I-KNOW 2002
Graz, Austria, July 11-12, 2002

Automated Retrieval of Information in the Internet by Using Thesauri and Gazetteers as Knowledge Sources

Prof. Dr. Wolf-Fritz Riekert
Fachhochschule Stuttgart – Hochschule der Medien (HdM)
University of Applied Sciences Stuttgart – School of Media

mailto:riekert@hdm-stuttgart.de
http://v.hdm-stuttgart.de/~riekert

CONTEXT

Concept and prototype development under commission of the German Federal Environment Agency, Berlin (1996-1998) by:

- Research Institute of Applied Knowledge Processing (FAW), Ulm (Riekert, Wiest, Fuchs, Klingler)
- Condat AG, Berlin (Nouhuys, formerly: CAdMAp GmbH)

The concepts described served as an input into the following German environmental information systems:

- German Environmental Information Network (GEIN)
- Geographical Information System Environment (GISU)

Software development and maintenance now:

- Ernst Basler + Partner (GISU)
- Sema Group (GEIN)

INFORMATION RESOURCES IN THE INTERNET

Categories of information resources:

- multimedia documents
- data
- application services

Supply exploding

- Problem: orientation (“lost in hyperspace”)
- powerful search tools required

SEARCH ENGINES

Search engines are based on a full text index which intentionally covers the whole Web

- Retrieval via Web browser (string search)
- Index maintained by “robots” “crawling” along hyperlinks
- No additional efforts required from information suppliers

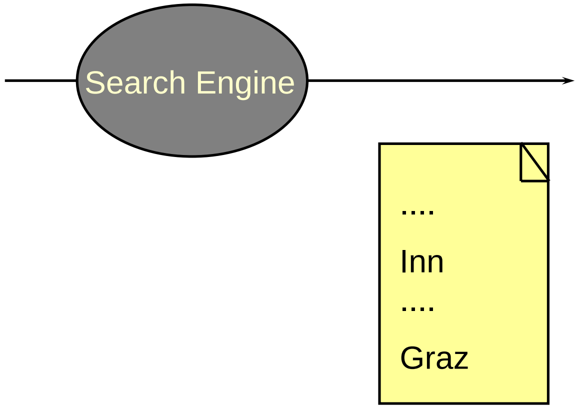
But:

- Search terms are interpreted only textually
- No semantic interpretation
- Full text index can only be used for textual resources

EXAMPLE

Query:

“Accommodation, Styria”



METAINFORMATION SYSTEMS

Metainformation systems support semantic criteria for indexing and retrieval:

- Thematic references (e.g., “Accommodation”)
- Spatial references (e.g., “Styria”)
- Temporal references (e.g., “July 11-12, 2002”)

Indexing (i.e., entering the metainformation) is done manually by the system administrator or information suppliers:

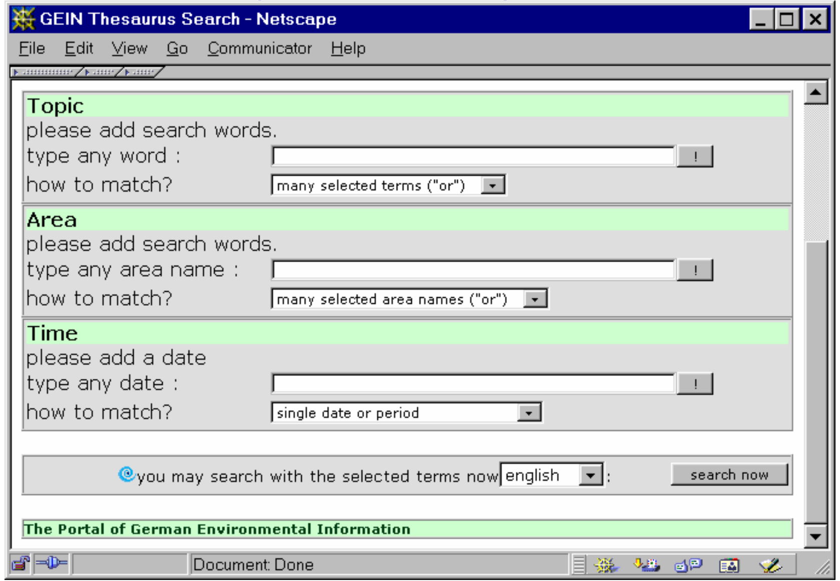
- Higher information quality (compared to search engines)
- Higher workload imposed on system administrator or information suppliers

