Establishing connections: Making resources available through the CLARIN infrastructure

Zeldenrust, Douwe
Meertens Institute
Royal Netherlands Academy of Arts and Sciences
Joan Muyskenweg 25
1019CJ Amsterdam
The Netherlands
douwe.zeldenrust@meertens.knaw.nl

Kemps-Snijders, Marc
Meertens Institute
Royal Netherlands Academy of Arts and Sciences
Joan Muyskenweg 25
1019CJ Amsterdam
The Netherlands
marc.kemps.snijders@meertens.knaw.nl

Abstract
This paper reports about establishing connections between datasets and the CLARIN infrastructure. One case is described in greater detail. This case involves the microtoponym collection of the Meertens Institute. The paper explains the technical aspects. It will conclude with a reflection on future steps in connecting research data and tools to infrastructures such as CLARIN.

Introduction
In the last five years investments in e-infrastructures for the humanities include the second European Strategy Forum for Research Infrastructures (ESFRI) roadmap. It earmarked amongst others projects CLARIN and DARIAH. In 2010 the EU made a report available concerning scientific data. It was recognized that, while these projects provide an essential layer in a collaborative data infrastructure, appropriate attention should be given to involving participants at the user and data generators level (European 2010). According to the report this can be done via user functionalities, data capture and transfer and Virtual Research Environments, VREs.

Project Alfalab (Royal Netherlands Academy of Arts and Sciences (KNAW), reached parallel conclusions (Zeldenrust, 2011). This resulted in the constructions of three VREs to address the problem of sustainability and applicability of tools tailored to heterogeneous data. At the same time the connections with large e-infrastructures such as CLARIN were established. This paper reports about connecting the dataset of one of the Alfalab VREs, namely the microtoponym collection of the Meertens Institute, to the CLARIN infrastructure. First it will describe the dataset and project Alfalab. Next it will go into technical detail concerning the connections. Finally this paper will conclude with a reflection on future steps in connecting research data and tools to infrastructures such as CLARIN.

A collection of microtoponyms and its history
The Meertens Institute, an institute of Royal Netherlands Academy of Arts and Sciences, studies the diversity in language and culture in the Netherlands. Onomastic variation is part of these studies. From beginning of the Second World War until the early nineties, the Meertens Institute has been gathering and processing information about the microtoponyms in the Netherlands. A microtoponym is the term for the names of small entities in both natural and man-made landscape. The first category covers all sorts of rugged features, such as moors, natural forests, marshes and streams etcetera. The second covers cultivated landscape and includes individual parcels as well as arable land, grazing land and man-made forests (Zeldenrust, 2005).

The microtoponym collection of the Meertens Institute consists mainly handwritten cards, which state the name, the location, the soil composition, the use and the source. Figure 1 shows a card of a microtoponym in the municipality of Heiloo called Amerika.1

---

1 This specific microtoponym Amerika is of a field used as grazing land. The municipality of Heiloo is in the North of Holland (Rentenaar, 1985).
With this conclusion in mind a pilot project was started in 2004 called: Digitization of rural Micro TOponyms (DIMITO). Its main goal was to explore the potential for digitization on the basis of a small sample from the available material and to find out whether it is useful and realistic to digitize the entire collection. In 2006 the project was brought to a successful conclusion. A working prototype with the microtoponyms of the municipality of Heiloo (about 400 microtoponyms and 9 maps) was constructed. This pilot established that, even though some aspects would present a problem, it would be possible to digitize the microtoponym collection and to set up a geographic information system. With such a system qualitative and quantitative research questions could be addressed (Zeldenrust, 2005; Zeldenrust, 2008).

