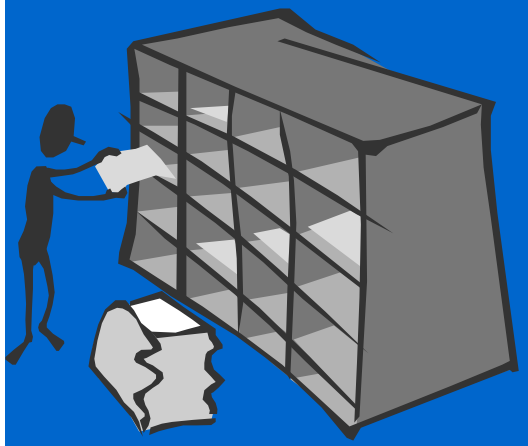
GERHARD

German Harvest Automated Retrieval and Directory

Automatisches Sammeln, Klassifizieren und Indexieren wissenschaftlicher Informationsressourcen im deutschen WWW



Dr. Kai-Uwe Carstensen, ISIV Osnabrück
Bernd Diekmann, bis Oldenburg
Gerhard Möller, OFFIS Oldenburg

http://www.gerhard.de

Motivation

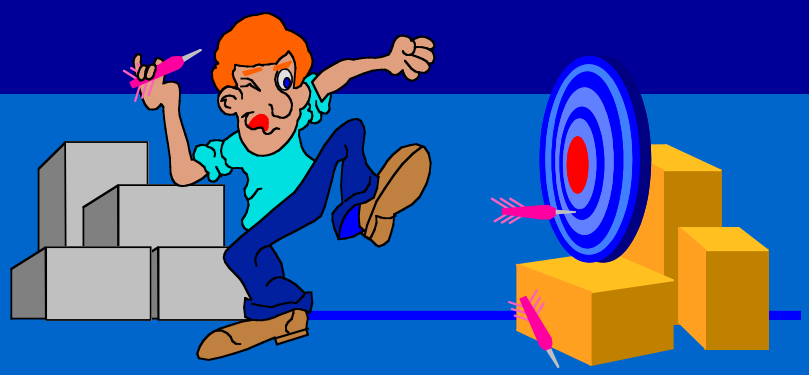
GERHARD

- Ausgangspunkt:
  - Interesse an wissenschaftlichen Informationen
    - aus dem deutschsprachigen Raum
    - im WWW
- Problem:
  - Fülle der Information im Web
  - Finden relevanter Information
    - eingeschränkt auf die Domäne
    - mit allgemeinen Suchmaschinen

* Unübersichtliche Flut an Suchresultaten

GERHARD

Ziel



GERHARD

- sammelt,
- klassifiziert und
- indexiert

automatisch wissenschaftlich relevante Information im deutschen World-Wide Web.

Aufgaben von GERHARD

GERHARD

- Integration von Suchen und Browsen
  - erlaubt Suche nach *Ähnlichkeit*
- Professionelles Klassifikationsschema
  - Intellektuell erstellte Ontologie (UDK)
- Automatische Klassifikation
  - der *einzige* Weg um *up-to-date* zu sein/zu bleiben
  - Muß sehr schnell sein (< 1 Sek/Dokument)

Verwandte Projekte

GERHARD

- Nordic WAIS/WWW Project (Lund)
  - Erster Versuch, 51 Klassen der UDK benutzt
- OCLC Project Scorpion
  - benutzt Dewey Decimal Classification (DDC)
  - Noch kein Browsing
- Desire II (EU Projekt)
  - Erst Arbeitsplan verfügbar

Architektur und Partner

GERHARD

BIS
Univ. Oldenburg

Sammeln der Information

SOIF

ISIV
Univ. Osnabrück

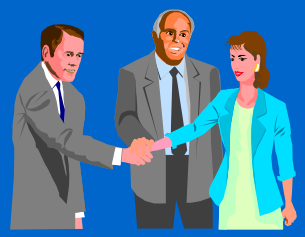
Automatische Klassifikation unter Verwendung der *UDK-ZH*

Klass.