EXAMPLES OF METAINFORMATION SYSTEMS

Examples of metainformation systems (taken from the environmental domain):

- GEIN: German Environmental Information System (Germany)
- GISU (Meta Component): Geographic Information System Environment (German Federal Environment Agency)
- UDK: Environmental Data Catalogue (Germany, Austria)
- CDS: Environmental Catalogue of Data Sources (European Environmental Agency)
- NGSC: National Geospatial Clearinghouse (USA)

METAINFORMATION SYSTEMS: EXAMPLE GEIN (WWW.GEIN.DE)


GEIN Thesaurus Search - Netscape
File Edit View Go Communicator Help
!
!
!
search now
Document: Done

Topic
please add search words.
type any word :
how to match? many selected terms ("or")

Area
please add search words.
type any area name :
how to match? many selected area names ("or")

Time
please add a date
type any date :
how to match? single date or period

you may search with the selected terms now english :

The Portal of German Environmental Information

SPECIFICATION AND PROCESSING OF SEMANTIC CRITERIA

Requirements

- Vocabulary for the specification of thematic, spatial and temporal references of information resources
- Techniques for the automated processing of thematic, spatial and temporal references

Approach

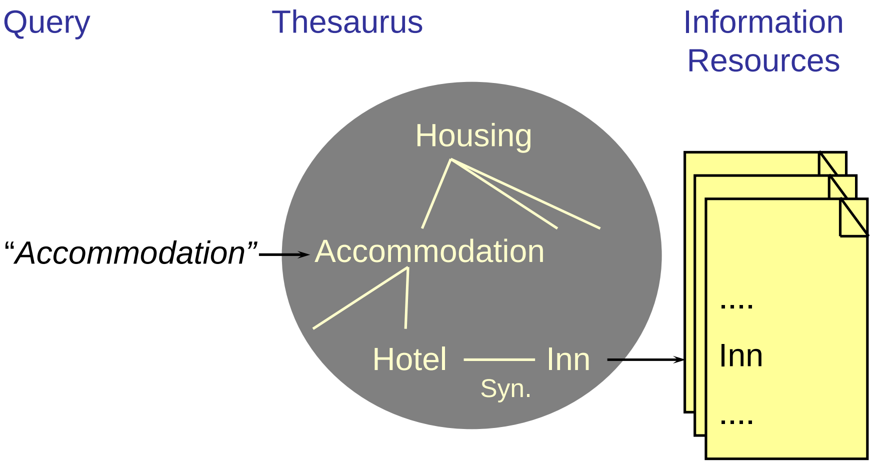
- Thesaurus to support specification and processing of thematic references
- analogously: „Gazetteer“ to support specification and processing of spatial references
- Handling of temporal references: relatively easy, not an issue of this talk

THESAURUS

A Thesaurus is a structured collection of terms with the following properties:

- Terms provide a controlled vocabulary for the specification of thematic references,
- Terms can be used for both indexing and retrieval.
- Terms are more than simple keywords.
- Terms form a semantic network established by:
  - ⇨ synonym relationship (inn - hotel)
  - ⇨ generalization hierarchy of broader / narrower terms (accommodation - hotel)
  - ⇨ linkage via related terms (accommodation - tourism)

THESAURUS-SUPPORTED QUERY PROCESSING



BLACK BOX SEARCH PROBLEM: A THESAURUS CAN HELP

Problem: Information resources are searched for by using a form in most metainformation systems (“black box search”)