**Alfalab and the construction of a microtoponym VRE**

To sustain the DIMITO project it needed a tailored framework that would be both innovative and practical. In 2009 project Alfalab was initiated at Royal Academy level that would turn out to provide such a very framework. Alfalab finished in September 2011 and was an initiative that networked the knowledge that is needed to develop and disseminate VREs. These VREs are online research platforms that are aimed at supporting a specific network of researchers, providing them with the tools and data they need to address the shared research objectives in a collaborative approach. A typical VRE will to this end combine communication tools (mail lists, forums, shared document portfolio) with tools to create, curate and analyze research data, as well as some form of publication tools (Zundert et al. 2009).

To reach the goals of the microtoponym VRE (called the GEOlab) first of all the entire collection had to be made digital available. The 170.000 cards with microtoponyms and the more than 2400 geographical maps were therefore scanned. The accompanying metadata of the maps is stored in a database. It contains data like publisher, date, municipality etc. The 170.000 cards are organized on base of municipality.

Second: for effective use of the data for research purposes it is required that the existing data could be enriched. This enrichment process consists of two steps. The first step is to enter the data from the cards in the database and the second step is to add the geographical information. Within the GEOlab these steps can be made with the two web based tools that have been developed: a georeference tool to position the maps and a geo-annotation tool and to locate the microtoponyms (the cards) and for data entry. The GEOlab will consist of two main sections: a central repository in which datasets are stored and managed and the web tools for enriching the data (Meertens, alfalab).

After the data has been enriched the collection can be used for traditional onomastic research and it will also be possible to introduce new research questions and possibilities. The following examples will make this clear. According to toponomologist Prof. Dr. R. Rentenaar the microtoponym *Amerika* was used for political motivation, migration related situations or for remote fields (1985). Rentenaar suggests that this particular field in Heiloo was...
called *Amerika* because it was grazing land situated on a large distance from the village. With a GIS and a digital, enriched collection it will not only be possible to calculate the distance of this and the other 157 to North America related microtoponyms to the nearest village. But if you add a map layer that consists of the transportation infrastructure you could calculate not only the distance in a strait line, but also what it will be if you walk (as farmers did in those days) on the roads. This data can be compared with the distances of other microtoponyms to villages (and farms if the data is available). The calculations will give new information and could sharpen the theory.

The digital microtoponym database and tools could also be used outside the traditional onomastic discipline. Rentenaar writes about names in literature that are used as microtoponyms. Eponymized place-names in connection with the name America are (next to the name *Amerika*) most of the time names of American States such as Florida and Maryland etc. Rentenaar also finds a name that is connected to the book of Karl May: ‘Der Schatz im Silbersee ‘ (The treasure of Silverlake). In Dutch it is called ‘De schat van het Zilvermeer’ and the name Zilvermeer is found in Nieuwe Biltdijk as the name of a pond. Rentenaar sees this as a literary place-name. According to Rentenaar this is a rare phenomenon in the eponymized place-names.

The digital collection will also make it possible to compare the microtoponyms with, for instance, databases that contain titles of literary work. That way we might find more eponymized place-names Rentenaar wasn’t aware of and verify if it is indeed such a rare phenomenon. But it also works the other way around: the microtoponyms can be used to give more information about names used in literature or other works and positions them. And not only certain episodes in literary work could be given a specific place. In a time before the Global Positioning System (GPS) and accurate maps, toponyms were often used to position an event. A database with toponyms (with exact positions) could be used to position events described in newspapers or other sources.

---

**Establishing the connections**

While the primary aim of Alfalab was to support a research network, starting from the researchers and the research questions involved, Alfalab’s secondary aim was to address the problem of dissemination of digital tools and data. The connection with CLARIN provided direct solutions for the data involved using the principles of persistent identifiers (PIDs), the CMDI (Component Metadata Infrastructure) metadata framework and OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting).

In many cases the scanned versions of the cards consist of two separate scans, one for the front and one for the backside, yielding a total of over 171,000 high resolution tiff image scans. To make these available in the GEOlab lower resolution JPEG images were created to facilitate loading and display in GEOlab web interface. To make all resources available through the CLARIN infrastructure each resource must be described using CMDI metadata, assigned PIDs, made web accessible and published through OAI-PMH.

