15 March 2021

English

United Nations Group of Experts on Geographical Names
2021 session
New York, 3 – 7 May 2021
Item 6(a) of the provisional agenda [*]
National and international standardization of geographical names: Names collection, office treatment, national authorities, features beyond a single sovereignty and international cooperation

Relevant resolutions: VII/9, VIII/6, VIII/9, VIII/10, IX/7

New National Geographical Names Archives Service

Submitted by Norway[**]

Summary:

The full report provides details on the planning, development and deployment of a new Norwegian national geographical names archives service. The service aims at collecting all known digital and digitized sources for geographical names in one single portal to facilitate the archiving of geographical names documentation. The service is built on open semantic principles using the International Committee for Documentation-Conceptual Reference Model ontology.[1] That will enable the service to exchange data with other data sets – geographical and non-geographical – through an application programme interface. The service currently exists in prototype form, with a limited number of data sets and only some 500,000 geographical name forms. It is envisaged that the service will contain in excess of 7 million name forms, of an estimated 2.5 million individual geographical names.

Historical forms and local pronunciation information often form the basis for determining the correct spelling of geographical names. The report gives a description of the use of the service in standardization matters and how it serves as a tool for regulators and the general public to source historical information about the spelling and pronunciation of Norwegian geographical names. It is envisaged that the service will make geographical name standardization decisions easier and improve transparency in decision-making.

One aspect that will need further legal investigation is a possible obstacle relating to the digital transformation of archives and collections resulting from the recent enforcement in Norway and the European Union of regulations relating to the General Data Protection Regulation. At worst, these regulations will not allow for an open and free exchange of geographical names information with regulators and the general public.

[*] GEGN.2/2021/1

[**] Prepared by Peder Gammeltoft, Norway.

[1] A tool for semantic information integration, it defines the underlying semantics of database schemata and document structures used in cultural heritage and museum documentation in terms of a formal ontology to enable semantic interoperability. See www.cidoc-crm.org/.

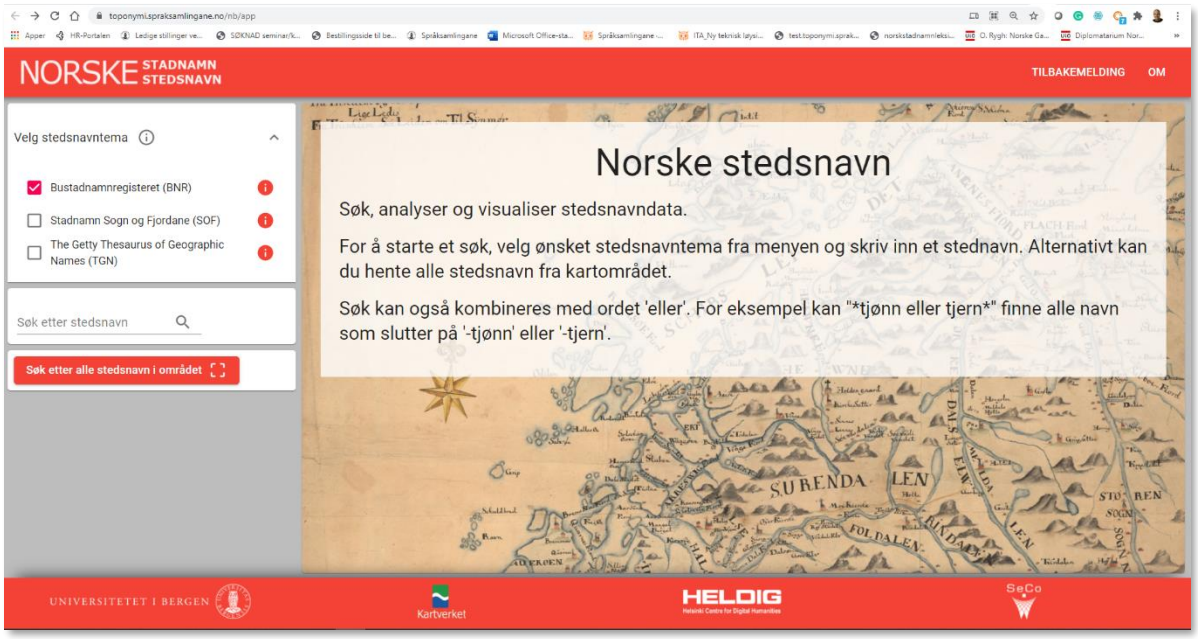
A New National Geographical Names Archives Service