- It is not clear which level of detail is required while specifying a query
  - ⇨ Many casual users dislike form-based search interfaces

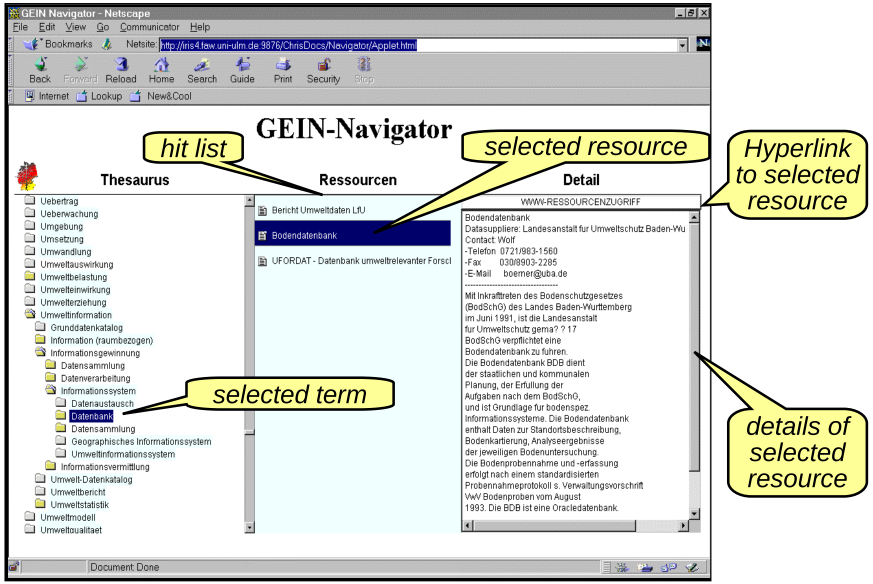
Requirement: Hierarchical directories to access the information resources

- However: Manual maintenance of hierarchical directories very time-consuming

Solution: Use a thesaurus for the automated generation of a hierarchical directory

Example: GEIN Navigator (prototype developed at FAW Ulm)

PROTOTYPICAL GENERATION OF A HIERARCHICAL DIRECTORY


GEIN Navigator - Netscape
File Edit View Go Communicator Help
Bookmarks
Netsite: http://iris4.faw.uni-ulm.de:9876/ChrisDocs/Navigator/Applet.html
Back Forward Reload Home Search Guide Print Security Stop
Internet Lookup New&Cool
Document: Done

GEIN-Navigator

hit list

selected resource

Hyperlink to selected resource

selected term

details of selected resource

Thesaurus

- Uebertrag
- Ueberwachung
- Umgebung
- Umsetzung
- Umwandlung
- Umweltauswirkung
- Umweltbelastung
- Umwelteinwirkung
- Umwelterziehung
- Umweltinformation
  - Grunddatenkatalog
  - Information (raumbezogen)
  - Informationsgewinnung
    - Datensammlung
    - Datenverarbeitung
    - Informationssystem
      - Datenaustausch
      - Datenbank
      - Datensammlung
      - Geographisches Informationssystem
      - Umweltinformationssystem
    - Informationsvermittlung
  - Umwelt-Datenkatalog
  - Umweltbericht
  - Umweltstatistik
- Umweltmodell
- Umweltqualitaet

Ressourcen

- Bericht Umweltdaten LfU
- Bodendatenbank
- UFORDAT - Datenbank umweltrelevanter Forscl

