On Modelling a Typology of Geographic Places for the Collaborative Open Data Platform histHub

Manuela Weibel, Tobias Roth

Schweizerisches Idiotikon, Auf der Mauer 5, Zurich, Switzerland

manuela.weibel@idiotikon.ch, tobias.roth@idiotikon.ch

Abstract. HistHub will be a platform for Historical Sciences providing authority records for interlinking and referencing basic entities such as persons, organisations, concepts and geographic places within an ontological framework. For the case of geographic places a draft of a place typology is presented here. Such a typology will be needed for semantic modelling in an ontology. We propose a hierarchical two-step model of geographic place types: firstly, a more generic type remaining stable over time that will ultimately be incorporated into the ontology as the essence of the identity of a place. Secondly, a more specific type closer to the nature of the place the way it is actually perceived by humans.

Our approach to prepare the existing data for the future place typology is decidedly bottom-up: we standardise the place types in our database of heterogeneous toponymic data using the place types already present as well as textual descriptions and name matches with typed external data sources. The types used in this standardisation process are basic conceptual units that are most likely to play a role in any place typology yet to be established. Standardisation at this early stage leads to comprehensive and deep knowledge of our data, which helps us developing a suitable place typology.

Keywords: Place Typology, Ontology, Collaborative Platform, Open Data, Interoperability, Text Mining.

1 Introduction

Historical Sciences in the digital age have become aware of their need to be able to digitally refer to uniquely identifiable and stable entities. The joint platform histHub will provide authority records for a number of basic entities in Switzerland. The entity we are working on at the moment is the geographic place. In our semantic modelling, a geographic place needs a place type so that in the future it can fit into a common ontology. We will show the characteristics of the place typology we are working on for this purpose. We will also describe how we bring our own database of toponymic data with about 0.5 million records in line with this new typology.

2 Context

2.1 histHub

The ultimate goal of histHub is a platform that offers authority records for interlinking and referencing historical content in an interoperable manner. HistHub is a joint project of four partners – the Historical Lexicon of Switzerland (HLS) [1], the Swiss Diplomatic Documents (DDS) [2], the Collection of Swiss Law Sources (SSRQ) [3], and the Swiss German Dictionary (Schweizerisches Idiotikon) [4], with its toponymy platform ortsnamen.ch [5] – all of them important actors in Switzerland's research community of Historical Sciences.

The project started in 2016 and will end in 2019. It aims at providing its data and services mainly on a Swiss national level. However, doing this will also allow for the connection to the international level. Interoperability and compliance to standards is ensured by using an ontology widely adopted in Historical Sciences (CIDOC CRM, see below). HistHub feels committed to the ideas of Open Science: we apply open licences to our own data, we promote standards, and we invite the community to participate.

HistHub will provide authority records for four different basic entity types used in Historical Sciences, i.e. persons, places, organisations and concepts. The focus of this paper as well as the current stage of the project lies on the process of creation of authority records for places. Well known existing datasets to consider when dealing with places in Switzerland – besides the ones contributed by the current histHub partners – are the ones provided by TOPOterm [6], swisstopo [7], and GeoNames [8]. Other projects and researchers are welcome to contribute their data, and histHub will give them the tools to do so.

2.2 Ontology / CIDOC CRM

We use an ontology for the semantic modelling of our toponymic data. To facilitate interoperability, the ontological conceptualisation will be embedded into a more widely adopted ontological framework. HistHub is mainly focused on Historical Sciences. This is why – based on ideas brought into the project by Francesco Beretta (for an application of CIDOC CRM in a similar context see [9]) – we plan to attach the histHub ontologies to CIDOC CRM [10].

CIDOC CRM (Conceptual Reference Model) is a “reference ontology for the interchange of cultural heritage information” [11]. It has been an ISO standard since 2006 and is being developed under the patronage of the International Committee for Documentation (CIDOC) of the International Council of Museums (ICOM). Memory institutions such as the British Museum and other organisations with a focus on history are the principal users of this ontology.

2.3 ortsnamen.ch

The Swiss German Dictionary (Schweizerisches Idiotikon) operates a platform of toponymic data as one of its sub-projects: ortsnamen.ch [5]. It is the platform where the regionally organised toponymic projects in German speaking Switzerland publish their data on the internet (while the primary goal for most of these projects is the publication of a printed book). Users can look up place names on ortsnamen.ch. Typically, they will find a description, geo-referencing, an etymological explanation, and historical evidence.