OFFIS

Navigation, Suche, dynamische Representation und Speicherung

SOIF



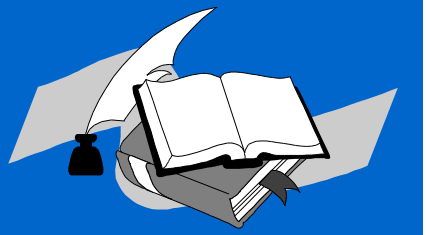
Sammeln der Information

GERHARD

- modifizierter Harvest robot
  - bald: *Combine*
- Konfigurations- und Statusdaten in einer Oracle Datenbank
- Voll konfigurierbar und kontrolliert durch eine Web-Schnittstelle

Universale Dezimalklassification

GERHARD



- Ursprung 1876 durch M. Dewey: *DDC*
- 1953: erste deutsche Ausgabe
- unbegrenzte *Erweiterbarkeit*
- Spezialversion von der ETHZ
  - *Drei-sprachig* (deutsch, englisch, franz.)
  - *~ 60.000 Einträge*
  - *13 verschiedene Relationen* zwischen Einträgen

Datenbank

GERHARD

- Strukturierte und Volltext-Daten
  - Verschiedene Datenbanktypen evaluiert
  - Oracle mit ConText Option und WebServer stellte sich als am geeignetsten heraus
- Konfigurierbar und kontrolliert durch eine Web-Schnittstelle

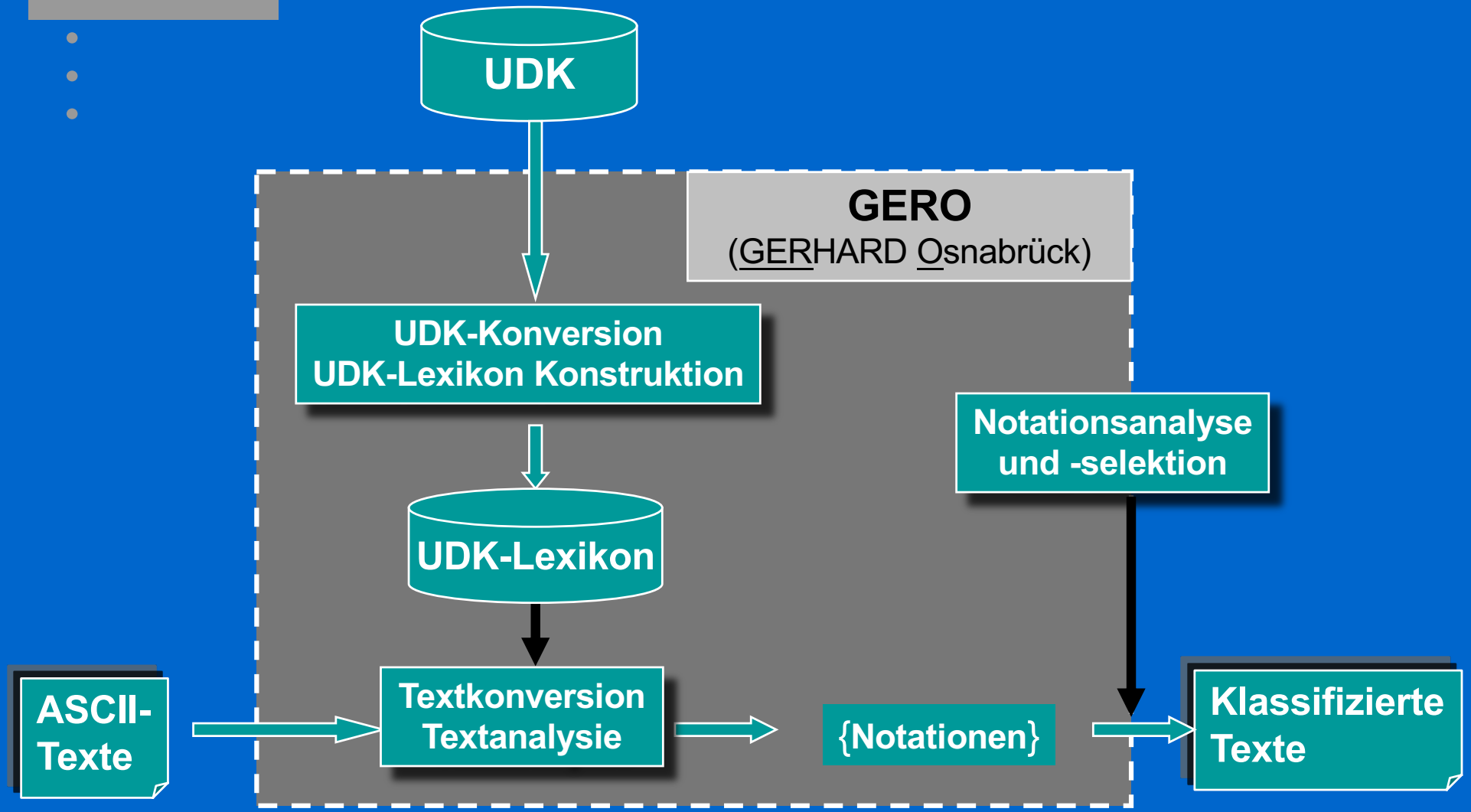
Klassifikation in GERHARD

GERHARD

- Verwendung einer Ontologie
  - Universelle Dezimalklassification (UDK)
- Pragmatischer Ansatz
  - maximale Qualität
  - minimaler Zeitverbrauch
  - Anwendung linguistischer Technologien