The background

In countries where geographical names standardization is tied to written standard based on local pronunciation more than on traditional spelling, access to detailed source material is of vital importance. The needs of traditional toponymic research and geographical names standardization converge on this issue on the need for reliable and easily accessible sources. The Norwegian Place-Name Act states that the regulation of spelling of place-names must be founded in the local inherited pronunciation. In reality, however, other factors such as tradition and historicity also play a role in the standardization of geographical names. Therefore, it is necessary to be able to provide source material for the authorities in charge of standardization – with information about pronunciation, wherever possible, and of earlier spellings.

In Norway, where standardization activities is carried out both at local government level and within state authorities like the Norwegian Language Council and, not least, the Norwegian Mapping Authority, digital single-point access is vital in order to secure a balanced and symmetrical standardization process accross administrations. However, since Norway has a long tradition of de-centralized geographical names collection and decentral university and county archives, the notion of a single-point of access is difficult to achieve.

Figure 1. The portal *Norske stedsnavn* / *Norske stadnamn* (Norwegian Place-Names (NPN)).



With the transfer of the Norwegian Place-Name Archives (being part of the Norwegian Language Collections)[2] from University of Oslo to University of Bergen, the possibility emerged of establishing a new, national resource for geographical names, scalable to any size and capable of handling and presenting both own data as well as external resources through webservices. The result became *Norske stedsnavn* / *Norske stadnamn* (Norwegian Place-Names (NPN)), a national, web-semantic portal for querying and viewing sources for geographical names.[3] There is an estimated 2,5 million geographical names recorded in Norway – of which 1 million is registered by the Norwegian Mapping Agency.

[2] The Norwegian Language collections houses and maintains large linguistic collections. It is the largest single collection of data in the country related to lexicography (dictionaries), dialectology, terminology and place-names, or toponymy.

[3] https://toponymi.spraksamlingane.no/nb/app

The purpose of the NPN portal service is to be able to query and display geographical names from all digitally available sources. In the Norwegian Place-Name archive alone, there is in excess of 7 million source forms that all function as documentation in toponymic research and geographical names standardization. In order to ascertain that the source-form in question relates to the name of a particular locality, geolocation is a central element in modern place-name applications – and this is a central feature of the NPN service. Therefore, it goes without saying that a powerful and central means of presenting and visualizing place-name forms from historical sources is essential.

Technologies used

The NPN service is built on open semantic principles using the the International Committee for Documentation-Conceptual Reference Model ontology (CIDOC-CRM) ontology.[4] This will enable the service to exchange data with other datasets – geographical as well as non-geographical – through an Application Programme Interface (API). NPN portal is itself a GitHub fork of the Finnish place-name portal nimisampo.fi, which was launched in February 2019, and was built by the Semantic Computing Research Group at the Aalto University[5]. Nimisampo.fi[6] is also an example of the SampoUI framework, a framework for developing user interfaces for web semantic portals. This portal framework offers a workbench on top of geographical place-name datasets. It provides search, visualization, and analytical tools on top of spatial place-name datasets, in a modern packaging, using a staple of open-source components. It builds on top of open-source components, such as Apache-Jena Fuseki, the Leaflet.js library for maps and React for the framework.

Our interest in this service was sparked by the portal itself, but also that it was using the linked data stack, which is a special focus area for the University of Bergen Library. In addition, language data lend themselves well to linked data technologies, and geographical names are especially well-suited for web semantic data structures because of them being intangible as well as physical object at one and the same time.

Initial development phase

After setting up the Nimisampo service in our own test environment, we loaded datasets representing envisaged data-retrieval environments. Thus, we added one self-hosted dataset, Bustadnamnregisteret (Settlement Names Archive), one endpoint-derived self-hosted dataset, Stadnamn Sogn og Fjordane (Place-Names Sogn and Fjordane), both configured as individual services in an Apache-Jena Fuseki-server environment. The Bustadnamnregisteret dataset was mapped to RDF, while the Sogn and Fjordane dataset already is an existing published RDF dataset dump. Unfortunately,

