An Ontology of German Place Names[*]

Sebastian Nagel[†]

May 2005

My research has given rise to the construction of an ontology containing geographic entities and its German names. The ontology includes (1) the linguistic features of place names, such as their inflectional morphology and characteristic syntactic behaviour, (2) associates place names to geographic entities, and (3) the relationship between one geographic entity and another.

1 A Dictionary of German Toponyms

A dictionary of German toponyms (place names) was first constructed so that their linguistic features could be systematically described, i.e. their morphology and syntactic behaviour. The dictionary's format is compatible with DELA (Courtois 2004), hence it can be used with Unitex and Intex (see Prolintex for a French equivalent). This software is used to implement local grammars which describe NPs and PPs containing toponyms and classifiers such as:

(1) rund 1500 Kilometer südsüdwestlich von Honolulu ‘about 1500 kilometres south-southwest of Honolulu’

(2) in der englischen Grafschaft Suffolk
in the English shire Suffolk
‘in Suffolk County, England’

The dictionary was constructed using toponyms extracted from freely available resources.[1] Because the list contains many errors and impurities such as spelling errors and common nouns denoting geographic features, e.g. *harbour*, the list was only used as input for a classifier based on local grammars which describe typical syntactic patterns a toponym may appear in. From a 35 GB corpus, a frequency list of occurences of these patterns was extracted.

With help of these patterns a preliminary classification of the toponyms was made based on their gender and associated determiner (see below). E.g., if the context of the toponym was found to contain “*in der X*” (‘in the$_{\text{fem}}$ *X*’) a counter was incremented signalling *X* is feminine and must be used with the definite article. The pre-classified entries were then manually corrected whereby ambiguities and overlap with homographic common nouns were resolved and inflectional information was added. Actually all toponyms in the corpus with a frequency of 100 or over have been entered in the dictionary.

Relational adjectives and the names of the inhabitants of cities and regions are added for approximately 1000 toponyms. This work had to be done manually with the aid of frequency lists and examples taken from the results of internet search engines.

1.1 Morphology

The inflectional morphology of German proper nouns is relatively simple compared to that of common nouns. No vowel gradation (‘Umlaut’) is observed at all. In $^2/_3$ of the dictionary entries, only

[*] To appear in *Corela – Cognition, Représentation, Langage* Numéros spéciaux, Le traitement lexicographique des noms propres.

[†] CIS, University of Munich, wastl@cis.uni-muenchen.de

[1] Among others: NGA:GNS, HistOrtsbuch, bund.de and Wikipedia.

the genitive form differs from the base form where an *-s* is appended to the word (*Berlin* → *Berlins*). Some toponyms, especially those consisting of two or more tokens, show more complex inflectional patterns. The 'Gulf of Persia' has eight different forms: *Persischer Golf*, *der Persische Golf*, *des Persischen Golfes*/*Golfs*, *Persischem Golf/e*, *dem Persischen Golf/e*.

Inflectional classes are described using the CISLEX standard (Maier-Meyer 1995) with some classes and extensions added to handle variants and multiword entries. Altogether 70 different inflectional classes were found.

Two other phenomena are also covered in the inflectional module: (1) Abbreviations of toponym parts (*Frankfurt am Main* → *Frankfurt a. M.* or *Frankfurt/Main*) and (2) deletion of diacritics (*Orleans* is used more frequent than *Orléans* in German texts).

As expected, the derivational morphology of toponyms shows even more variation than inflectional morphology. The most frequent suffix used to form relational adjectives and names of inhabitants is an *-er* appended to the toponym, e.g. *Berlin* → *Berliner*. The final *e* or *en* in the topoynm is often ellipsed before application of the suffix, e.g. *München* → *Münchner*, *Bremen* → *Bremer*, but *Essen* → *Essener*. The bulk of suffixes occur only once (e.g. *Monaco* → *Monegasse*). In total over 300 derivational suffixes were found: 110 to derive the adjective and 200 to derive the name of male and female inhabitants. Because the derivational endings were so idiosyncratic, no attempt was made to systematize them (similar to inflectional codes). Instead, each derivation was given the status of a lemma and was linked to the toponym it was derived from.

