BOBCATSSS 2002
Portorož, Slovenia, January 28-30, 2002

FACHHOCHSCHULE STUTTGART
HOCHSCHULE DER MEDIEN

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

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

ACKNOWLEDMENTS

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

- Research Institute of Applied Knowledge Processing (FAW), Ulm (with the author as the project leader)
- Condat AG, Berlin (formerly: CAdMAp GmbH, Berlin)

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

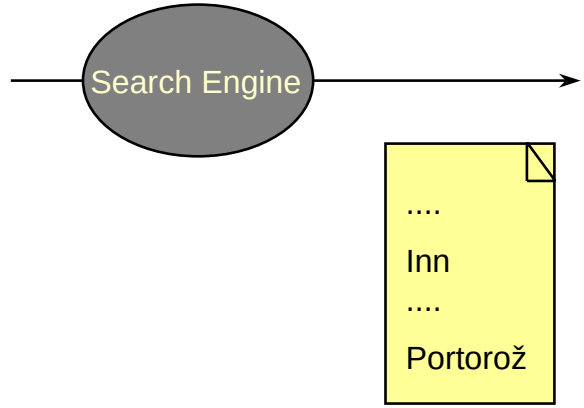
But:

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

EXAMPLE

Query:

"*Accommodation, Slovenian Adria"*



METAINFORMATION SYSTEMS

Metainformation systems support semantic criteria for indexing and retrieval:

- Thematic references (e.g., "Inn")
- Spatial references (e.g., "Slovenian Adria")
- Temporal references (e.g., "January 28-30, 2002")

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

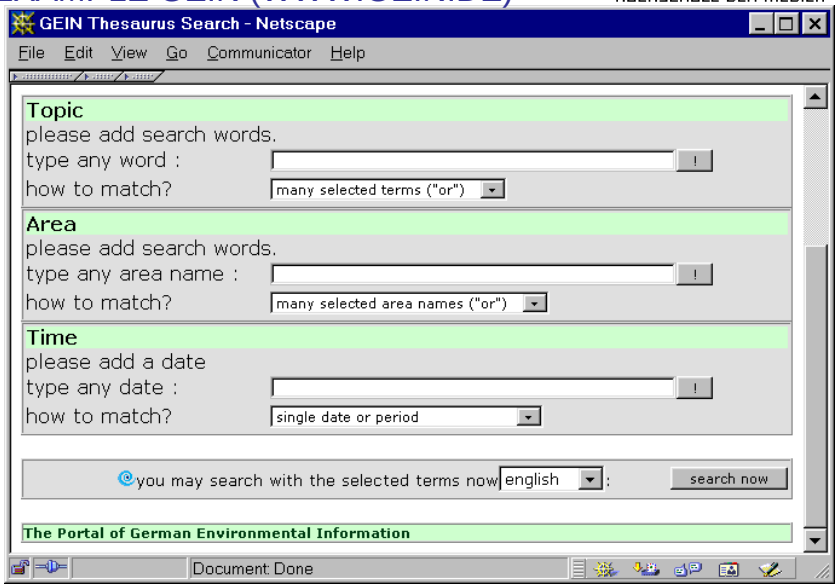
- Higher information quality (compared to search engines)
- Higher workload imposed on system administrator or information suppliers

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)



SPECIFICATION AND PROCESSING OF SEMANTIC CRITERIA

FACHHOCHSCHULE STUTTGART
HOCHSCHULE DER MEDIEN

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

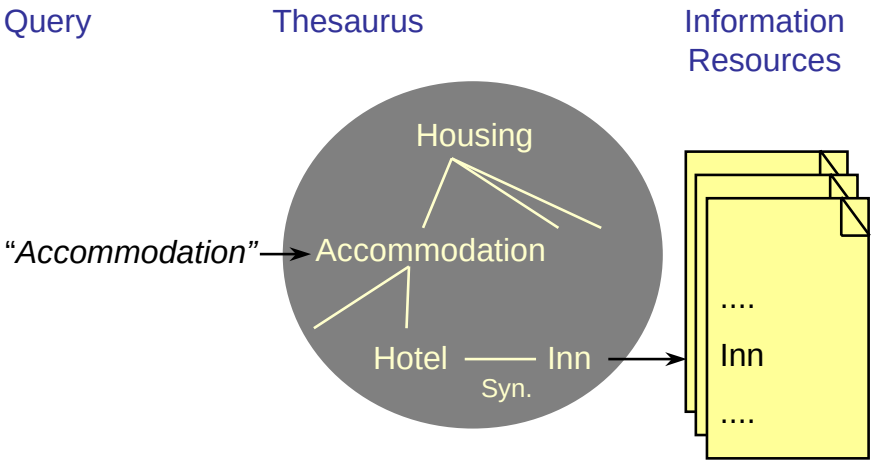
FACHHOCHSCHULE STUTTGART
HOCHSCHULE DER MEDIEN