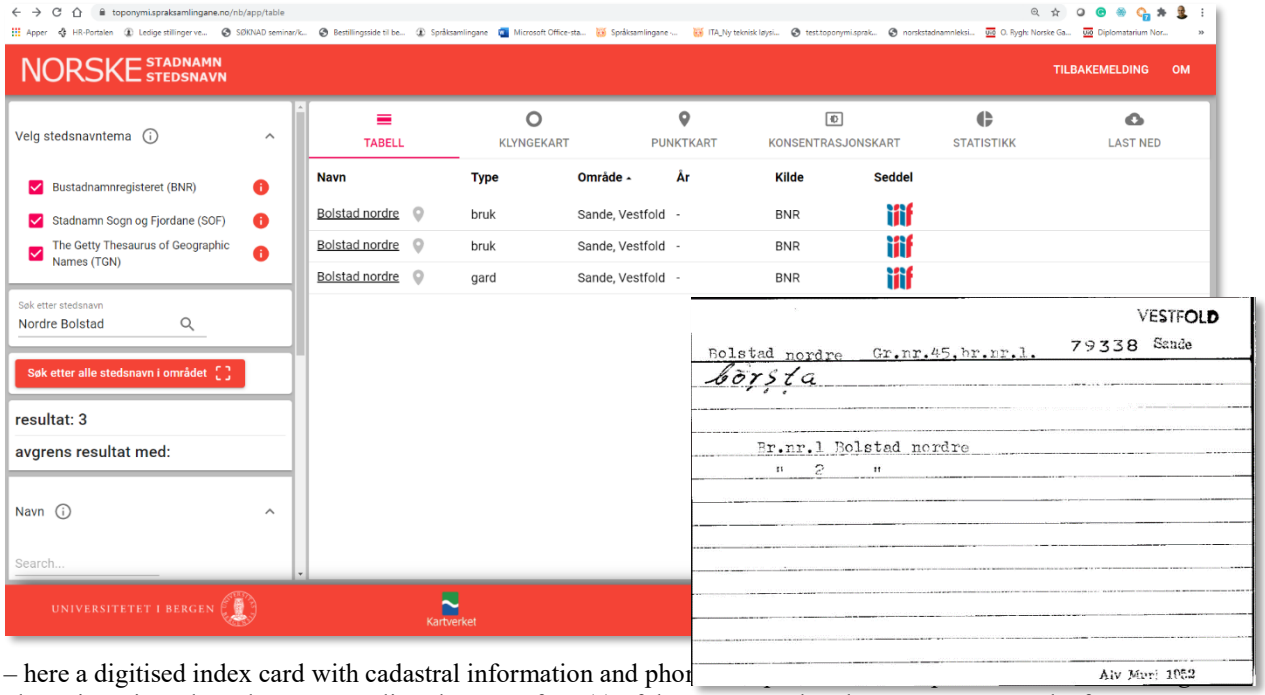
[4] The CIDOC Conceptual Reference Model (CRM) is a tool for semantic information integration. It defines the underlying semantics of database schemata and document structures used in cultural heritage and museum documentation in terms of a formal ontology to enable semantic interoperability. http://www.cidoc-crm.org/.

[5] https://seco.cs.aalto.fi/applications/sampo/?print=1

[6] https://nimisampo.fi/fi/app/table

it does not have an available endpoint as by yet, why we had to set up our own server environment for the dataset.

Figure 2. List search result view from the NPN portal with inset digitized index card. The figure shows a typical tabular search result representation. The head form links to the Sazuko Trifid landing page, the pin indicates that the entry has coordinates, and the IIIF icon on the right side (*Seddel*) will activate images associated with the entry



– here a digitised index card with cadastral information and phor
phonetic script. The columns *Navn* lists the name form(s) of the query result, whereas *Type* states the feature type, *Område* the administrative information given in the dataset. The column *År* states the year of the source, if known, and *Kilde* gives the source abbreviation of the dataset from where the name form is derived.

To test how datasets from an external source worked, we included one external Sparql[7] endpoint-configured RDF-service (Getty Thesarus of Geographic Names). The Getty Sparql-endpoint was included also as dataset in the original Nimisampo application. Since it also contains Norwegian place names, we included it as an example of using external endpoints. The test datasets comprise 500,000+ source forms for geographical names. In a web semantic environment, data is infinitely interlinked via triples. The NPN test datasets contain 217,353,538 triples or facets. The number of triples is comparable to the Finnish Nimisampo portal's 241,068,456 facets, although this portal contains in approximately 2 million geographical name forms. This suggests that the NPN portal is using a richer dataset.