1.2 Syntactic Features

As in other European languages such as English, French and Italian, German proper nouns are not normally used with a **determiner** (definite article). This, however, is not the general rule for toponyms. The use of a determiner is, in fact, lexicalized and hence part of the wording of a toponym. The presence of the definite article must therefore be explicitly coded in each toponyms entry:

(3) a. in Frankreich 'in France' (`+DetZ`)
b. in der Türkei 'in Turkey' (`+Det`)

The determiner *der* in (3b) is compulsory. It can only be omitted in texts written in a telegraphic style, e.g. in head-lines:

(4) Vier Tote bei Terroranschlag in Türkei
'four dead in terrorist attack in Turkey'

The use of the definite article differs depending on the type of toponym. Approximately 40%of geographic regions (e.g. states, islands etc.) are used with the definite article. However, only a small number of towns, villages and cities (0.005%) require the definite article. In contrast, almost all names of mountains and mountain ranges and all names of bodies of water are used with the definite article. Hence, except for the last class (see below), no rule describing the use of determiners with German geographic names could be established.

All German toponyms agree in terms of linguistic **gender** with determiners, adjectives and, anaphoric as well as relative pronouns etc. associated with the toponyms: *der Rhein* 'the Rhine' (masc.) vs. *die Donau* 'the Danube' (fem.). The inflectional ending on these associated word classes are used to determine the gender of a toponym. For some toponyms, however, it is impossible to determine their gender by the ending on the associated word types: (1) for toponyms with an compulsory classifier, the classifier governs the gender of the phrase, and (2) for 'pluralia tantum', gender is consistently unmarked in plural number. For such toponyms, the gender is marked as 'unspecified' (see below for examples).

Some toponyms are only found in the plural (**'pluralia tantum'**):

(5) in den Niederlanden 'in the Netherlands'

However, for some 'pluralia tantum', a reference to a single element of the 'collective' toponym is possible:

(6) a. * eine der Niederlande
(*'one of the Netherlands')
b. Curaçao ist eine der niederländischen Antillen. 'Curaçao is one of the Netherlands Antilles.'

In this case, the gender of the 'plurale tantum' will be included in the dictionary.

Administrative regions are often named after their capital or an important city in them. In German, therefore, the **classifier** is required to reference to the region:

(7) a. in München 'in the city of Munich'
b. im Landkreis München 'in Munich county' (`+oblCl`)

Toponyms containing declined adjectives have different forms depending on the presence of a zero, definite, or indefinite article:

(8) a. im Persischen Golf
'in the Gulf of Persia'
b. zwischen Persischem Golf und Rotem Meer 'between the Gulf of Persia and the Red Sea'

All the features described above are useful for disambiguating between toponyms and homonymous common nouns as well as between different toponyms with the same form:

(9) a. in Essen
'in (the city of) Essen' (`+DetZ`)
b. im Essen 'in the food'

(10) a. in Gera
'in (the town of) Gera' (`+DetZ`)
b. in der Gera
'in the Gera river' (fem., `+Det`)

Coordination below the token or word level is handled directly in the lexicon. This is a practical solution justified by the fact that only a small number of toponyms (forming parts of a larger entity or located near each others) can be coordinated in this way:

(11) a. Ober-, Mittel- und Unterfranken
'Upper, Middle, and Lower Frankonia'
b. Ober- und Unterammergau
'Oberammergau and Unterammergau'

Variants with the same form but different syntactic behaviour are entered as separate lemmata:

(12) a. `Wedding,.EN+Topon+Oikon+Det:M{NS13}`
→ *im Wedding* 'in Wedding (a district of the city of Berlin)'
b. `Wedding,.EN+Topon+Oikon+DetZ:N{NS2}`
→ *in Wedding*

In the ontology, however, they are connected to one and the same geographic entity.