Detail

WWW-RESSOURCENZUGRIFF

```
Bodendatenbank
Datasuppliere: Landesanstalt fur Umweltschutz Baden-Wu
Contact: Wolf
-Telefon  0721/983-1560
-Fax      030/8903-2285
-E-Mail    boerner@uba.de
---------------------------------
Mit Inkrafttreten des Bodenschutzgesetzes
(BodSchG) des Landes Baden-Wurttemberg
im Juni 1991, ist die Landesanstalt
fur Umweltschutz gema? ? 17
BodSchG verpflichtet eine
Bodendatenbank zu fuhren.
Die Bodendatenbank BDB dient
der staatlichen und kommunalen
Planung, der Erfullung der
Aufgaben nach dem BodSchG,
und ist Grundlage fur bodenspez.
Informationssysteme. Die Bodendatenbank
enthalt Daten zur Standortsbeschreibung,
Bodenkartierung, Analyseergebnisse
der jeweiligen Bodenuntersuchung.
Die Bodenprobennahme und -erfassung
erfolgt nach einem standardisierten
Probennahmeprotokoll s. Verwaltungsvorschrift
VwV Bodenproben vom August
1993. Die BDB ist eine Oracledatenbank.
```

A PROCEDURE TO GENERATE A HIERARCHICAL DIRECTORY

- Create a “weeded” thesaurus consisting of all relevant terms, i.e.:
  - ⇨ take all terms used as an index for existing information resources,
  - ⇨ add recursively all broader terms,
  - ⇨ disregard all other terms
- Display thesaurus in a hierarchical presentation (Windows Explorer-like), starting from “toplevel terms”
- Special highlighting indicates which terms
  - ⇨ directly lead to hits,
  - ⇨ possess narrower terms leading to hits
- Provide navigation paths to the metainformation records and from there to the original information resources

METAINFORMATION SYSTEMS VS. SEARCH ENGINES

Metainformation system:

- Easy retrieval by using semantical criteria
- But: Indexing very expensive for administrators or information suppliers

Search engine:

- Indexing very easy, no work imposed on suppliers
- But: only textual processing of search criteria

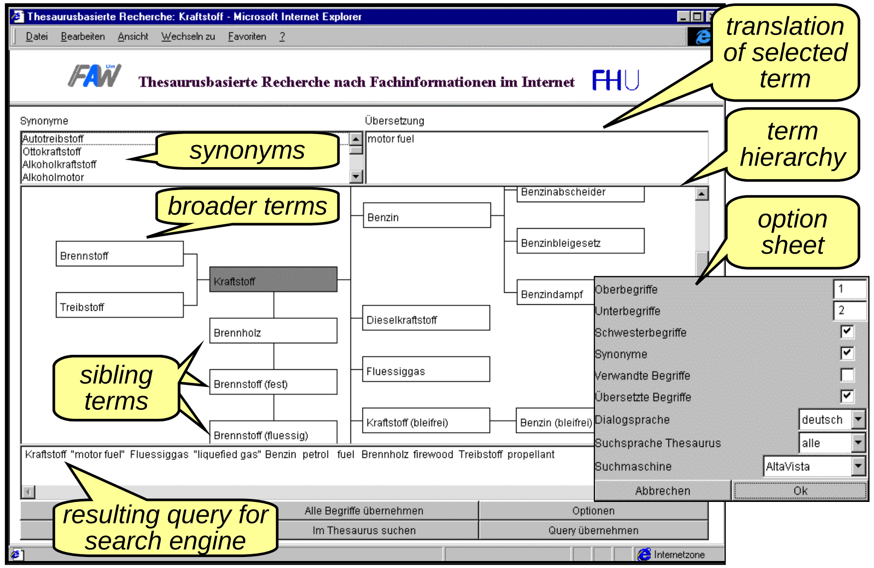
Synthesis:

- Combination of the advantages of search engines and metainformation systems: Thesaurus-based preprocessor for search engines

COMBINE THE ADVANTAGES

	Indexing inexpensive	Semantic processing of search terms
search engine	✗	—
metainformation system	—	✗
search engine with thesaurus-based preprocessor	✗	✗