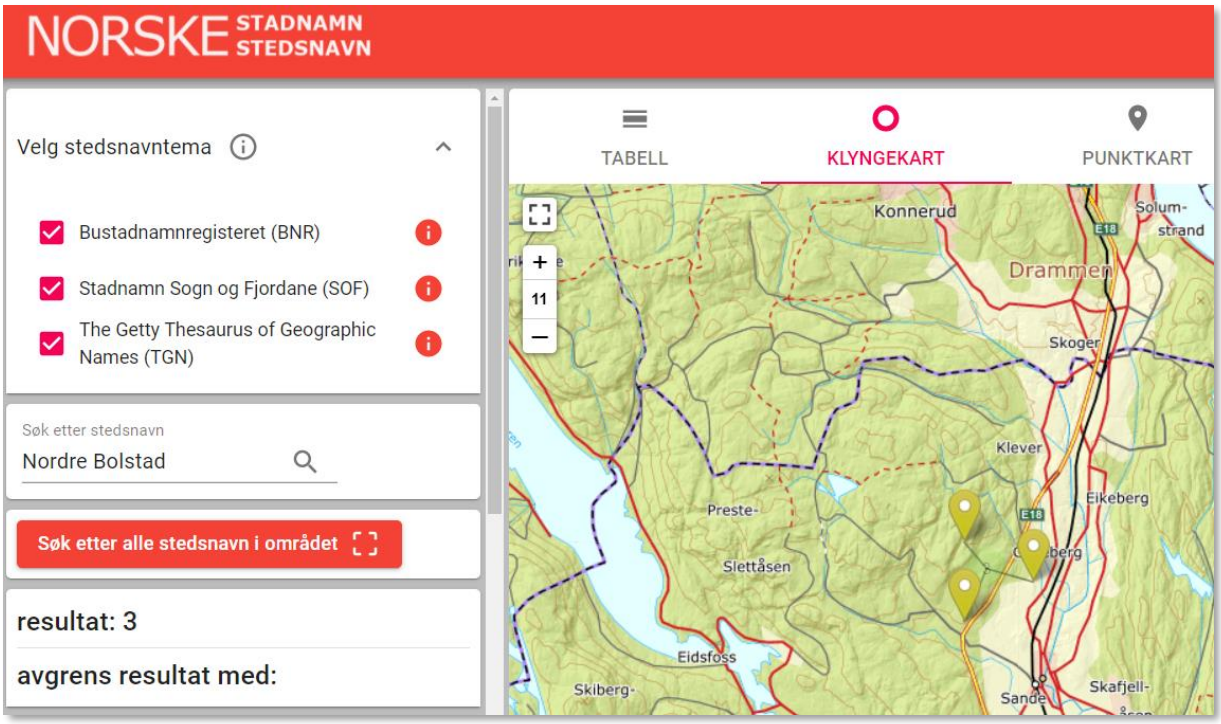
The in-house dataset, Bustadnamnregisteret, also contains digitised index cards which are available from the application, at the same time as having a regular landing page similar to that of the Getty Gazetteer. The dataset from Sogn and Fjordane does not have a landing page as such, so one is directed to their local URI of the individual place-name entries. These are the main differences the user can observe between the three test datasets.

So far we have made minimal changes to the SampoUI design, so the interfaces of Nimisampo and the NPN service are currently instantly recognisable. Our changes, wherever they occur, are under the skin, so to speak. Owing to some peculiarities of representing place-names in different Norwegian sources, we have had to tweak the geographical names search considerably and implement a specialised kind of search.

[7] https://www.w3.org/TR/rdf-sparql-query/

Speaking of search, there are two main ways of querying the data. Either by means of a string-search, a "search by name", or a geographical search, activated by means of a "search within an area" from a specially developed map search window. The former query type will find all entries in the datasets, whereas the latter, being coordinates reliant, will only retrieve entries with coordinate facets. In an ideal world, both searches would be able to retrieve similar results, but the nature of geographical names data does not allow for this.

Figure 3. Cluster map result view of same query result as in figure 2.



When the search is completed, there are facets which can be applied for filtering the query result. Only facets found in the aggregated query result from the selected datasets can be used to filter results. The NPN test portal has a number of test visualization available, a list view, a point map view, a cluster map view, a heat map view and a statistics module. A decision on which views to implement in the final result and how to, is still not made.

