# Geoobjekt-Identifikation und Klassifizierung via Oracle-Geocoder

**Andreas Bartels**
**Disy Informationssysteme GmbH**
**Karlsruhe**

**Schlüsselworte**
Datenbank, Spatial, LOCATOR, GEOCODER, Georeferenzierte Adressdatenbestand (GAB), Koordinatenreferenzsysteme, BETA2007, Datenmigration.

**Einleitung**
Für eine von uns durchgeführte Datenmigration mussten Geoobjekte von Gebäuden aus Teilgebieten über ganz Deutschland verteilt zusammengeführt und über ihre Adresse identifiziert und dann nach Nutzung klassifiziert werden. Sowohl die Geoobjekte als auch die Adressen und Nutzungsinformationen stammten aus unterschiedlichen Quellen und waren deshalb bezüglich den Datenformaten und Strukturen nicht homogen. Um die Geoobjekte mit den Nutzungsinformationen zu verknüpfen, stand uns der georeferenzierte Adressdatenbestand des Bundesamts für Kartografie und Geodäsie (BKG) für ganz Deutschland zur Verfügung.

In diesem Manuskript wird der Homogenisierungsprozess für die Gebäude und Nutzungsinformationen, der Aufbau eines für die Aufgabe ausreichenden Geocoder-Datenbestandes skizziert.

**Datenbasis**
Als Datenbasis des Projektes wurden uns vom Kunden, Datenbestände von 50 verschiedenen Quellen geliefert. Dieser Umstand brachte mit sich, das Daten in unterschiedlichen Datenformaten, Redundant, Teils widersprüchlich und in unbekannter Qualität verarbeitet werden mussten.

Für die Gebäude wurden uns 30.000 Dateien (ca. 90 Mio. Gebäude) geliefert. Die Geographischen Daten der Gebäude wurden zudem in 13 unterschiedlichen Koordinatenreferenzsystemen ausgeliefert die Teilweise nicht bekannt waren.

Zur Identifikation der zu Klassifizierenden Gebäude, bekamen wir überwiegend Excel und CSV Dateien mit Adressen, aus 30 unterschiedlichen Quelle. Diese waren unterschiedlich Strukturiert, teilweise sogar ohne Struktur, so das die Bestandteile einer Adresse nur durch parsen ermittelt werden konnten.

Wie oben schon aufgeführt wurde uns der Georeferenzierte Adressdatenbestand des Bundesamts für Kartografie und Geodäsie (GAB) zum verknüpfen der Geoobjekte mit den Sachdaten geliefert. Dieses sind etwa 30 Mio. Datensätze. Die zwar in einem einheitlichen Datenformat ausgeliefert werden. Da die Datensätze aber aus 15 Bundesländern und über 4 unterschiedliche Verfahren erstellt werden. Liegen auch hier redundante Daten mit unterschiedlicher Qualität vor. Die Qualität der Daten wird aber in den Daten durch die Angabe des Ermittlungsverfahrens bestimmt.

**Datenmigration**

Unsere Basis Strategie bei Datenmigration, in unterschiedlich Formaten, ist die zusammen Führung dieser in einer Datenbank. In allen von uns durchgeführten Projekten haben wir dabei eine Oracle

Datenbank eingesetzt. Die Daten werden im ersten Schritt soweit möglich eins zu eins übernommen. Nur da wo es notwendig bzw. nicht verlustbehaftet ist, wie z.B. bei Geometrien, transformieren wir die Daten in das von der Datenbank vorgegebene Format.

Mit den folgenden Schritten werden die Daten Validiert, Homogenisiert und dann Zusammen geführt. Bei jedem Schritt werden Herkunft und aktueller Zustand der Daten protokolliert. Zusätzlich wird Protokolliert ob bei dem Verarbeitungsschritt ein Fehler aufgetreten ist und welcher.

Durch den Import der Daten, in die Datenbank kann deren Funktionalität für die oben aufgeführten Schritte verwendet werden. Dieses sind nicht nur die SQL-Funktionen, Skalierbarkeit und auch das verschieben bzw. Verteilen von Daten auf andere Rechner oder Speicher ist bei solchen Projekten von Bedeutung.

Die Protokollierung ermöglicht es uns, mit Fehlerfreien Daten die Verarbeitung fortzuführen, während wir uns Parallel dazu, um die Bereinigung der Fehler kümmern und deren Weiterverarbeitung danach fortführen.

Für das gesamte Projekt kamen wir auf ungefähr 250 Verarbeitungsschritte. Das hier skizzierte Teilprojekt benötigte ungefähr 100 davon.