THESAURUS-BASED PREPROCESSOR FOR SEARCH ENGINES



Thesaurusbasierte Recherche: Kraftstoff - Microsoft Internet Explorer

Thesaurusbasierte Recherche nach Fachinformationen im Internet

Synonyme: Autotreibstoff, Ottokraftstoff, Alkoholkraftstoff, Alkoholmotor

Übersetzung: motor fuel

Oberbegriffe: 1
Unterbegriffe: 2
Schwesterbegriffe ☑
Synonyme ☑
Verwandte Begriffe ☐
Übersetzte Begriffe ☑
Dialogsprache: deutsch
Suchsprache Thesaurus: alle
Suchmaschine: AltaVista

Kraftstoff "motor fuel" Fluessiggas "liquefied gas" Benzin petrol fuel Brennholz firewood Treibstoff propellant

translation of selected term

synonyms

term hierarchy

broader terms

option sheet

sibling terms

resulting query for search engine

SPATIAL REFERENCES IN TRADITIONAL SYSTEMS

Problem: Spatial references in traditional systems are handled very poorly (if they are handled at all):

- Rigid vocabulary
  - ⇨ Usually only one single spatial reference system supported (coordinates only, names only)
- No intelligence
  - ⇨ It cannot be recognized if one region encloses another

Solution: Specification of spatial references through geographic objects (geoobjects)

- Geoobjects are more than names or coordinates
- They possess both names and coordinates
- Geometrical and topological relationships can be computed

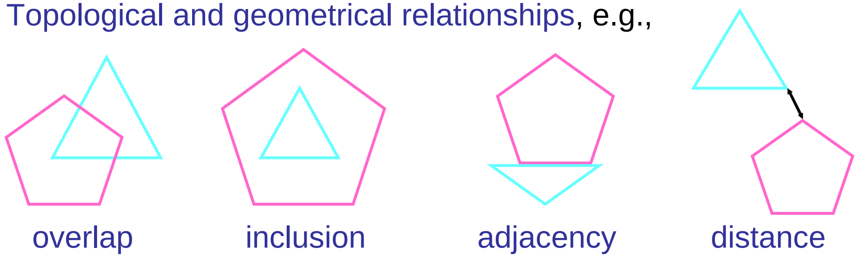
GAZETTEER

A Gazetteer is a structured collection of geographic data objects (geoobjects)

- Geoobjects can be used to specify spatial references
- Spatial reference: n:m-relationship between information resources and geoobjects in the gazetteer
- Geoobjects may possess the following properties:
  - ⇨ name (e.g., “Styria”)
  - ⇨ geometry (e.g., coordinates describing a polygon)
  - ⇨ type (e.g., “state”)
  - ⇨ unique identifier (e.g., administrational code)
  - ⇨ optional: hierarchy (e.g., administrational hierarchy)
  - ⇨ optional: synonymous names, translated names

TOPOLOGICAL AND GEOMETRICAL RELATIONSHIPS

Topological and geometrical relationships, e.g.,

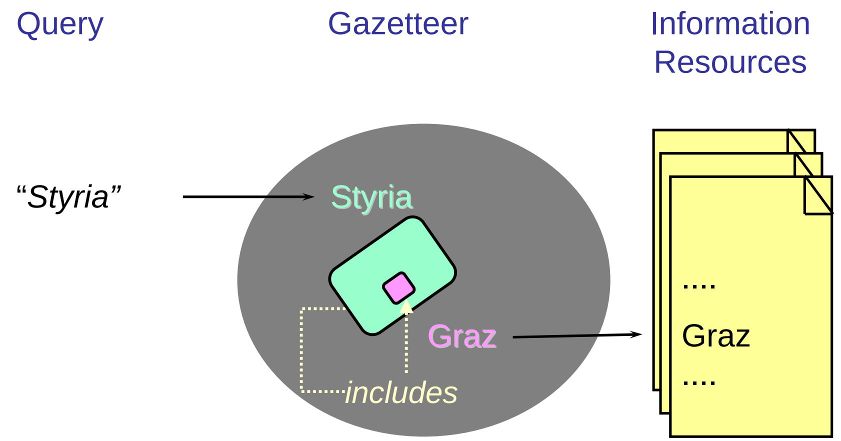


can be derived from geometry (i.e., coordinates)

