier, a URI [rfc2396]. For example the municipality of Abcoude between 1941 en 2011 is available at and is identified with the URI "http://www.gemeentegeschiedenis.nl/gemeentenaam/Abcoude_2". Using a process referred to as "content negotiation" a client application obtain the information available in the appropriate data format, e.g. RDF-XML, JSON or – in case of a web browser, HTML. Some examples of URIs:

http://www.gemeentegeschiedenis.nl/gemeentenaam/Schoten
http://www.gemeentegeschiedenis.nl/gemeentenaam/rdfxml/Schoten

http://www.gemeentegeschiedenis.nl/amco/10382
http://www.gemeentegeschiedenis.nl/cbscode/1173

There are also URIs that provide a list of municipalities in one province, for instance:
http://www.gemeentegeschiedenis.nl/provincie/Noord-Holland
http://www.gemeentegeschiedenis.nl/provincie/rdfxml/Noord-Holland
http://www.gemeentegeschiedenis.nl/provincie/json/Noord-Holland
http://www.gemeentegeschiedenis.nl/provincie/csv/Noord-Holland

We envision that the historical database containing toponyms and the geographic representation of Dutch municipalities with various types and links between the entries, should be available to the general public in a variety of ways. At the moment, the website provides a simple search on the names of municipalities. Historians can use this to standardise their dataset, but have to search for every entry one by one. Therefore, extra functionality is needed to improve the usability and applicability of the website.

We distinguish two types of future functionality: services that assist historians to *standardise* the geographical entities in their datasets and services that assist them to *process* their standardised datasets. The remainder of this paper is used to explain and describe these services.

2. Standardisation services

In order to obtain a dataset with standardised forms of the original geographical entity in the historical document, historians should add an additional field in their records. The original records contain the contemporary and authentic description of the geographical name that needs standardising. In the additional field, Gemeentegeschiedenis.nl will store the interpretation of this original description by means of a normalised toponym or code. The main advantage of using a service platform approach is that new knowledge about toponyms and relations is continuously incorporated into the database. Therefore, the geocoding success rates of the service platform will improve over time without additional efforts on the part of the historians.

2.1 Historians standardise the geographical descriptions in their datasets

An important service entry of Gemeentegeschiedenis.nl is the search interface. Users can enter a search term and various names of municipalities and settlements are returned. If any of the codes

(Amsterdamse Code, Kloeke, CBS) is provided instead, the matching municipality or settlement is returned.

For every municipality, the search returns the following information

1. name
2. URI
3. Amsterdam Code
4. CBS code
5. Geoname identifier
6. Province name
7. Geographic coordinates of its delineation

For every settlement, the search result returns the following information:

1. name
2. all municipalities it was located in, including the time period, URI and codes
3. Province name
4. Country
5. Geoname identifier
6. Geographic coordinates of the location

The user interface provides widgets that enable historians to narrow down the search result set by selecting a year, a region, and/or a standard. Thus, the responsibility remains with the historians for the correct interpretation of the historical source and for the selection of the right standard and standardised form of the geographical description.

2.2 Indexing or large-scale data entry projects

Obviously, adding a standardised geographical code or name can also be performed at data entry. In that case it would be convenient to incorporate the search into the user interface of the data entry software. In The Netherlands an interesting example where this service could come in handy is http://www.velehanden.nl. On this crowd-sourcing platform, archival services can start a data entry project on a specific archival source.

To implement such a service we need to introduce an *Application Programming Interface (API)*. Using this API a request can be sent to Gemeentegeschiedenis.nl that subsequently returns a result set in a specified format (JSON or RDF-XML). The result set can be used to build a selection option in the data entry interface.

The input-variables of this function are:

1. q; mandatory: the search-term
2. std; the requested standard (name (default), amco, cbscode, kloeke, geonames)
3. format; the requested dataformat (json (default), rdfxml)
4. date; the period in which the name must exist (available values: 1812-now)

An example request URL could look like this:

http://www.gemeentegeschiedenis.nl/q=abcoude&std=cbscode&format=rdfxml

A more general service would be a SPARQL-endpoint. A SPARQL-endpoint can be used to formulate a specific query, that is not provided by the API. That way the current crude Semantic Web functionality of RDF-XML export is extended.

2.3 Automatic standardisation

Additionally, the API can be used to provide automatic standardisation. Therefore, the result set is organised according to a certain relevance function: the standardised form with the highest probability will be presented as first entry in the result set. The automatic detection can be improved by two extra variables, that provide information about the context of the source:

5. place: the location where the source was created
6. target area: the area where the result must be located

Results are selected that are located in the requested geographical target area and existing in the requested time period, prior to the end date of the period in which the source was created. This last variable can be requested in date.

The order in which the result set is presented could be determined as follows. Of course, the name with the smallest amount of differences with the search-term is highest on the result-list. This difference can be expressed in the Levenshtein distance [levenshtein1966]. When more results are delivered, the geographical unit that has the smallest geographic distance to the place of origin, is higher on the result list.

An extra service could enable historians to upload a CSV file. As they indicate which column contains the geographical name to be standardised and the standard in which it is to be standardised, the service platform returns a CSV file that now contains an extra column that holds the automatically generated standardisation. In this case, only the result with the highest probability is returned.

