

Mapping Languages: Erfahrungen aus dem Projekt dbo@ema

Norbert BARTELME³, Günther FLIEDL², Marcus HASSLER², Heinrich C. MAYER², Jost NICKEL⁴, Johannes SCHOLZ³, Jürgen VÖHRINGER² und Eveline WANDL-VOGT¹

Institut für Österreichische Dialekt- und Namenlexika (I DINAMLEX)¹ / Institute für Angewandte Informatik (AINF) der Universität Klagenfurt² / Institut für Geoinformation der TU Graz³ / Forschungszentrum „Deutscher Sprachatlas“⁴

Zusammenfassung

In Österreich haben Dialekte eine große Bedeutung im täglichen Sprachgebrauch. Das „Wörterbuch der bairischen Mundarten in Österreich“ (WBÖ) ist das lexikographische Standardwerk der österreichischen Dialekte. Die digitalisierte Datensammlung wird in ein relationales Datenbanksystem übertragen und mit geographischen Daten verknüpft. Damit wird die Grundlage für ein webbasiertes Informationssystem gelegt, dessen primäres Zugriffsmedium eine interaktive Karte – ein WebGIS – ist. Darüber hinaus entsteht eine weitere Anwendung für die Administration und Eingabe von heterogenen dialektologischen Daten, die ebenfalls mit einem GIS versehen wird, um die Georeferenzierung von Belegen zu erleichtern.

1 Einleitung und Ausgangssituation

Das Projekt „Datenbank der bairischen Mundarten in Österreich electronically mapped“ (dbo@ema) hat die Zielsetzung, die umfangreiche Sammlung an Dialektdaten der Österreichischen Akademie der Wissenschaften (ÖAW) einem breiten Publikum zugänglich zu machen. Primäres Zugriffsmedium der linguistischen Daten ist eine interaktive Karte, die in einem WebGIS eingebettet ist.

Die Ausgangssituation des vom FWF finanzierten Projektes dbo@ema ist eine Sammlung analoger Belegzettel. Ein Belegzettel wird für jedes erfasste Mundartwort angelegt und enthält neben dem Lemma (Stichwort), der Lautung, dem Sammlernamen sowie der Bedeutung auch eine Angabe über den Belegort. Letzterer dokumentiert die Ortschaft oder die Gemeinde, deren Mundart der Beleg repräsentiert. Die Geschichte der Sammlung geht vom Jahr 1911 aus und ist im Laufe der Jahre auf zirka 5 Millionen Datensätze angewachsen. Historisch bedingt wurde nicht nur im heutigen Österreich erhoben, sondern auch die deutschsprachigen Gebiete der Habsburgermonarchie miteinbezogen. Eine festgelegte hierarchische Gliederung des wissenschaftlich bearbeiteten Gebietes lässt die eindeutige räumliche Zuordnung von Belegen zu. Im Zuge der Publikation des Beihefts 2 (INSTITUT FÜR ÖSTERREICHISCHE DIALEKT- UND NAMENLEXIKA 2005) zum Wörterbuch der bairischen Mundarten (WBÖ) (KRANZMAYER, E. / AB BD. 3 KOMMISSION FÜR MUNDARTKUNDE UND NAMENFORSCHUNG /AB BD. 4 INSTITUT FÜR ÖSTERREICHISCHE DIALEKT- UND

NAMENLEXIKA 1963f), wurden Kartenwerke der räumlichen Einteilung der bairischen Sprachgebiete veröffentlicht, die Grundlage für das gegenwärtige Projekt sind. Abb. 1 zeigt die bereits festgelegte hierarchische Gliederung des Bearbeitungsgebietes in Regionen (1. Ordnung unterhalb der Landesebene) und Kleinstgebiete am Beispiel des Bundeslandes Steiermark. Trotz der bereits publizierten, gut gestalteten Karten gibt es momentan an der ÖAW noch keine digitalen Geodaten, mit denen die Belege georeferenziert werden können.