There are nearly 0.5 million place-name records on ortsnamen.ch at present. Coverage is thus more fine-grained than e.g. the respective place-name dataset of the Swiss Federal Authorities: about 0.3 million records for the whole of Switzerland vs. 0.5 million records for about 80% of the territory of German speaking Switzerland on ortsnamen.ch. This large coverage is a good reason for taking ortsnamen.ch as one of the main providers for digital humanities' authority records for places in Switzerland.

In addition to its website, ortsnamen.ch is about to make its data available through web services (REST API). For better and more refined access, data is being restructured: historical-evidence details that were mostly stored in flat text fields are transformed to proper records with year, name form, etc. [10].

3 A Two-Step Model of Geographic Place Types

HistHub aims at merging data from various origins into one database. Furthermore, we want the data to be enriched continually by a growing community. Users can add, interlink, and reference data, as well as suggest corrections. Particular emphasis lies on putting the data into a diachronic context by documenting the history of every instance of place, person, organisation, and concept [11].

3.1 Identity Markers of Geographic Places

When working with historical geographic data it is often difficult to judge whether two place mentions refer to the same instance. A place can change name, extension, and function over decades and centuries. Rather philosophical-ideological questions also have to be taken into consideration: should a handful of small prehistoric lake settlements in the lower area of the Lake Zurich be considered precursors of the modern city of Zurich or should we treat them as independent instances?

Our goal is to enable consistent handling of new records coming from different users. In a collaborative effort with the whole team we identified the need for a feature that denotes the essence of a place and remains consistent over time. Instances of one place must share this essential feature. In consequence, places that differ in the essential feature have to be added as independent records to the model (see Ex. b).

In the process of identifying the essence of a place, we approached a solution by looking at different examples, such as the following ones:

(a) The *Highland Tolbooth Church* of Edinburgh was built between 1839 and 1844. The building was given different roles ever since: not only is it a multi-functional complex (called *The Hub*) with a library and rooms for conferences and weddings today, it was also home to the Scottish Parliament in the 19th century. The place took a series of different functions over time, but it always remained: a building.

(b) *Schlossacker* is the name of an industrial area in the city of Winterthur. The toponym reveals that the modern district was built on a piece of land that used to be *farmland* (German *Acker*). Name and geographic extension of the place have remained identical over time, yet this time, the identities of the *district* and the former *farmland* are not compatible: the form has changed from a cultivated piece of land to a city part with streets, buildings and other infrastructure.

When evaluating these cases we felt that we had come closer to our solution of how to determine the essence of a place: the essence can be described by an identity marker. This identity marker, in turn, is closely related to the type attribute of a place (*building, farmland, district*):

The Hub of Edinburgh can be considered an instance of the same place that also describes the former *Highland Tolbooth Church* since they both share the type *building*. In case the building had been torn down and given place to a town square, the place would have undergone a major identity change, visible in the place type, changing from *building* to *square*. According to the ideas of our team, these cases would have to be represented in two separate (but semantically interlinked) records.

We saw that the place type was important. Still, this was not yet the final solution. Considering the actual place type to be the identity marker, the different place instances in the example of Zurich (*small settlements* vs *city*) have to be modelled as separate records. Yet, not every historian would agree with this approach. Since we do not want to dictate to the users how to handle critical cases like this one, we looked for a model that is flexible enough to allow for different options.

The key lies in clustering the types into slightly more generic groups. In the latest example, *small settlement* and *city* can be grouped to the cluster *populated place.* We consider these clusters of place types to be identity markers.

As a conclusion, we think that the essence of the identity of a place is marked by a more generic place type. Two places whose types share the same identity marker might be instances of the same place.[1] Two places with different identity markers, however, can never be instances of the same place.

[1] We allow the community to decide autonomously whether they want to create independent or interlinked records.

3.2 Extending the Ontology and Building a Taxonomy

We elaborated a two-step model that takes into account this distinction between more generic (e.g. *populated place*) and more specific (e.g. *city, settlement, town, hamlet*) place types:

The identity markers (clustered groups of types) will be integrated into the ontology by forming new ontological classes. The final integration will be implemented within a separate project milestone. The more detailed types form part of a typological taxonomy yet to be elaborated and integrated into the ontology.