**Zusammenführung der Geographische Gebäudedaten**

Zur Zusammenführung der Geographischen Gebäudedaten mussten diese wie oben beschrieben erst in die Datenbank importiert werden. Nach dem Import wurden die Daten Validiert und soweit möglich automatisch Korrigiert. Bei den Geographischen Daten wurden in diesem Schritt noch die Stützpunkte ausgedünnt. Danach wurden die Daten in eine einheitliche Tabellenstruktur überführt. Um die Daten zusammenführen zu können, mussten die Geographischen Daten noch in ein einheitliches Koordinatenreferenzsystem überführt und Redundante Daten aussortiert werden.

Um die Ansprüche des Kunden an die Genauigkeit der Koordinaten erfüllen zu können musste die Transformation mittels des BETA2007 Gridshiftverfahrens durchgeführt werden. Dieses konnten wir in der Datenbank mittels Spatial-Funktionalität durchführen. Voraussetzung dafür war allerdings eine Datenbank der Version 11GR2.

Das Aussortieren der Redundanten Gebäudedatensätze erfolgte anhand, der bei der Validierung bestimmten Qualität der Daten, nach einer vom Kunden vorgegebenen Priorisierung. Hierbei wurden im ersten Schritt nicht einzelne Datensätze gelöscht. Wenn die Qualität der Daten einer Quelle zu schlecht war wurde sie gegen die der niedriger priorisierte Quelle ausgetauscht. In einem zweiten Schritt wurden dann die Redundanten Datensätze in den überlappenden Randbereichen der ausgewählten Daten, nach der gleichen Priorisierungsverfahren aussortiert.

**Erstellung der Klassifikationsadressenlisten**

Zur Erstellung der Klassifikationsadressenlisten wurde nach dem oben beschriebenen Verfahren, pro Klassifikationstype eine Adressdatentabelle erzeugt. Dabei mussten wir zusätzlich Adressen mit Hausnummerintervallen aufteilen. Anhand der Herkunft der Quelle, war es uns noch möglich die Adresse um das Bundesland zu erweitern, was für die spätere Nutzung des Geocoder sinnvoll ist.

**Klassifizierung**

Bei der Klassifizierung mussten wir, anhand der Adressen in den Klassifikationsaddressenlisten die Geobjekte in dem migrierten Gebäudedaten selektieren und in der dafür vorgegeben Spalte Kategorisieren. Da die Gebäudedaten keine Adressdaten beinhalteten, musste dieses durch räumliche

Selektion der Geoobjekte erfolgen. Zu diesem Zweck, wurde uns vom Kunden, der georeferenzierte Adressdatenbestand (GAB) des Bundesamts für Kartografie und Geodäsie (BKG) für ganz Deutschland zur Verfügung gestellt.

| | | | | | |
|---|---|---|---|---|---|
| 2110766825 | 21 | Hasensprung | 10318 | Berlin | Karlshorst |
| 2110766824 | 17 | Hasensprung | 10318 | Berlin | Karlshorst |
| 2110766823 | 11 | Hasensprung | 10318 | Berlin | Karlshorst |
| 2110766822 | 8 | Hasensprung | 10318 | Berlin | Karlshorst |
| 2110766819 | 1 | Hasensprung | 10318 | Berlin | Karlshorst |
| 2110863678 | 40 | Kol.Frieden Neuer Weg | 12099 | Berlin | Westend |
| 2110863677 | 38 | Kol.Frieden Neuer Weg | 12099 | Berlin | Westend |
| 2110863676 | 31 | Kol.Frieden Neuer Weg | 12099 | Berlin | Westend |
| 2110863675 | 13 | Kol.Frieden Neuer Weg | 12099 | Berlin | Westend |
| 2110863674 | 12 | Kol.Frieden Neuer Weg | 12099 | Berlin | Westend |
| 2110863673 | 11 | Kol.Frieden Neuer Weg | 12099 | Berlin | Westend |
| 2110850781 | 33 | Kol.Frieden Neuer Weg | 12099 | Berlin | Westend |
| 2110850773 | 25 | Kol.Frieden Neuer Weg | 12099 | Berlin | Westend |
| 2110850758 | 10 | Kol.Frieden Neuer Weg | 12099 | Berlin | Westend |
| 2110850757 | 9 | Kol.Frieden Neuer Weg | 12099 | Berlin | Westend |
| 2110850756 | 8 | Kol.Frieden Neuer Weg | 12099 | Berlin | Westend |
| 2110826078 | 21 | Grabbeallee | 13156 | Berlin | Niederschönhausen |
| 2110826079 | 27 | Grabbeallee | 13156 | Berlin | Niederschönhausen |
| 2110826081 | 30 | Grabbeallee | 13156 | Berlin | Niederschönhausen |
| 2110826085 | 43 | Grabbeallee | 13156 | Berlin | Niederschönhausen |
| 2110826086 | 66 | Grabbeallee | 13156 | Berlin | Niederschönhausen |
| 2110826087 | 80 | Grabbeallee | 13156 | Berlin | Niederschönhausen |
| 2110861414 | 53 | Grabbeallee | 13156 | Berlin | Niederschönhausen |
| 2110850548 | 39 | Kol. Eichtal | 14050 | Berlin | Charlottenburg |
| 2110850547 | 38 | Kol. Eichtal | 14050 | Berlin | Charlottenburg |
| 2110850546 | 36 | Kol. Eichtal | 14050 | Berlin | Charlottenburg |
| 2110850545 | 35 | Kol. Eichtal | 14050 | Berlin | Charlottenburg |
| 2110850544 | 34 | Kol. Eichtal | 14050 | Berlin | Charlottenburg |
| 2110850543 | 31 | Kol. Eichtal | 14050 | Berlin | Charlottenburg |
| 2110850542 | 30 | Kol. Eichtal | 14050 | Berlin | Charlottenburg |