Linguistische Klassifikation

GERHARD



UDK

GERHARD

- Hierarchische Kodierung klassifizierender Information

  z.B.:

  5: ”Mathematik/Naturwissenschaften“
  51: " Mathematik“
  53: ”Physik“
  536.1: ”Hitzetheorie“
  536.13: ”Hitzepotential“

- Breite Abdeckung von Wissensbereichen
  (~60000 Einträge, 27MB Daten)

- Mehrsprachiger Ansatz (Deutsch, Englisch, Französisch)

UDK-Source

GERHARD

001Z ~03
002DDUEBERSETZUNGEN / TECHNISCHE U. NATURWISSENSCHAFTLICHE
003DETRANSLATIONS / TECHNICAL AND SCIENTIFIC
004DFTRADUCTION / SCIENTIFIQUE ET TECHNIQUE

001Z ~30-022
002DDALTHOCHDEUTSCH
003DEOLD HIGH GERMAN
004DFANCIEN HAUT-ALLEMAND
005TDALTHOCHDEUTSCH/WOERTERBUECHER
005TDSPRACHEN, WOERTERBUECHER/ALTHOCHDEUTSCH
005TDDEUTSCH/ALTHOCHDEUTSCH
006TEOLD HIGH GERMAN/DICTIONARIES
006TELANGUAGES, DICTIONARIES/OLD HIGH GERMAN
006TEGERMAN/OLD HIGH GERMAN
007TFHAUT-ALLEMAND/ANCIEN
007TFLANGUES/HAUT-ALLEMAND, ANCIEN
007TFALLEMAND/HAUT-ALLEMAND, ANCIEN
007TFVIEUX HAUT-ALLEMAND
007TFDICTIONNAIRES/HAUT-ALLEMAND, ANCIEN

UDK-Source

GERHARD

001Z 519.767.6

002DDMETHODEN DER GEWINNUNG SEMANTISCHER INFORMATION AUS EINEM TEXT (MATH. LINGUISTIK)

003DESEMANTIC INFORMATION, EXTRACTION FROM TEXTS (MATH. LINGUISTICS)

004DFMETHODES D'EXTRACTION D'INFORMATION SEMANTIQUE D'UN TEXTE

005TDINFORMATION/SEMANTISCHE INFORMATION (MATH.LINGUISTIK)

005TDSEMANTISCHE INFORMATION (MATH.LINGUISTIK)

006TESEMANTICS/METHOD OF EXTRACTION OF INFORMATION FROM TEXTS (MATH.LINGUISTICS)

007TFINFORMATION/SEMANTIQUE, METHODES D'EXTRACTION (LINGUISTIQUE MATH.)

007TFSEMANTIQUE/EXTRACTION D'INFORMATION SEMANTIQUE D'UN TEXTE (LINGUISTIQUE MATH.)

001Z 519.768

002DDMASCHINELLE UEBERSETZUNG / MATH. LINGUISTIK

003DEMACHINE TRANSLATION / MATH. LINGUISTICS

004DFTRADUCTION ASSISTIE PAR ORDINATEUR

005TDUEBERSETZUNGEN/MASCHINELLE UEBERSETZUNGEN

005TDMASCHINELLE UEBERSETZUNG/SEMANTISCHE INFORMATIONSVERARBEITUNG

006TETRANSLATION/COMPUTER (MATH.LINGUISTICS)

006TECOMPUTER TRANSLATION (MATH.LINGUISTICS)

006TECAT (COMPUTER AIDED TRANSLATION)

007TFTAO (TRADUCTION ASSISTIE PAR ORDINATEUR)

UDK-Konversion

GERHARD

- …eine Menge automatischen Editierens…
- Entfernen von Stopwörtern, Abkürzungen etc. aus den UDK-Einträgen
  - Verwendung der CELEX-Datenbank (Max Planck Institut, Nijmegen)
- Extraktion *nützlicher* natürlich-sprachlicher Ausdrücke