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

FACHHOCHSCHULE STUTTGART
HOCHSCHULE DER MEDIEN



BLACK BOX SEARCH PROBLEM: A THESAURUS CAN HELP

FACHHOCHSCHULE STUTTGART
HOCHSCHULE DER MEDIEN

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

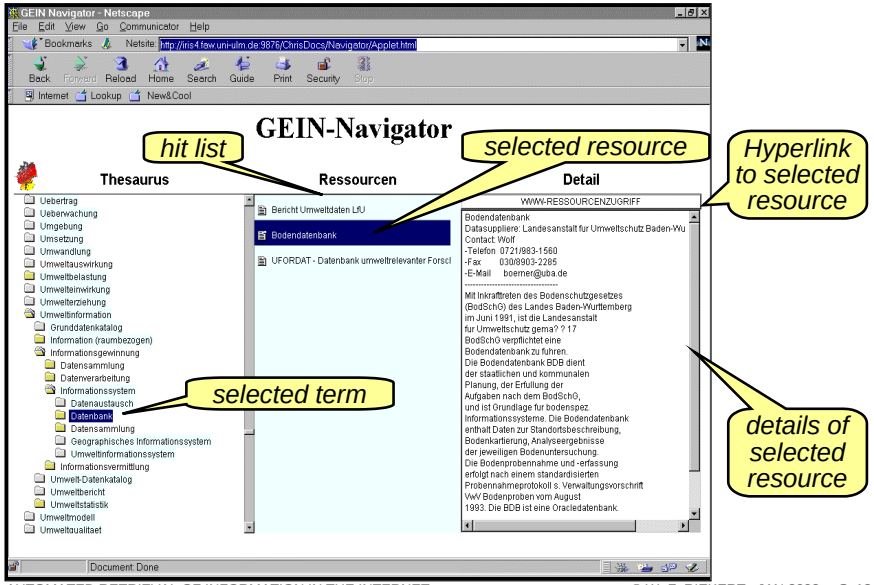
Requirement: Hierarchical directories to access the information resources

- 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



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

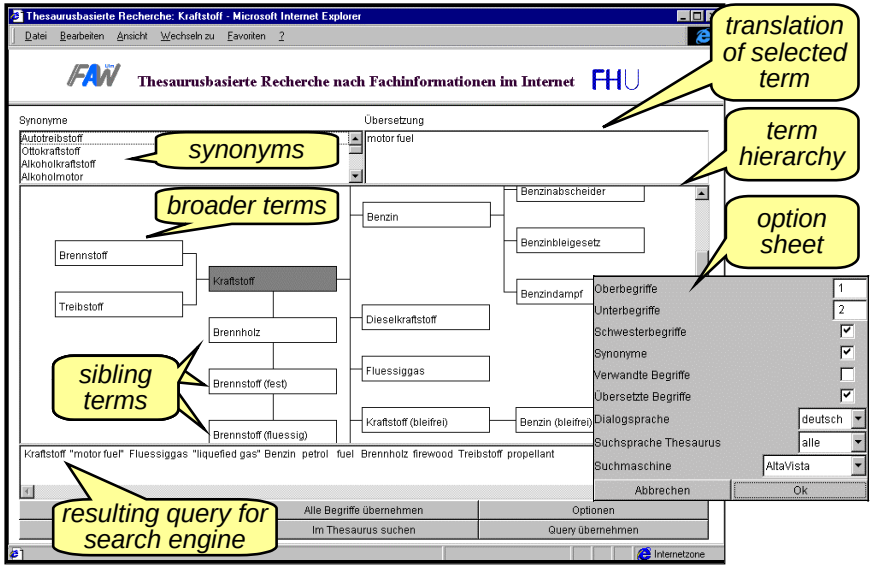
Synthesis:

- 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



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

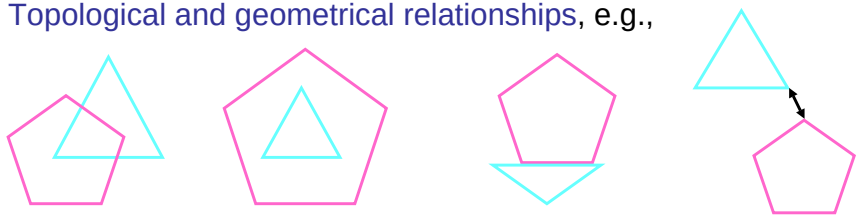
GAZETTEER

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., “Slovenia”)
  - ⇨ geometry (e.g., coordinates describing a polygon)
  - ⇨ type (e.g., “country”)
  - ⇨ unique identifier (e.g., country code)
  - ⇨ optional: hierarchy (e.g., administrational hierarchy)
  - ⇨ optional: synonymous names, translated names