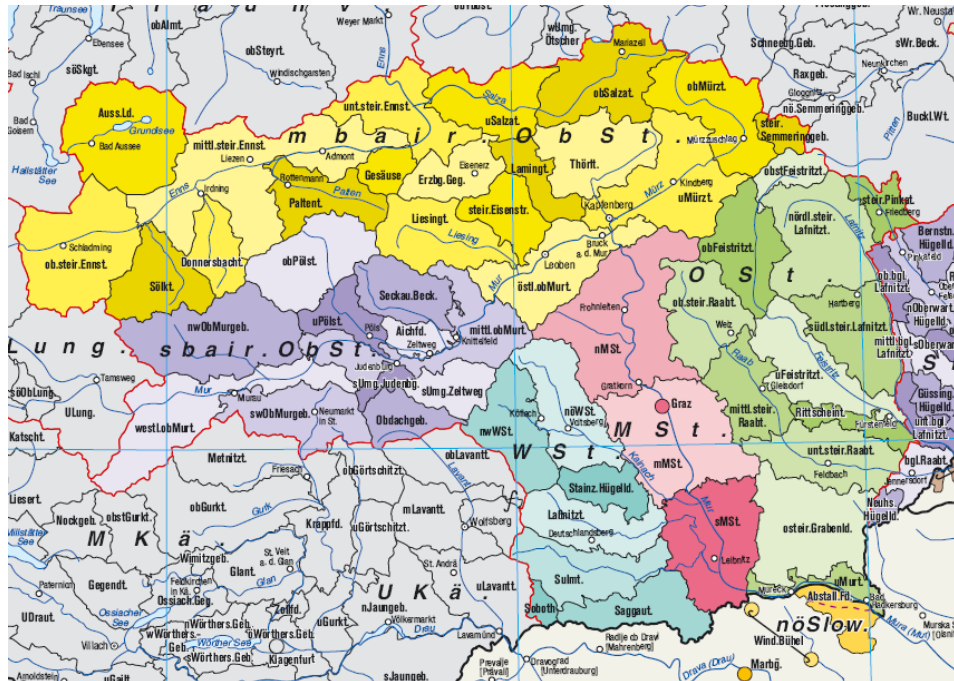


Abb. 1: Bearbeitungsgebiet in der Steiermark und im nordöstlichen Slowenien mit eingezeichneten Gebieten und Teilgebieten.

Seit 1993 erfolgt eine sukzessive Digitalisierung der Belegzettel mittels des Tübinger Systems von Textverarbeitungs-Programmen (TUSTEP). Die so erzeugten Daten sind auf die Bedürfnisse der wissenschaftlichen Bearbeitung von Textdateien zugeschnitten (ZENTRUM FÜR DATENVERARBEITUNG UNI TÜBINGEN 2008), und somit nicht direkt für die Speicherung in relationalen Datenbanken geeignet.

2 Räumliche Referenzierung der Belegdaten

Die Zuordnung der Belegzettel zu Orten erfolgt im rezenten System durch eine Reihe von Datenbanken – die sogenannten Lokalisationsdatenbanken. Diese sind im System TUSTEP abgelegt und einer laufenden Weiterentwicklung unterzogen. Die Lokalisationsdatenbanken setzen sich aus den folgenden Detaildatenbanken zusammen: die Belegorte- (ODb), die

Gemeinden- (GemDb) und die Gebietsdatenbank (GebDb). Die ODb hat die Namen sämtlicher im Quellenmaterial aufscheinender Belegorte als Inhalt, und diese werden über die GemDb einer Gemeinde bzw. über die GebDb einem für das WBÖ definierten Mundartgebiet zugeordnet (WANDL-VOGT 2006).

Die Gebietsinformationen sind hierarchisch gegliedert, wobei jedes Gebiet rekursiv in mehrere Teilgebiete zerlegt werden kann. Die Teilgebiete auf unterster Hierarchiestufe bestehen in der Regel aus Gemeinden. Die Gebietsinformationen weisen jedoch nicht einen einzigen hierarchischen Baum auf, sondern vier hierarchische Bäume:

- Administrativ: orientiert sich an den administrativen Grenzen
- Dialektologisch: hat als primäre Gliederungseinheit linguistisch-geographische Regionen, wie z.B.: Südbairisch
- WBÖ Sonderregionen: umfasst jene Regionen, die sich der dialektologischen und administrativen Hierarchie entziehen, wie z.B.: das Ennstal
- Verweisregionen: für die Verwaltung alternativer Regions-Objekte