In addition to map views, there is also a download facility where query results may be downloaded. This is in compliance with the open-data strategy of the Language Collections and the University of Bergen. The download is currently in the form of CSV-files with geo-location, administrative information and central facet URIs.

All the datasets published have configurations containing an endpoint, a name, and the query. From now on, all which is needed to add additional datasets is to publish the dataset to a Sparql endpoint, configure it and write a query to be able to display the similar query results.

Digitally available data

As mentioned, local pronunciation information, and to some extent historical forms, form the basis in determining the standardized spelling of geographical names in Norway. The NPN portal enables the geographical names experts a single-point of access to geographical names information.

Digitally available datasets increase continually. The Norwegian Place-Name Archive have the following datasets available:

Geographical names volumes

- O.Rygh: Norske Gaardnavne (Norwegian Farm Names) 0,19 mill.

Digitised geographical names archives:

- Norwegian Place-Name Archive 0,70 mill.
- Settlement Name Archive 0,24 mill.
- Uni. of Bergen GN Archive 0,30 mill.
- Uni. of Stavanger GN Archive 0,15 mill.
- Uni. of Tromsø GN Archive 0,22 mill.
- NTNU GN Archive 0,10 mill.

Cadastres:

- 1836 Cadastre 0,11 mill.
- 1886 Cadastre 0,21 mill.
- 1950 Cadastral proposition 0,78 mill.
- Cadastre ca. 2010 1,30 mill.

Norwegian Mapping Agency datasets

- Previous GeoID series (2016) 1,00 mill.
- Current dataset 1,10 mill.

Other national datasets

- Geonames 0,75 mill.
- OpenStreetMap, settlements 0,03 mill.

Censuses

- 1900-Census 0,28 mill.
- 1910-Census 0,30 mill.

Total appx. 7,9 mill.

All the datasets from the Norwegian Place-Name Archive are supplied with geographical names URIs cadastral location URIs as well as URIs to enable linking of geographical names between different datasets and to other web semantic services. External datasets do not, however, currently have the same level of linking. The system of individual URIs for both name and location is developed specifically to cater for geolocation-enabled geographical names services, to represent the complexities of geographical names. The principles have been outlined in the recent UNGEGN Bulletin, 58[8]: *Name versus place – an unresolved problem with geodata*, pp. 19-20.

In the pipeline is also data enrichment and georeferencing of additional datsets, such as Censuses from 1865, 1875 and 1920, Norwegian postal address books from 1901 and 1972. To this will also eventually be added, the microtoponym geographical names collection campaigns deposited in the Norwegian Place-Name Archive as well as the geographical names collection campaigns of the 1930's carried out by school children (and their relatives) deposited in place-name archive of the University of Bergen. These datasets are expected to add in the region of 1 million additional name forms to the portal.

[8] https://unstats.un.org/unsd/ungegn/pubs/Bulletin/UNGEGN_bulletin_no.58_May2020.pdf

Development and launch

The Norwegian Geographical Names portal *Norske stedsnavn* / *Norske stadnamn* is planned for official launch in connection with the centenary celebrations of the Norwegian Place-Name Archive in October 2021, as announced on the UNGEGN homepage. The majority of digitally available datasets are expected to be part of the portal by then, albeit probably not all.

One aspect which will be needing further legal investigation later in the year, is the possible hinderance of digital transformation of archives and collections resulting from recent enforcement of GDPR (European Union General Data Protection Regulation,[9] EU 2016/679) regulations in Norway and the European Union. In its worst consequence, these regulations will not allow for an open and free exchange of geographical names information to regulators and the general public. However, this issue is still being explored by the University of Bergen.

Point for discussion

(a) The Group of Experts is invited to express its views on the need for historical and linguistic documentation of geographical names in matters of standardization.

(b) Express its views on the way forward concerning development of web-semantic technologies for geographical names standardization management.

(c) Express its support for geographical names management as also belonging to the scientific domaine.

(d) Give examples of similar work carried out in other countries.

The Group of Experts is requested to:

(1) Take note of the effort UNGEGN has made to encourage the use of web semantic technologies in geographical names standardization management.

(2) Endorse UNGEGN to continue, and appropriately accelerate, encouraging member states to implement the principles laid out in UNGSGN I/4, Recommendation B, in digital geographical names standardization management systems.

[9] Please see: https://eur-lex.europa.eu/eli/reg/2016/679/ for the legal text.