2 Extending the Dictionary to an Ontology

The dictionary was extended to an ontology by changing the lemmata in the dictionary into to instances of a lemma class in the ontology. The class of 'geographic entities' was introduced as central concept and all linguistic entities were linked to geographic entities.

The ontology was developed using PROTÉGÉ, an open source ontology development platform supporting features such as unicode, multiple superclasses and constraint checking.

2.1 The Linguistic 'Branch'

Central to the linguistic branch of the ontology is the concept of a 'lemma' as per common linguistic definition: a paradigm of word forms represented by one base form. All forms share the same syntactic features (e.g. gender, use of definite article), although they may differ in terms of certain grammatical categories, such as case and number. To keep the size of the ontology small, an inflectional key is included instead of the entire paradigm. Hence an instance of a lemma corresponds to a single line in the (uninflected) DELAS-dictionary.

Lemmata are divided into various subclasses: parts-of-speech and further subdivisions for constraint checking, slot overrides etc. E.g. out of 1000 hydronyms, not a single example could be found of one being used with zero determiner. Hence, it can be taken as a rule: when adding a new hydronym, the slot for the syntactic feature `+Det`/`+DetZ` will be pre-filled with `+Det`.

Relations between lemmata are handled directly by using slots which link lemmata: each noun has a slot for derivational adjectives, each toponym has a slot for the name of its inhabitants.

2.2 Geographic Entities

The second 'branch' of the ontology contains geographic entities as real-world-objects. Instances of lemmata fill the various slots in a geographic entity:[2]

- the common name (required)
- the official name (if different from the common name)
- historical names
- other variant names (including abbreviations and exonyms)

Other slots describe the relations between geographic entities:

- administrative subdivision: x is subject of y.
- x is capital of y
- x contains y[3]
- river x flows into body of water y

For an example of how slots are filled, see fig. 1.

2.3 Classification of Entities

Geographic entities are grouped into classes such as 'settlement', 'region', 'body of water', 'mountain'. The classification is based on the following two principles:

1. linguistic motivation, i.e. how a human would classify the toponym: people "do" similar things with the entities of one class. Top-level classification principles like 'natural'/'artificial' are then irrelevant (cf. Bauer 1998: 55-6). A person can swim, fish etc. in a lake (natural) as well as in a reservoir (artificial). Hence, the distinction between lake and reservoir is made on the lowest level, i.e. the class 'reservoir' is a subclass of 'lake'.
2. practicability: the classification should be self-evident to a high degree, i.e. the person performing the classification should not have to think long about it. In particular, this means avoiding splitting continua into different classes, as one can always debate the difference between a 'village' and a 'town', or a 'river' and a 'creek'.

The classes are structured within a taxonomy graph.[4] A part of the taxonomy is visible in the left window in fig. 3.

2.4 Time: Historical Names and Places

Time is handled on two levels (cf. Axelrod 2003; Tran, Grass & Maurel 2004):

1. on the linguistic level proper noun lemmata can be associated with a time period during which they were used or a valid official name: the city located at 59°54'20"N, 30°16'9"E was named *Saint Petersburg* (1701-1914,1991-), *Petrograd* (1914-1924) and *Leningrad* (1924-1991). See fig. 2 for a graphical representation.
2. geographic entities which no longer exist can be marked as 'historical' and associated with the time period they were existing: the Swiss canton *Ausserschwyz* existed only in 1831.