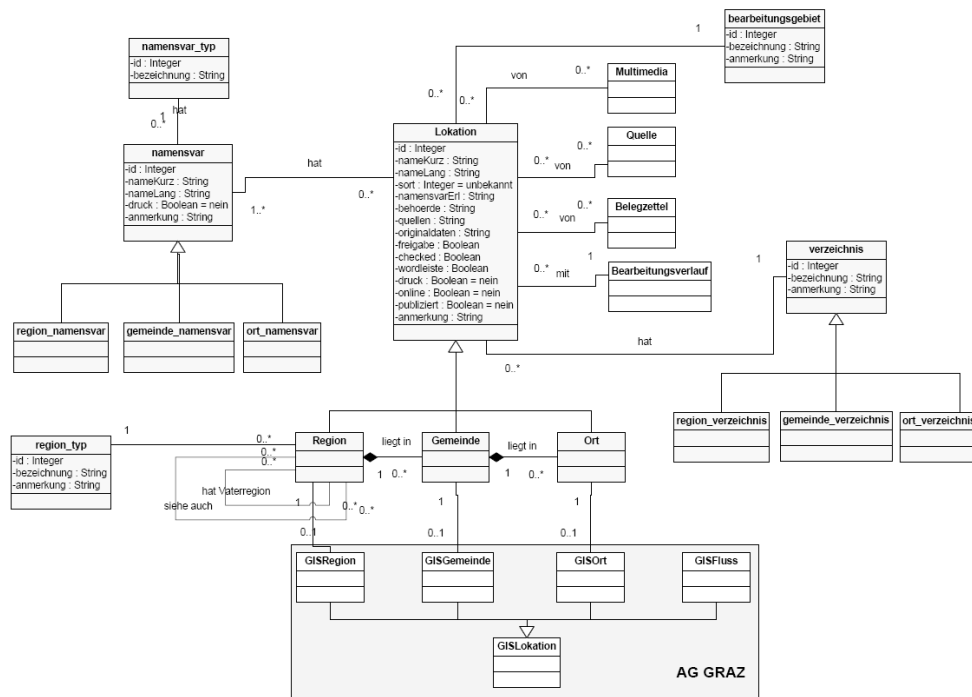


Abb. 2: ER Modell für die Speicherung der Lokalisationsdaten. Die räumlichen Komponenten sind im grauen Rechteck zusammengefasst und mit AG Graz beschriftet.

Im Zuge der Aufbereitung der Lokalisationsdaten für die digitale Speicherung sowie der nachfolgenden Präsentation in einem WebGIS wird ein relationales Datenbankmodell aufgebaut, welches die Möglichkeit der Speicherung von attributiven und räumlichen Daten ermöglicht. Ein Entity-Relationship (ER) Modell (siehe Abb. 2) zeigt den Ausschnitt des Datenbankmodells, der für die Lokalisationsdaten zuständig ist. Hierbei sind die räumlichen Komponenten innerhalb des grauen Rechtecks dargestellt, und mit AG GRAZ beschriftet. Die Daten des Systems TUSTEP werden in einem semi-automatischen Prozess in die erstellte DB-Struktur migriert und mit den räumlichen Daten verbunden. Die georeferenzierten Daten umfassen folgende Objekte: Gemeindegrenzen (Polygone), Ortschaften (Punkte) sowie Gebiete der GebDb (Polygone).

Die Abgrenzung der Gebiete erfolgt im derzeitigen System linienscharf, d.h. es existieren keine Übergangszonen zwischen den Gebieten. Da die Sprache sich nicht an genau definierte Grenzen hält, ist diese Form der Modellierung keine akkurate Repräsentation der realen Welt. Für die Modellierung von solcherart „weichem“ Datenmaterial, also die Modellierung von natürlichen Phänomenen wie Sprache, kann eine fuzzy-basierte Zuordnung von Sprachbelegen zu Gebieten der GebDb aussagekräftigere Ergebnisse liefern (WORBOYS 2000; BURROUGH & MCDONNELL 1998; SALLES & BREDWEG 2006; ZADEH 2008).