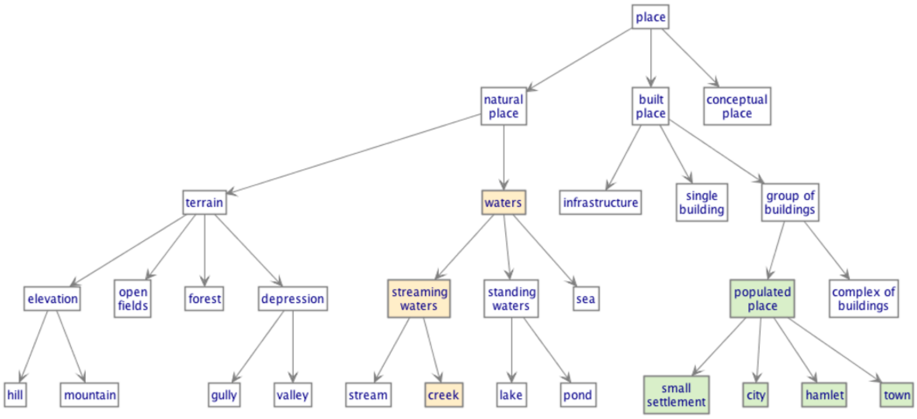

Fig. 1. Exemplary and strongly simplified extract of the place type taxonomy.

We want to put the new typology (*histHub-geotype*) on a broad basis by taking into account existing typologies and by involving not only all project partners but also external experts. The goal is to create a multilingual place typology with historical and contemporary type and function designations.

4 Data Cleansing: Type Assignment in ortsnamen.ch

Given the heterogeneous sources of the toponymic data on ortsnamen.ch, no systematic place type assignment had been deployed so far. To benefit from a future expanded ontology as described in Chapter 3, the existing database had to be updated with place types. In this Chapter, we describe how we used data from internal and external sources to make assumptions as to the type of a place.

4.1 Current State of the Data

Amongst several other fields, toponyms in the ortsnamen.ch database possess fields for the coordinates of a place and its description. These entries are substantial for the process of type extraction. 65% out of a total of nearly 0.5 million toponyms come with complete coordinate indications. 60% provide a textual description. Only 9% of the records have neither coordinates nor a description.

4.2 Type Assignment Based on a Reference Database

At first, we compared the existing toponyms and their coordinates with the information in a type-based dataset of the Swiss Federal Authorities (*SwissNames3D* [13]). For every match, we extracted the place type indications from the reference database:

Table 1. Enriching the norm database with place types.

Name of Database	Place Name	Coordinates	Type
ortsnamen.ch	*Klein Ruchi*	*723'130, 191'930*	-
swissNames3D	*Chli Ruchi*	*722'963.80, 191'929.48*	*Hauptgipfel (‚main summit')*
histHub norm database	*Klein Ruchi*	*723'130, 191'930*	*Hauptgipfel*

Thus, 36% of the overall places could be assigned a type. The remaining places without a successful match traced back to deviating or vague coordinate indications and strongly deviating name spelling.

4.3 Type Assignment through Text Analysis

Another significant source providing insights concerning the place type of a toponym is its description text: the authors of the source material made no effort with respect to type assignment or consistent typology. However, they frequently referred to one or several type designations in the description.

To profit from these text snippets, we built a list-based algorithm that linguistically analyses the toponyms and their descriptions: by part-of-speech tagging the snippets we were able to focus on the nouns, thus reducing potential sources of errors. All (but a few[2]) nouns – as well as every place name – were then checked against a list of type designations (for more details about the list see Chapter 4.4). The result of this proceeding was a set of suggested types for every toponym, paired with a probability score that rates the following cases (with decreasing weight):

- Number of matches coincides with number of nouns in description
- Corresponding matches in name and description
- Early appearance in description
- Match either in name or in description
- Discount for different type matches

[2] We created a separate list of stop words, mainly consisting of prepositions and cardinal points. In case a noun was preceded by a bigram of tokens that includes one of the stop words, the noun in question was not considered to be a potential type designation any further.

Table 2. Type suggestions based on text analysis (matches highlighted).

Place Name	Description	Matched Types	Suggested Type	Score
*Mühlen**weg***	***Waldung***	*Weg, Wald (forest)*	Wald	0.99
*Ärbs**bach***	***Bach** von der Quelle im Oberstafel bis zur Einmündung in den Sernft*	*Bach (creek)*	Bach	0.98
*Chulm**flue***	*Grosse **Felswand** am Niederbauen (Se. 307).*	*Fels (rock), Felswand (rock face)*	Felswand	0.86
Hinter Rosegg	*Zur Hauptsache ebenes, sumpfiges **Weideland** im Talgrund; im Bereich der **Talflanke** leicht ausgewölbter **Hang**.*	*Weide, Flanke, Hang*	Weide	0.76

This procedure allowed automatic tagging of another 26% of the overall data and generated suggestions for 10% of the toponyms, which will have to be checked manually due to a low matching score.