2.5 Classifiers

Classifiers of geographic entities, such as *capital city*, *seaport*, were also added to the ontology as subclass of the class 'N' (nouns). Three additional slots take (1) classes, (2) geographic entities as instances of classes and, (3) slots. In (1), all instances of a class, in (2) only the specified instances can be used with the given classifier. Hence, the classifier *country* can be used with all instances of the class 'country', whereas *republic* is only appropriate for several instances of this class. Linking slots with classifiers (3) is useful for paraphrases. E.g., a rule could be formulated which specifies that all instances which have the slot 'capital' filled can be paraphrased by using a classifier which governs the instance whose slot is filled. Hence, *Paris* can be paraphrased as *the capital city of France* according to: 'Capital(France) = *Paris*' and 'classifies_slot(Capital) = *capital city*'.

No method has yet be devised to model the relations between classifiers (such as 'all instances of a

[2] Hence, the geographic entity can be thought of as the 'meaning' (or 'referent') of one or more lemmata.

[3] Containment is closely related to and almost always prerequisite for an administrative subdivision.

[4] Since multiple superclasses are allowed, the taxonomy forms a graph, not a tree. An example is the class 'canal', which is thought of as a 'waterway' and is accordingly assigned to the superclasses 'body of water' and 'traffic route'.

‘capital’ are instances of a ‘city”). A graph containing these relations would provide a second classification network next to the classes of geographic entities defined in the ontology. In contrast to our other classification it would be language specific and may show some degree of ‘fuzziness’, a phenomenon which should be avoided in ontologies for practical reasons.

3 Some Statistics and Future Development

The ontology contains 17 000 geographic entities and 21 000 lemmata (18 000 toponyms, 1 000 adjectives, 1 000 male and 1 000 female inhabitants, and 1 200 classifiers). The inflected dictionary contains 180 000 forms. Additions are continuously being made.

The ontology is being developed as part of a larger-scale project which aims to analyse German sentences with locations (including place names) as arguments. The recognition of place names in texts is one important subtask.

4 Bibliography

Axelrod, Amittai E.
2003 On building a high performance gazetteer database. *Workshop on the Analysis of Geographic References, NAACL’03.* http://www.metacarta.com/kornai/NAACL/WS9/Conf/ws910.pdf

Bauer, Gerhard
^{2}1998 [^{1}1985] *Namenkunde des Deutschen.* Germanistische Lehrbuchsammlung 21. Berlin.

BUND.DE
Städte, Kreise &Gemeinden. http://www1.bund.de/nn_518/Content/Verwaltung-in-Deutschland/SKG/SKG/SKG-knoten.html__nnn=true

Courtois, Blandine
2004 Dictionnaires électroniques *DELAF* anglais et français. In: Leclère et al. (eds.): *Lexique, syntaxe et lexique-grammaire; syntax, lexis &lexicon-grammar* 113–123.

HistOrtsbuch
Rademacher, Michael: Deutsch-österreichisches Ortsbuch 1871-1945. http://www.literad.de/geschichte/ortsbuch39.html

Intex
INTEX: an Linguistic Development Environment. http://intex.univ-fcomte.fr/

Jones, Christopher B.; A. I. Abdelmoty; G. Fu
2003 Maintaining ontologies for geographical information retrieval on the web. *Proceedings of OTM Confederated International Conferences CoopIS, DOA, and OOBASE* 934–951. http://www.geo-spirit.org/publications/SPIRIT_maintaining_ontologies.pdf

Maier-Meyer, Petra
1995 *Lexikon und automatische Lemmatisierung.* CIS-Bericht-95-84. München.

Maurel et al. [Maurel, Denis; Mickaël Tran; Duško Vitas; Thierry Grass; Agata Savary]
2004 Prolexbase : Proposition d’une ontologie multilingue des noms propres. Rapport interne du Laboratoire d’Informatique de l’Université de Tours 274. http://tln.li.univ-tours.fr/Tln_Biblio/2004RapportOntologieProlexbase.zip

Maurel, Denis; Odile Piton
1998-1999 Un dictionnaire de noms propres pour INTEX: les noms propres géographiques. In: Fairon, Cédrick (ed.): *Analyse lexicale et syntaxique: le système INTEX* 279–289.