3. Processing services

Data sets that contain standardised geographic names can be processed more consistently. For example, one could create a service that converts the standardised geographic names in a dataset into another standard. Furthermore, historians would want to visualise their data set on a map in order to identify geographic patterns.

3.1 Converting from one standard into another

Combining various datasets with differently standardised geographical names can be very complex. A service enabling historians to convert from one standard into another, could be very useful. Historians would upload a CSV file and indicate the column that contains the standardised name and select the standard and period in which it is to be converted. A new CSV file is automatically generated that contains an extra column holding the additional standardised value.

Historians have to be aware that the different standards are not equivalent: a settlement that is converted into a municipality results in the loss of precision of the geographical location. However, the service platform enables historians to convert whatever they like.

3.2 Enrichment with geographic coordinates

For visualization purposes the standard codes can be converted into geographical coordinates: in a GIS these coordinates can be used to draw a settlement or municipality on a map, together with the information related tot he settlement or municipality. For every geographic entity the service platform is able to provide the geographic coordinates. Uploading a file containing multiple geographic entities, the standard forms of the toponymns are subsequently enriched with the coordinates of their geographic representation in batch.

3.3 Visualizing information

One of the many functions a GIS can perform is the visualisation of geographic datasets. However, the operation of a GIS system can be fairly cumbersome for the uninitiated. A future functionality of the service platform to automatically generate a map based on historical data sets that have been uploaded could be very useful, either to draw the main conclusions in a historical research project, or to assist in the decision whether to involve professional mapping expertise.

Historians would upload their data sets containing both the unique identifiers for the geographic entities, i.e. the Amsterdam Code, CBS code or Kloeke code, and the corresponding attribute values to be represented on the map. Once the specific year for the dataset has been selected, the values will be matched to the contemporary geographic delineations of the municipalities. Depending on the nature of the values the appropriate map type is created to visualise the data set geographically.

In case the dataset contains nominal values, a chorochromatic map is created. Municipalities that have the same nominal value, are filled with the same colour. An example is a map coloured-coded according to the largest political party per municipality. If the data set contains ordinal or relative values per municipality, a choropleth map is created. Municipalities with increasing values are filled with colours of increasing saturation or lightness. An example is a map that shows the relative growth of the population per municipality. Finally, if the data set contains absolute values, a proportional symbol map is created. An example is a map that shows increasing number of conscripts per municipality using increasing sizes of bars, squares or circles.

The resulting map could be easily downloaded in SVG or PDF formats for inclusion in scientific publications or presentations. Interactive versions of the maps could be embedded in third-party websites using a JavaScript-based mapping API that would request map images from our platform.

4 Future work

This paper describes the functionality we envision to be useful for historians with interest in a geographical component in their research. Hopefully, these services will be implemented for our service platform in the near future. Another area of future growth of our service platform lies in the addition of new names of settlements and their spellings variants. Of particular interest are Dutch exonyms for foreign settlements – e.g. the Dutch toponym “Jarmuiden” refers to the British port of Yarmouth – and foreign exonyms for Dutch settlements, e.g. French toponym “Nimegue” refers to the Dutch city “Nijmegen”. This enables historians to use historical sources from all around the world to combine information about the Netherlands.

We have already learned that people are very eager to help improving the information on Gemeentegeschiedenis.nl is incorrect, particularly the maps. Additional functionality for visitors to provide feedback on the information, or even to help us correct it, could prove to be very helpful. New information can be added as well. New names, with different spellings could be provided by historians doing research in a particular source, with unknown names.

The service platform could also be extended to smaller geographic entities, like neighborhoods, streets or even individual houses. That way, we gradually create a historical gazetteer or geocoder. We hope and aim that these services will enable historians to interpret, enrich and visualise their historical data in new and innovative ways. Moreover, the standardisation services provide them with a means to actually publish their historical data as linked open data.

References

References; url checked on 11 october 2013

[rfc2396] Berners-Lee, T. et al. (1998). *Uniform Resource Identifiers (URI): Generic Syntax*, http://www.ietf.org/rfc/rfc2396.txt

[kloeke1926] Kloeke, G.G., De totstandkoming van de “kaart van het Nederlandsche taalgebeide ten behoeve van het dialectgeografisch onderzoek” in: *Handleiding bij het Noord- en Zuid-Nederlandsch Dialectonderzoek*, pp. 57-65. http://www.meertens.knaw.nl/projecten/mand/LITkloeketotstandkoming.html

[meer2006] Meer, A. v.d. & Boonstra, O. (2006), *Repertorium van Nederlandse gemeenten 1812-2006*. Den Haag: DANS, Data Archiving and Networked Services, http://www.dans.knaw.nl/content/categorieen/publicaties/dans-data-guide-2

[boonstra1992] Boonstra, O., NLKAART. A dynamic map of the Netherlands, 1830-1980, In: J. Smets, ed., *Histoire et Informatique V. Montpellier 1992*, 315-324. The shape files of this research are deposited at DANS: https://easy.dans.knaw.nl/ui/datasets/id/easy-dataset:43063

[levenshtein1966] Levenshtein VI (1966). "Binary codes capable of correcting deletions, insertions, and reversals". *Soviet Physics Doklady* ***10****: 707–10.*

[southall2011] Southall H., R. Mostern & M.L. Berman (2011) On Historical Gazetteers. International Journal of Humanities and Arts Computing 5 (2), pp.127–145.