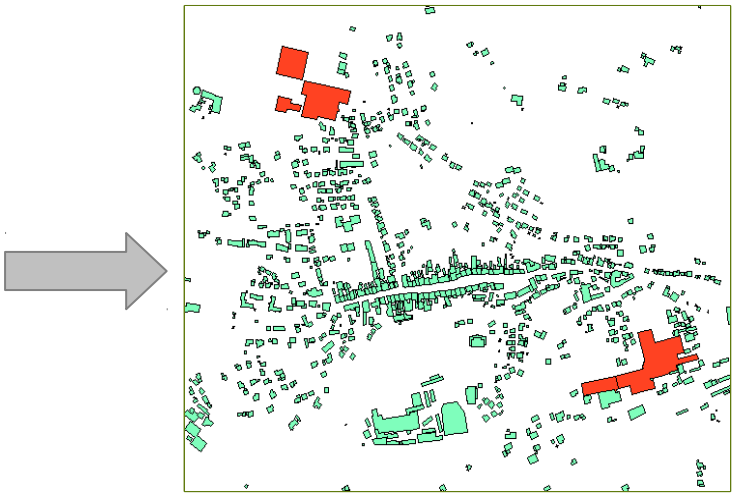

*Abb. 1: Räumliche Gebäudeklassifizierung anhand von Adressen*

Leider ist es nicht möglich mittels folgenden SQL-Statment das entsprechende Gebäude zu selektieren.

```
SELECT a.id FROM GEBAEUDE a, GAB b, ADRESSE c
            WHERE b.STREET = c.STREET
              AND b.HSNR = c.HSNR
              AND b.PLZ = c.PLZ
              AND b.CITY = c.CITY
              AND SDO_RELATE(a.GEOMETRY,
                             SDO_GEOMETRY(2001, NULL,
                             SDO_POINT_TYPE(b.X, b.Y, NULL),
                             NULL, NULL),
                  'mask=anyinteract') = 'TRUE';
```

Folgende Gründe sprechen dagegen:

1. Rechtschreibfehler in den Datenbeständen
2. Unterschiedliche Schreibweisen von Straßennamensergänzungen (z.B. Straße, Strasse, Str.)
3. die Adresse ist nicht eindeutig (z.B. wegen fehlender PLZ)
4. die Koordinate aus dem GAB-Daten liegt nicht im Gebäude sondern im zugehörigen Flurstück oder sogar auf der Straße vor dem Gebäude.

Um Probleme dieser Art zu lösen, bietet die Spatial-Erweiterung von Oracle das Paket Geocoder an. Das wir in dem Projekt verwendet haben.

Mittels des Geocoder Paketes, konnten wir die Probleme 1-3 lösen. Für das 4-te Problem mussten wir einen eigenen Mechanismus entwickeln.

**Geocoder**

Der Oracle Geocoder ist nur eine Interpretierende Software, die Daten die benötigt werden können bei

mehreren Anbietern erworben werden. In dem Projekt haben wir die notwendigen Daten aus den GAB-Daten generiert. Der Geocoder bietet neben der Räumlichen Bestimmung von Adressdaten auch noch Routing Funktionalität an. Da diese Funktionalität für uns nicht von Interesse war, haben wir die dafür notwendigen Daten nicht generiert.

Die für uns wichtigen Daten sind die in den Tabellen GC_AREA_DE, GC_ROAD_DE, GC_ROAD_SEGMENT_DE, GC_ADDRESS_POINT_DE und GC_POSTAL_CODE_DE

Die Tabelle GC_POSTAL_CODE_DE ermöglicht die Gruppierung der Adressen die zu einer Postleitzahl gehören.