With the help of these relationships,
a flexible geographic vocabulary can be used for

- indexing and
- retrieval purposes

GAZETTEER-BASED RETRIEVAL



GAZETTEER: APPLICATIONS

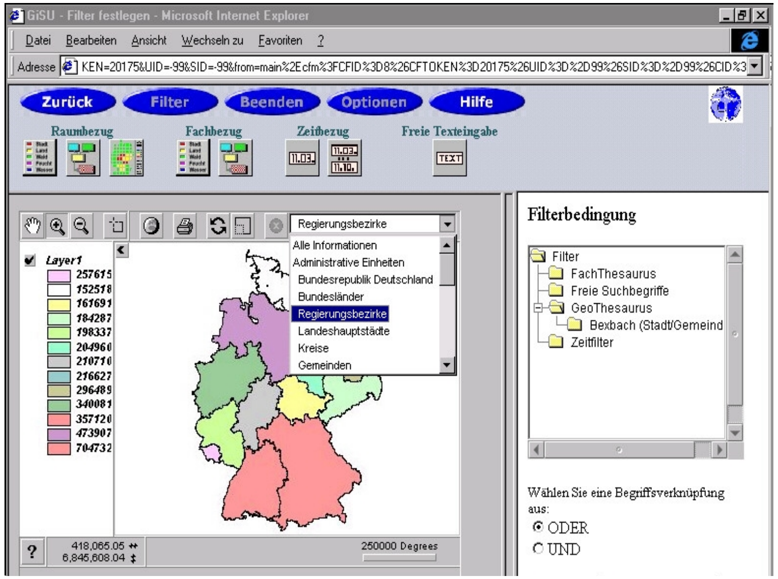
- Browser for geographical names
- Specification of spatial references on a cartographic interface
- Automated extension of queries: information resources in a certain geographic or topological neighborhood can be included into the scope of the query
- Easy transition between textual and geometrical representation of spatial references
- Text analysis for the automated spatial indexing of textual information resources
- Graphical display of spatial references as "footprints", e.g. to visualize a search result

THE GERMAN "GEOTHESAURUS": AN EXAMPLE FOR A GAZETTEER

The German Federal Environment Agency developed a Gazetteer known as "Geothesaurus"

- It contains about 100 000 administrational, topographical and environmental entities in the form of geoobjects
- The geometries are rastered in a 3×3 km² grid
  - ⇨ The whole geothesaurus can be represented in a relational database (no "geographic information system" required)
- Application in two German environmental metainformation systems:
  - ⇨ GEIN (German Environmental Information Network)
  - ⇨ GISU (Geographic Information System Environment)

EXAMPLE GISU: DISPLAYING THE GAZETTEER AS A MAP



GiSU - Filter festlegen - Microsoft Internet Explorer

Adresse KEN=20175&UID=-99&SID=-99&from=main%2Ecfm%3FCFID%3D8%26CFTOKEN%3D20175%26UID%3D%2D99%26SID%3D%2D99%26CID%3

Regierungsbezirke

Alle Informationen
Administrative Einheiten
Bundesrepublik Deutschland
Bundesländer
Regierungsbezirke
Landeshauptstädte
Kreise
Gemeinden

🟪	257615
□	152518
🟨	161691
🟩	184287
🟩	198337
🟩	204960
⬜	210710
🟦	216627
🟫	296489
🟩	340081
🟥	357120
🟪	473907
🟥	704732