4.4 Remarks on the Typology

The taxonomic decisions concerning the type clustering (cf. Chapter 3.2) will form the base for the new place typology *histHub-geotype* that will be elaborated in the course of the upcoming project milestone [11].

The final version of the typology still pending, we opted for building a broad list of types as the base for the text analysis described in Chapter 4.3. This list consists of a wide variety of type denominations and regular expression search patterns and was built recursively by systematically evaluating nouns in the description texts. Examples of search patterns are:

Table 3. Examples of search patterns applied to the data.

Search Expression	Explanation	Example
[aä]cker$	matches tokens ending in *-acker* or *-äcker* (‘farmland’)	*Stampf**äcker***
(?<!zusammen)hang$	matches tokens ending in *-hang* (‚slope’) unless preceded by *zusammen-* (*zusammenhang* ‘context’)	*Wald**hang,*** *Weide**hang***
schul(e|haus)	matches *Schule* (‘school’) or *Schulhaus* (‘schoolhouse’)**.**	*Lärchen**schulhaus***

This approach lets us prepare our data by enriching it with standardised types. The list shows a very low level of type abstraction, which allows us to eventually map the types to a more elaborate typology (*histHub-geotype*): As a result of the text analysis, the place called *Ärbsbach* (see Table 2) is assigned the standardised type *Bach* (German for ‘creek’). Depending on the level of detail of the yet to be established typology, it may later be assigned the types *Bach* ‘creek’, *Fliessgewässer* ‘streaming waters’ or simply *Gewässer* ‘waters’ (see Fig. 1).

5 Conclusion

After intensive discussions of the ontological modelling of geographic places, we realised that the place typology we planned to build was crucial for our ontology. For the reasons explained in Chapter 3, we proposed a hierarchical two-step model of geographic place types: a detailed, fine-grained type that precisely denominates the character of a place instance, e.g. *park, billabong, hamlet*, and clustered groups of types that remain stable over time, e.g. *infrastructure, water, populated place*. The latter will ultimately be incorporated into the ontology. This allows for a better handling of diachronic data.

In Chapter 4, we presented a bottom-up approach to prepare our data with regard to the previously described findings: we standardised the place type information in our database of heterogeneous toponymic data. This was done by comparing the data to a reference database on the one hand, and by semi-automatically extracting type designations from internal information like the description text on the other hand.

Our approach can also be seen as an example of how to infer structured data from unstructured and related (structured) data. This newly gained data will be used to improve interoperability in the context of the histHub platform. Within histHub, this data can again be used in much the same way in order to infer knowledge through linking following standards.

References

1. Historical Lexicon of Switzerland (HLS), http://www.hls-dhs-dss.ch, last accessed 2017/10/20.
2. Swiss Diplomatic Documents (DDS), https://www.dodis.ch, last accessed 2017/10/20.
3. Collection of Swiss Law Sources (SSRQ), https://www.ssrq-sds-fds.ch, last accessed 2017/10/20.
4. Swiss German Dictionary (Schweizerisches Idiotikon), https://www.idiotikon.ch, last accessed 2017/10/20.
5. ortsnamen.ch, https://www.ortsnamen.ch, last accessed 2017/10/20.
6. TOPOterm, http://www.topoterm.ch, last accessed 2017/10/20.
7. Bundesamt für Landestopografie swisstopo, https://www.swisstopo.admin.ch, last accessed 2017/10/20.
8. GeoNames, http://www.geonames.org, last accessed 2017/10/20.
9. Beretta, F., Alamercery, V., Bruseker, G., Doerr, M.: Aligning the symogih.org ontology with the CIDOC CRM: towards a CRM extension for historical data. Presentation at the 38th joined meeting of the CIDOC-CRM SIG and ISO/TC46/SC4/WG9 and the 31th FRBR - CIDOC-CRM Harmonization meeting, Apr 2017, Heraklion, Greece (2017).
10. Roth, T.: Isolation and Mapping of Place-Name Forms in Toponymic Data. In: Dipper, S., Neubarth, F., Zinsmeister, H. (eds.): Proceedings of the 13th Conference on Natural Language Processing (KONVENS) (= Bochumer Linguistische Arbeitsberichte 16), https://www.linguistics.rub.de/bla/016-konvens2016.pdf, pp. 221–225 (2016).
11. Consortium Historicum: histHub – Vernetzte und normierte Daten für die historischen Wissenschaften. Project proposal, unpublished (2017).
12. ISO: ISO 21127:2014 Information and documentation – A reference ontology for the interchange of cultural heritage information, ISO standard (2014).
13. swissNAMES3D, https://shop.swisstopo.admin.ch/de/products/landscape/names3D, last accessed 2017/10/20.