UDK-Lexikon Konstruktion

GERHARD

- Verwendung von Lingsoft® für die
  - morphologische Reduktion von Wörtern in der UDK
  - Generierung aller möglichen Endungen eines Wortstamms
- Konstruktion eines UDK-Lexikons
  - Abbildung von Wortstämmen+Endungen auf UDK-Notationen
  - Mehrwort-Einträge sind möglich
- Kompilation des UDK-Lexikons in einen Erkenner
  - Endlicher Automat

Beispiele

GERHARD

old~~A high~~A german~~A/ dictionary~dictionaries~~N
uebersetzung~~S/technisch~~A u.~~ABK naturwissenschaftlich~~A

old high german dictionary
old high german dictionaries

technisch uebersetzung
naturwissenschaftlich uebersetzung

Einträge im UDK-Lexikon

althochdeutsch:-:~30-022
old high german:-:~30-022
ancien haut-allemand:-:~30-022
althochdeutsch woerterbuch:-:~30-022
althochdeutsch sprach:-:~30-022
althochdeutsch woerterbuecher:-:~30-022
old high german dictionary:-:~30-022
old high german dictionaries:-:~30-022

technical translation:-:~03
scientific translation:-:~03
traduction scientifique technique:-:~03

semantisch information:-:519.767.6
semantisch informationsverarbeitung:-:519.768

gen:xxx e s en:575.113.1

Textanalyse

GERHARD

- Textaufbereitung
  - Entfernung von Stopwörtern
- Text als Eingabe des Erkenners
  - Flexibles Erkennen,
  - sogar von Mehr-Wort-Ausdrücken
- Ausgabe: gefundene Notationen

Statistische (Prä)Analyse

GERHARD

- Ausnutzung der hierarchischen Notationsstruktur

“gute Notation“:

$$\sum_{i=1}^{depth} \frac{i * counter(not[i]) * lwf}{fn} > t$$

depth (of analysis) := **average_notation_length**

lwf (length weighting factor) := longest_match_for_prefix_not[1..i]/average_match_length

t (threshold) := **average_match_length /average_notation_length**

fn (found notations) := **counter(bag of found notations)**

Notationsauswahl

GERHARD

- Auswahl von
  - Notationen auffälliger cluster und/oder
  - einzelnen Notationen guter (d.h., langer, Mehr-Wort-) Treffer

- Zusätzliche Selektion durch Relevanzgewichtung
  - Auftreten im Titel vs. Textkörper
  - Häufigkeit des Auftretens
  - Abwerten geografischer Termini
  - Reduktion auf ungefähr 2/3 der vorher ausgewählten Notationen

Benutzerfreundliche Schnittstelle

GERHARD

- Vorgabe: so einfach wie möglich
- Suche in Dokumenten und der Klassifikation
- Browsing
  - Nur die UDK-Einträge mit ihrer Information werden angezeigt
  - Nahtlose Integration von Suche und Browsing
- Änderung der Sprache unmittelbar möglich
- Kontext-sensitive Hilfe

Suchresultate

GERHARD

Anfrage: althochdeutsch (alle Begriffe)

zugeordnete Dokumente 1 bis 25 (von insgesamt 105)

- meipubl.htm
- etymologisches w rterbuch des althochdeutschen (ewa)
- auswahl der publikationen von prof. dr. bergmann zum althochdeutschen
- (stefanie stricker)
- althochdeutsches woerterbuch (goettingen)
- prof. dr. rolf bergmann, ls f r dt. sprachwissenschaft und ltere dt. literatur, 29.06.1995
- e. leiss: entstehung des artikels im deutschen
- uni-bamberg: lehrstuhl f r deutsche philologie des mittelalters
- deutschsprachige w rterb cher
- deutschsprachige w rterb cher
- lehrangebot f r das ss 1998
- byq-bzz altdeutsche und althochdeutsche sprache und literatur
- deutschsprachige woerterbuecher - projekte an akademien, universitaeten, instituten -
- deutsche philologie
- (anmerkungen zur graphik auf der homepage)
- byq-bzz altdeutsche und althochdeutsche sprache und literatur
- rechtschreibreform - geschichte

Zukunftsmusik

GERHARD

- Verbesserung der Klassifikation
  - durch mehr (Katalog-/statistische) Daten
  - Unterschiedliche Behandlung von Dokumenten verschiedenen Typs
- effizientere Volltextbehandlung
- Interessenprofile
- Verbesserung von Ähnlichkeitssuche und -navigation
- Viele, viele andere kleine Änderungen