418,065.05 ↔
6,845,608.04 ↕

250000 Degrees

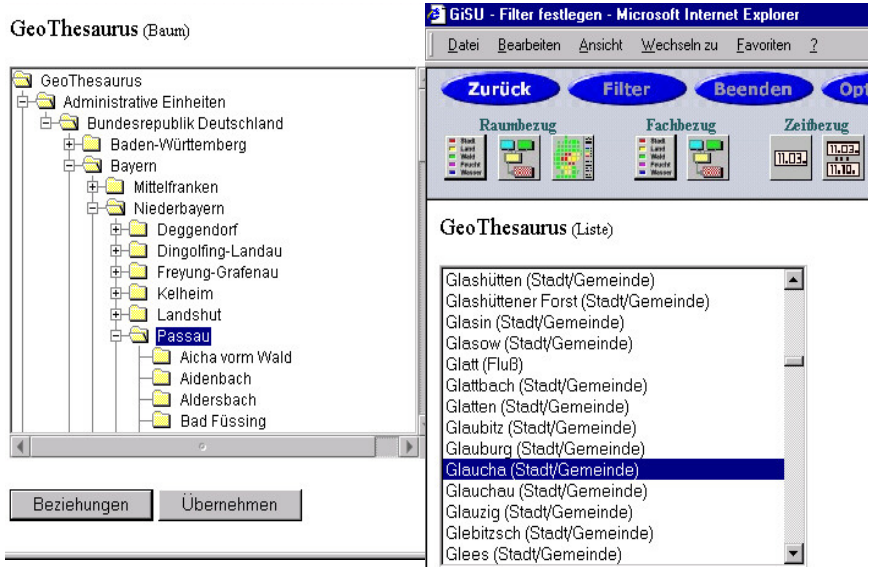
Filterbedingung

Filter
FachThesaurus
Freie Suchbegriffe
GeoThesaurus
Bexbach (Stadt/Gemeind
Zeitfilter

Wählen Sie eine Begriffsverknüpfung aus:

◉ ODER
○ UND

GISU: DISPLAYING THE GAZETTEER AS A TREE OR AS A LIST OF TERMS


GiSU - Filter festlegen - Microsoft Internet Explorer
Datei Bearbeiten Ansicht Wechseln zu Favoriten ?
Zurück
Filter
Beenden
Opt
Raumbezug
Fachbezug
Zeitbezug
11.03.
11.10.
Beziehungen
Übernehmen

GeoThesaurus (Baum)

- GeoThesaurus
  - Administrative Einheiten
    - Bundesrepublik Deutschland
      - Baden-Württemberg
      - Bayern
        - Mittelfranken
        - Niederbayern
          - Deggendorf
          - Dingolfing-Landau
          - Freyung-Grafenau
          - Kelheim
          - Landshut
          - Passau
            - Aicha vorm Wald
            - Aidenbach
            - Aldersbach
            - Bad Füssing

GeoThesaurus (Liste)

Glashütten (Stadt/Gemeinde)
Glashüttener Forst (Stadt/Gemeinde)
Glasin (Stadt/Gemeinde)
Glasow (Stadt/Gemeinde)
Glatt (Fluß)
Glattbach (Stadt/Gemeinde)
Glatten (Stadt/Gemeinde)
Glaubitz (Stadt/Gemeinde)
Glauburg (Stadt/Gemeinde)
Glaucha (Stadt/Gemeinde)
Glauchau (Stadt/Gemeinde)
Glauzig (Stadt/Gemeinde)
Glebitzsch (Stadt/Gemeinde)
Glees (Stadt/Gemeinde)

RESULTS

- Metainformation systems and search engines can be enhanced considerably by thesauri and gazetteers
- New attractive user interfaces: maps, directories, network graphics instead of blackbox search
- Flexible vocabulary for the specification of thematic and spatial references
  - ⇨ Automated reformulation, extension, and translation of terms and geographic locations
- Thesauri and gazetteers are knowledge structures which are relatively stable and application-independent
  - ⇨ maintenance relatively inexpensive
  - ⇨ reusable in multiple applications
- Investments in thesauri and gazetteers pay