TOPOLOGICAL AND GEOMETRICAL RELATIONSHIPS

Topological and geometrical relationships, e.g.,



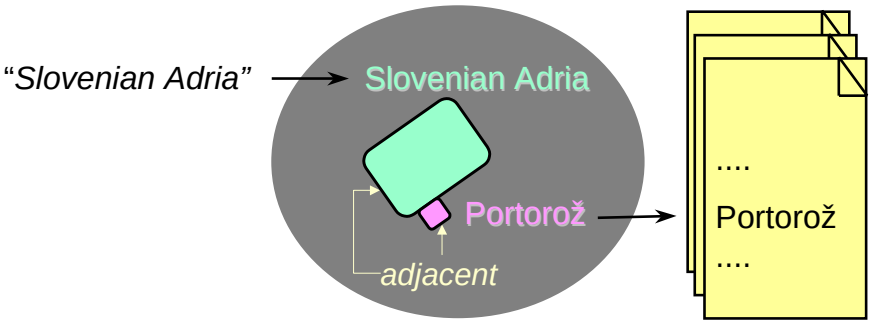
can be derived from geometry (i.e., coordinates)

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

- indexing and
- retrieval purposes

GAZETTEER-BASED RETRIEVAL

Query Gazetteer Information Resources



GAZETTEER: APPLICATIONS

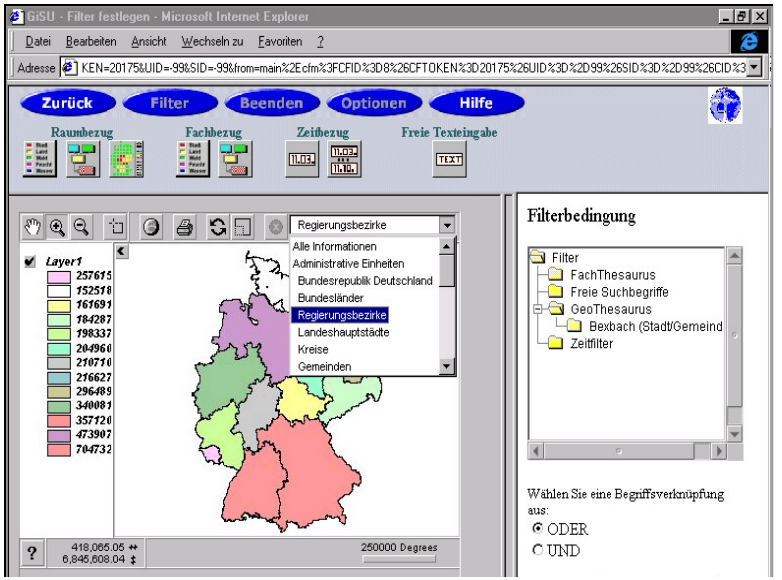
- Browser for geographical names
- Specification of spatial references on a cartographic interface
- Automated extension of queries: information resources in a certain geographic or topological neighborhood can be included into the scope of the query
- Easy transition between textual and geometrical 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

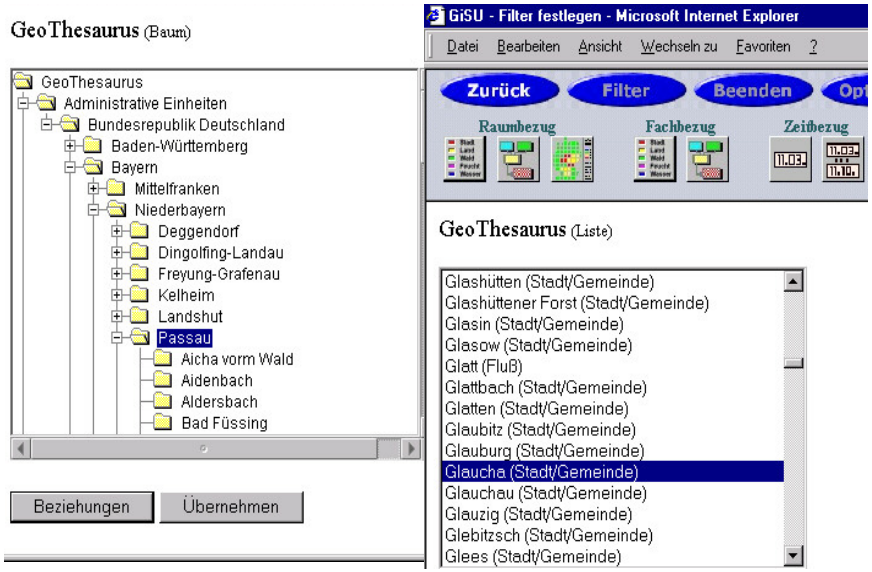
The German Federal Environment Agency developed a Gazetteer known as "Geothesaurus"

- It contains about 100 000 administrational, topographical and environmental entities in the form of geoobjects
- The geometries are rastered in 3×3 km² squares
  - ⇨ 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: DISPLAYING THE GAZETTEER AS A TREE OR AS A LIST OF TERMS



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