Die Tabelle GC_AREA_DE ermöglicht die Gruppierung von Adressen nach Verwaltungseinheiten. Unterstützt werden 4 Gliederungsstufen, Staat, Länder, Gemeinden und Ortsteile. Im Projekt haben wir auf die Ortsteile verzichtet.

Bei den beiden Tabellen ist zu beachten das sie als Geometrien nicht den Umriss der Flächen halten sondern den Centroid der Fläche. Diese Koordinate ist dann das Ergebnis des Geocoder, wenn z.B. nur die PLZ oder der Gemeindename aufgelöst werden konnte. Bei mehrdeutigen Gemeindenamen wird je nach Art der Anfrage, der Centroid über alle Treffer für das Ergebnis gebildet.

Die Tabellen GC_ROAD_DE und GC_ROAD_SEGMENT_DE ermöglichen die Gruppierung der Adressen nach Straßennamen. Die Tabelle GC_ROAD_SEGMENT_DE in Kombination mit der oben nicht aufgeführten Tabelle GC_INTERSECTION_DE bilden die Grundlage für die Routing Funktionalität.

Die Tabelle GC_ADDRESS_POINT_DE beinhaltet alle Adressen mit zugehöriger Koordinate. Sie beinhaltet zudem noch weitere räumliche Informationen wie Straßenseite und Relative Position im Straßenabschnitt.

Nach dem die Daten aufgebaut sind muss der Geocoder noch konfiguriert werden. Dafür müssen die beiden vorkonfigurierten Tabellen GC_PARSER_PROFILES und GC_PARSER_PROFILEAFS angepasst werden.

Um einen für unsere Anforderungen ausreichend Funktionsfähigen Geocoder in einer Oracle Datenbank der Version 11GR2 aufbauen zu können, mussten wir zudem noch zwei Patsches einspielen.

Probleme hatten wir mit Datensätzen deren Hausnummern einen Zusatz hatten (z.B. 16A) und Straßennamen mit den Zeichen ÄÖÜäöüß. Das Problem mit den Sonderzeichen konnte durch die Patches und das ändern der Tabelle GC_PARSER_PROFILES gelöst werden. Das Problem mit den Hausnummerzusatzzeichen haben wir nicht gelöst. Da wir aber nur sehr wenig Adressen mit Hausnummerzusatzzeichen auflösen mussten, haben wir nicht viel Energie in die Lösung investiert.

Bei benutzen des Geocoder ist zu beachten, das er Ergebnisse interpoliert, wenn er keinen direkten Treffer ermitteln kann. Die Qualität des Ergebnisses kann mit Hilfe der Werte Matchcode, Matchvector und Errormessage beurteilt werden, die im Ergebnis mit geliefert.

```
select b.STREETNAME, b.HOUSENUMBER, b.POSTALCODE, b.SETTLEMENT,
b.MUNICIPALITY, b.REGION, b.COUNTRY, b.MATCHCODE, b.MATCHVECTOR,
b.ERRORMESSAGE
from (select SDO_GCDR.GEOCODE_ALL(user, SDO_KEYWORDARRAY('AMSELWEG 10',
'09244 LICHTENAU'), 'DE', 'DEFAULT') adress from dual) a, table(a.adress)
b;
```

Die Anfrage liefert folgendes Ergebnis:

```
AMSELWEG, 10, 09244, LICHTENAU, SACHSEN, DE, 1, ??0101010010??
404?, ??X?#ENUT?B281CP?
```

Am Matchcode = 1 lässt sich in diesem Fall ablesen, das alle Angaben mit einem Datensatz über einstimmen. Anhand des Matchvector und der Errormessage kann man dann ermitteln welche Angaben gepasst haben bzw. nicht.

**Fazit**

Zwar mit ein bisschen Glück, die Benötigten Patches der Datenbank für die Koordinatentransformation und den Geocoder waren zu Beginn des Projektes verfügbar, und auch Unterstützung von Oracle konnten wir dieses Projekt erfolgreich mit unser Strategie durchführen. Es hat sich aber auch gezeigt das man oft nur mit Ausreichend Zeit zum Experimentieren und kompetenten Ansprechpartnern zum Ziel kommt. Da z. B. für den Geocoder zwar die Verwendung und die einzelnen Strukturen gut Dokumentiert sind. Um aber die Datenbasis für den Geocoder aufzubauen Fehlen entsprechende Beispiele und wichtige hinweise auf mögliche Fallstricke.

**Kontaktadresse:**
Andreas Bartels
Disy Informationssysteme GmbH
Erbprinzenstraße 4-12
D-76133 Karlsruhe

Telefon: +49 (0) 721-1600 6213
Fax: +49 (0) 721-1600 605
E-Mail andreas.bartels@disy.net
Internet: www.disy.net