3 WebGIS zur Visualisierung von linguistischen Daten

Das so erstellte räumliche Datenbanksystem dient als Grundlage für den Aufbau eines webbasierten GIS. Dieses System dient als primäres Zugriffsmedium für die interessierte Allgemeinheit und wird weitgehend aus Open Source Komponenten aufgebaut. Im Zuge des Projektes wurde die Funktionalität des Systems festgelegt, die den möglichen Interaktionen der Benutzer entspricht. Das System besteht aus folgenden Komponenten:

- Relationales Datenbankmanagementsystem (RDBMS) mit räumlicher Erweiterung
- Mapserver mit OGC-konformen WMS und WFS Dienst
- Geo-Portal für die Visualisierung und des WMS und Abfrage des WFS Dienstes und zur Visualisierung der damit verknüpften attributiven linguistischen Daten.

Die grundlegende Systemarchitektur ist in Abb. 3 dargestellt. Die konkrete Implementierung basiert auf einer MySQL-Datenbank, dem UMN Mapserver und dem Geo-Portal Framework Mapbender. Der Mapbender wird in eine Portalseite, die grundlegende Such- und Visualisierungsfunktionen für die attributiven Daten bereithält, integriert. Dies ist aufgrund der Open Source Komponenten von Mapbender und die standardisierten Schnittstellen der eingesetzten Serversoftware problemlos möglich.

Das WebGIS wird nicht nur zur Dissemination der Forschungsergebnisse eingesetzt, sondern findet auch bei der Dateneingabe Anwendung. Um die Georeferenzierung von Belegen zu vereinfachen, wird für die Administrationsoberfläche – die für die ÖAW entwickelt wird – eine Anbindung an das WebGIS realisiert. Dadurch können z.B.: Orts – und Gemeinidenamen schneller gefunden, evaluiert und richtig referenziert werden, was zu einer Qualitäts – und Effizienzsteigerung führt.

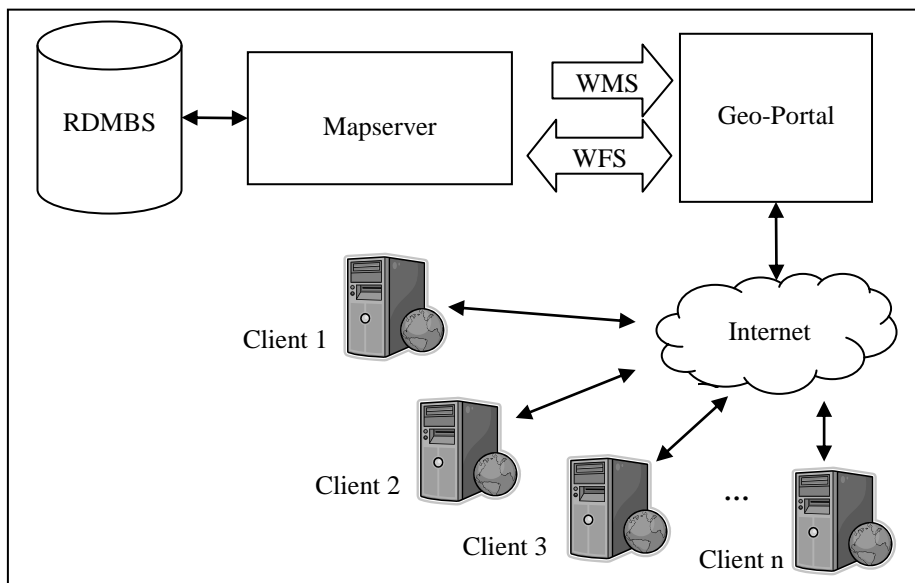


Abb. 3: Generische Systemarchitektur des WebGIS zur Visualisierung der räumlichen Daten.

4 Zusammenfassung und Ausblick

Das Ziel des interdisziplinären Forschungsprojektes ist es, dialektologische Daten mit Raumbezug der Öffentlichkeit zugänglich zu machen. Gleichzeitig wird ein effizientes Werkzeug für das Management des umfangreichen Datenbestands der ÖAW geschaffen, das ein großes Vernetzungspotential für räumliche Daten aller Art enthält und somit als Hilfsmittel für laufende, wie auch als Kristallisationskeim für künftige Projekte dienen kann.