Metadata preparation included assessing the availability of relevant metadata information. Since all cards consist of hand written material, the content of the fiches themselves could not be used to generate metadata information. Field names and field locations, i.e. geo coordinates, were derived from contextual information, sometimes even from the file names of the scanned images. Additional techniques were used to extract additional information form the images themselves by looking into the image (EXIF) header information. Based on the available information a custom metadata schema was devised to describe these types of resources. It was decided to use a single metadata document to describe each fiche where each metadata document could contain up to four links to related resources (up to two links to the TIFF front and back card scans and up to two links to the low resolution JPEG versions).

When constructing the metadata schema, care was taken to make the structure as compliant as possible to already existing components in CLARIN’s Component Registry (Catalog Clarin; Broeder, 2010). New components were registered where needed and combined to create a custom ToponymProfile. All metadata field definitions that could not be matched to those already available in the ISOCat Data Category Registry were created and linked to the components’ specifications in accordance to CLARIN guidelines (ISOcat).

Once the metadata creation process was completed each metadata document and corresponding resource was prepared for further publication. Metadata documents and resources were assigned persistent identifiers using the persistent identifier services provided through the EPIC consortium. A local handle service was installed under our own handle prefix (10744) allowing us to manage handles from the institute under a separate handle prefix. The advantage of this approach is that any handle created at the Meertens Institute can be managed in a secure environment without interfering with handles requested by other institutes. As a backup scenario this also allows

---

Figure 3: Screen shot detail of GEOlab environment with card overlay onto Google maps
us to migrate to a different local handle server without many problems should the situation arise.

To make each of the resource available in a web accessible manner all resources were placed on a web server with the possibility to enforce authorization control using the Single Sign On features provided by Shibboleth. This means that, where needed, resources can be individually protected and access may be granted to individual users within the CLARIN federation. While the resources for the DIMITO and Alfalab project are freely available we were able to experiment with the principles of IdPs (Identity Providers) and SPs (Service Providers) and set up an access control system for all resources.

To allow others to harvest our metadata records an OAI-PMH provider was set up serving DCMI (Dublin Core Metadata Initiative) and CMDI documents. Metadata documents are published through this as a standard procedure in the publication process so they may be integrated into central catalogues such as CLARIN’s Virtual Language Observatory. To facilitate the OAI-PMH publication process a CMDI to DCMI conversion was made which was realized through a standard stylesheet (XSLT) transformation.

For internal and external use the metadata records are stored separately in an eXist database which is made web accessible to ensure that all metadata documents are accessible using the assigned PIDs. Metadata documents are furthermore indexed and made searchable using SOLR, a popular open source search platform. It features full-text search, faceted search and geospatial search. More specifically for the microtoponym data, but in fact for all Meertens data latter functionality offers an interesting alternative method for locating data sets.

Conclusion
The microtoponym dataset used in Alfalab’s GEOlab is connected to the CLARIN infrastructure and therefore available for future research. In that sense the goal envisioned in the 2010 EU report had been reached. But, as the EU report rightly mentioned, tools are part of the infrastructure as well. In the case of the GEOlab: next to the dataset the lab contains web-based tools for adding geographical annotations to linguistic and historical data (a georeferencing tool and a geo-annotation tool). Questions about storage, access and sharing of these tools remain a challenge that still lie ahead and need to be solved.

References


Marc Kemps-Snijders, Menzo Windhouwer, Sue Ellen Wright (2008). Putting Data Categories in their Semantic Context
Archives:
The archive of the Meertens Institute, collection no. 49, field name map collection ca. 1860-1964.
The archive of the Meertens Institute, collection no. 99, field name collection 1941-1992.
Websites:
www.meertens.knaw.nl/books/veldnamen (accessed October 06, 2011).