NGA:GNS
National Geospatial-Intelligence Agency: GEOnet Names Server (GNS): Names Files of Selected Countries. http://earth-info.nima.mil/gns/html/

Piton, Odile; Denis Maurel
2001 Les noms propres géographiques et le dictionnaire Prolintex. *Quatrièmes journées Intex, Bordeaux, 11-12 juin (à paraître dans les Presses Universitaires de Franche-Comté).* http://grelis.univ-fcomte.fr/intex/downloads/Odile%20Piton.pdf

Prolintex
Laboratoire d’Informatique de l’Université de Tours: Prolintex. http://tln.li.univ-tours.fr/Tln_Prolintex.html

Protégé
The Protégé Ontology Editor and Knowledge Acquisition System. http://protege.stanford.edu/

Tran, Mickaël; Thierry Grass; Denis Maurel
2004 An ontology for multilingual treatment of proper names. *Ontologies and Lexical Resources in Distributed Environments (OntoLex 2004), in Association with LREC2004, Lisboa, Portugal, 29 may* 75–78. http://tln.li.univ-tours.fr/Tln_Biblio/2004ontolex.zip

Unitex
Unitex – Corpus Processor. http://www-igm.univ-mlv.fr/~unitex/

Wikipedia
Wikipedia – Die freie Enzyklopädie. http://de.wikipedia.org/

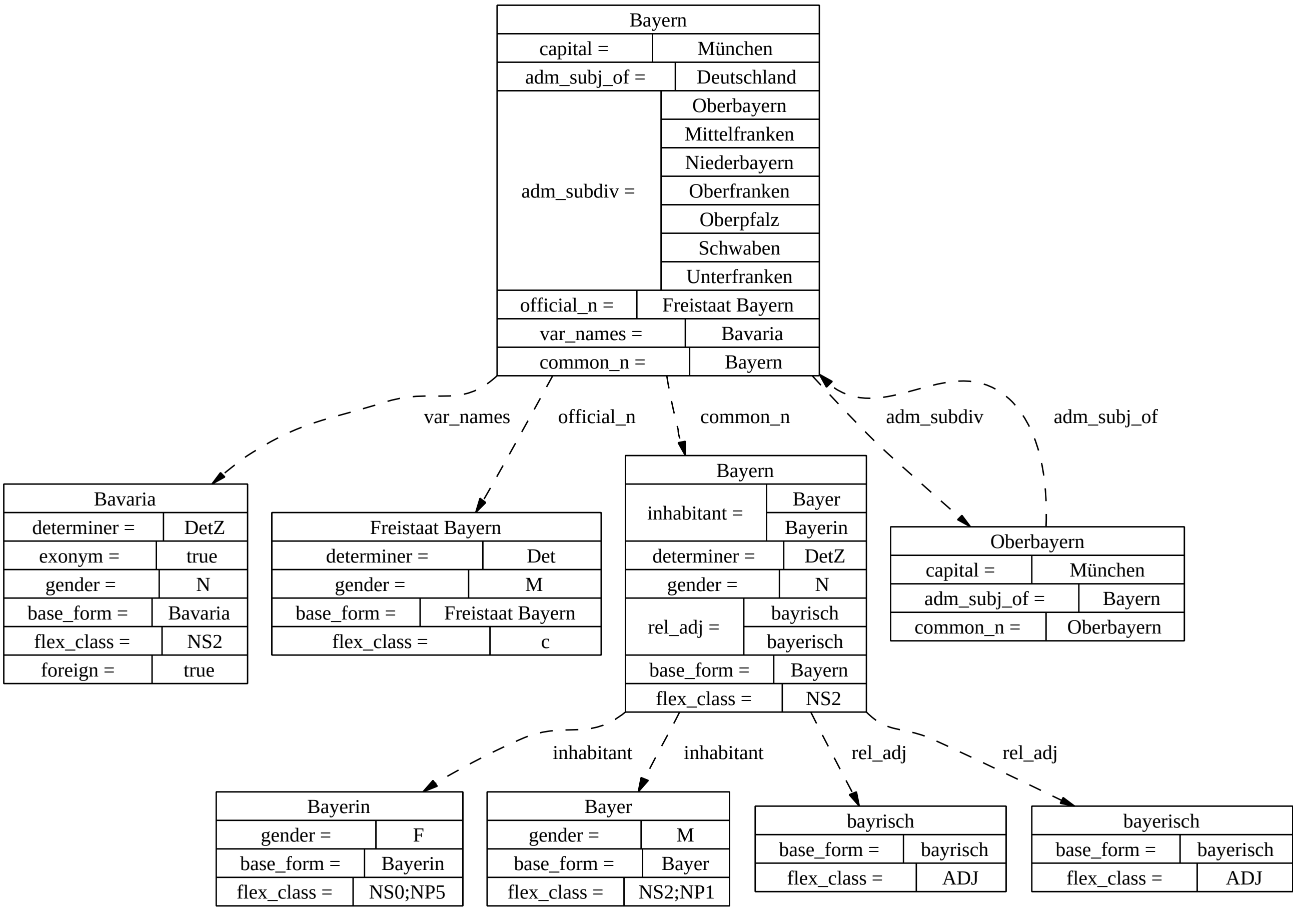

Figure 1: The geographic entity ‘Bavaria’ and selected related entities. Each box represents one instance (a geographic entity or a lemma). The cells of one instance are filled with slot names and values. Arrows indicate relations between instances.

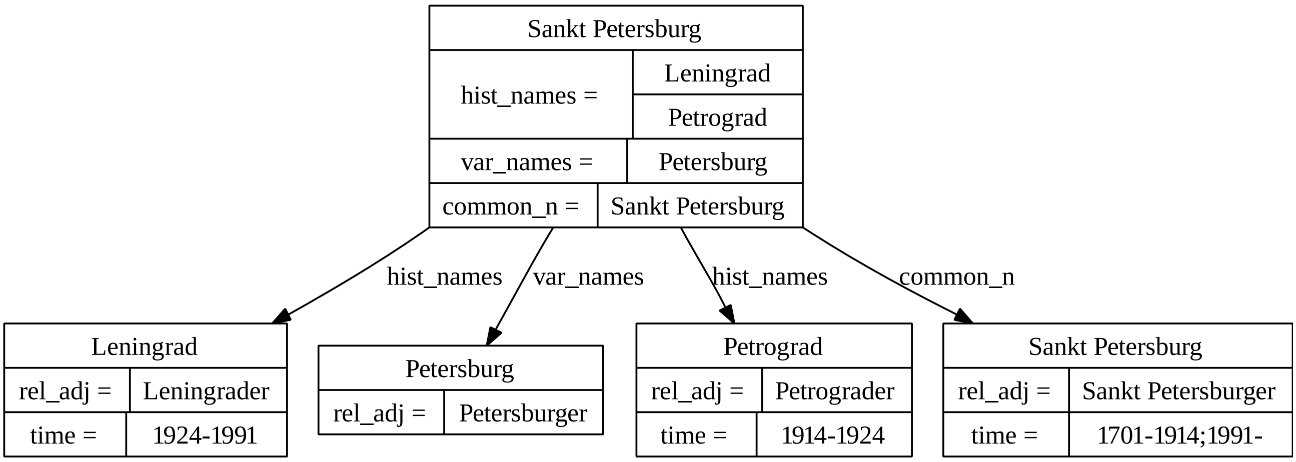

Figure 2: The geographic entity ‘Saint Peterburg’ and its (historical) names.