Durch den Einsatz von interaktiven Karten als primäres Zugriffsmedium ist es möglich, komplexe Zusammenhänge aussagekräftig zu visualisieren, und somit für den interessierten Laien verständlich zu machen. Durch die weite Verbreitung der digitalen Globen, wie Microsoft Virtual Earth oder Google Earth, ist die breite Bevölkerung im Umgang mit Kartenportalen vertraut. Das senkt die Hemmschwelle, diese Applikation zu verwenden.

Großes Augenmerk wurde im Laufe des Projektes auf die Migration der bestehenden rein attributiven Datensätze in das neue relationale Datenbankschema gelegt, das eine gezielte Abfrage und Datenmanipulation erst möglich macht. Durch die Speicherung von geographischen Daten und die Verbindung dieser mit den entsprechenden attributiven Daten wurde die Grundlage für ein webbasiertes Informationssystem mit integriertem GIS gelegt. Im weiteren Verlauf des Projektes wird die Referenzimplementierung fertig gestellt und der Öffentlichkeit präsentiert werden.

Hinweis

Dieses Projekt wird vom FWF, Projektnummer: L323-G03, gefördert.

Weiterführende Informationen zum Projekt dbo@ema: <http://www.wboe.at>

(Mitarbeiter/Innen: <http://www.wboe.at/de/scientists.aspx>).

Literatur

- INSITUT FÜR ÖSTERREICHISCHE DIALEKT- UND NAMENLEXIKA (2008), Datenbank der bairischen Mundarten in Österreich (DBÖ). - Wien
- BURROUGH, P. & MCDONNELL, R. (1998), Principles of Geographical Information Systems.- New York.
- INSTITUT FÜR ÖSTERREICHISCHE DIALEKT- UND NAMENLEXIKA (2005), Wörterbuch der bairischen Mundarten in Österreich. Beiheft Nr. 2. Erläuterungen zum Wörterbuch. Lautschrift, Abkürzungsverzeichnis, Literatur- und Quellenverzeichnis, Gebietsverzeichnis, Gemeindeverzeichnis (mit einer Übersichtskarte und 6 Detailkarten zum Gebietsverzeichnis) (Bayerisch-österreichisches Wörterbuch: I Österreich).- Wien.
- KRANZMAYER, E. / AB BD. 3 KOMMISSION FÜR MUNDARTKUNDE UND NAMENFORSCHUNG /AB BD. 4 INSTITUT FÜR ÖSTERREICHISCHE DIALEKT- UND NAMENLEXIKA (1963f), Wörterbuch der bairischen Mundarten in Österreich (WBÖ).- Wien.
- SALLES, P. & BREDWEG, B. (2006), Modelling population and community dynamics with qualitative reasoning. – In: Ecological Modelling Nr. 195 (2006), S. 114-128.
- WANDL-VOGT, E. (2006), Von der Karte zum Wörterbuch. Überlegungen zu einer räumlichen Zugriffsstruktur für Dialektwörterbücher. Dargestellt am Beispiel des Wörterbuchs der bairischen Mundarten in Österreich (WBÖ).In: ELISA CORINO, CARLA MARELLO, CRISTINA ONESTI (ED.): ATTI DEL XII CONGRESSO INTERNAZIONALE DI LESSICOGRAFIA. TORINO, 6-9 Settembre 2006. European Association for Lexicography. Volume II.- Torino, S. 721-732.
- WORBOYS, M. (2000), GIS: A Computing Perspective.- London.
- ZENTRUM FÜR DATENVERARBEITUNG UNI TÜBINGEN (2008), Tübinger System von Textverarbeitungs-Programmen. – Web: <http://www.zdv.uni-tuebingen.de/tustep/> (26.01.2008).
- ZADEH, L. A. (2008), Is there a need for fuzzy logic? – In: Information Sciences (forthcoming – corrected proof), doi:10.1016/j.ins.2008.02.012.