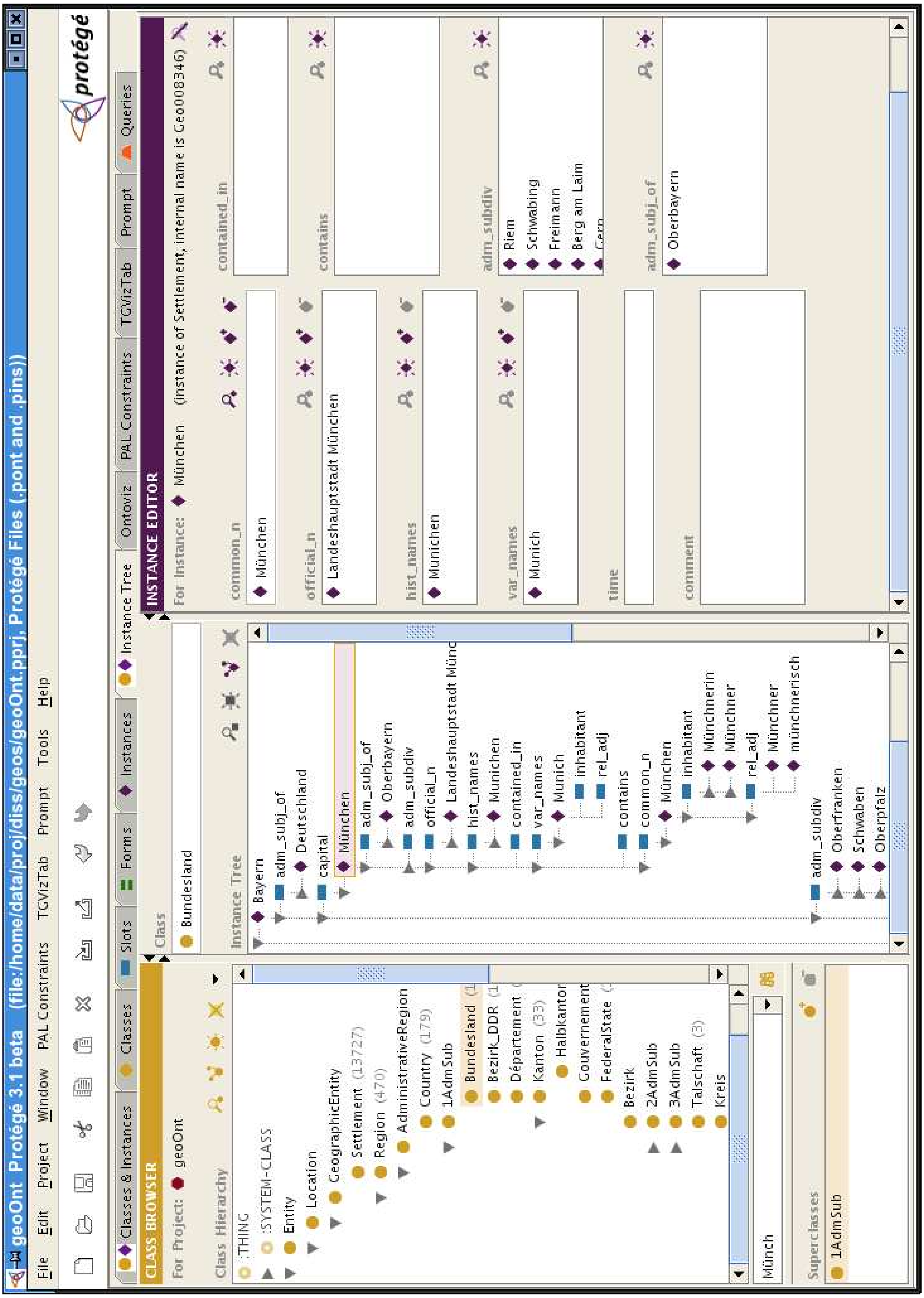

Figure 3: Screenshot of Protégé. The left window shows the class hierarchy, the middle window the relations between the geographic entity ‘Bavaria’ and other instances. The right window shows all slots of the selected instance ‘Munich’.