

# Suchstrategien und Wissen erschließen im Web

Autor: Christian Spohn Datum: 20.09.2005

## Einleitung:

- Vorstellung einer **neuartigen Technologie**, die es erlaubt, **Wissen** auch von schwach strukturierten Dokumenten (i. e. nicht annotierten, reinen Texten) zu **erschließen**.
- Es werden die Basisbausteine dieser Technologie beschrieben. Diese Basisbausteine können auch zu komplexen Auswertungen verknüpft werden. „Kaskadierende“ Auswertungen oder Verknüpfung von Ergebnissen mehrerer Abfragen, i.e. Definitionen von Suchstrategien werden damit möglich (Automatic Readers). Komplexe inhaltliche Auswertungen, d.h. Auswertungen, die die Vielfalt der sprachlichen Möglichkeiten berücksichtigen, sind durchführbar (wissenschaftliche Suche). Dabei werden nicht nur Dokumente, sondern insbesondere die resultierenden, kontextsensitiven Textstellen zurückgeliefert (inhaltliche Suche, erschließen von Wissen).
- Da diese Technologie quasi vollautomatisch ablaufen kann, ist sie namentlich für die Verarbeitung großer Datenmengen d.h. auch für das Web geeignet.

## Warum neuartig:

Herkömmliche Information Retrieval Systeme verwenden

„... **mathematische Modelle, die weder natürliche Sprachen noch andere kognitive Fähigkeiten des Menschen adäquat modellieren...**“

(aus Mandl, Tolerantes Information Retrieval)

Anders ausgedrückt : Sprache hat nur bedingt oder herzlich wenig mit Mathematik bzw. mit Statistik zu tun, **d.h. ein anderes Modell ist gefordert:**

Dieses Modell steht quasi auf 3 Säulen:

- Informationsexplizierung statt Informationsreduktion
- Sprachkonformes Modell
- Automatisches Handling sehr großer Datenmengen muss möglich sein

Neuartige Technologie:

- neu von ihrem Ansatz
- neu von ihrer Umsetzung
- neu von den daraus resultierenden Eigenschaften

## Was soll unter „Wissen erschließen“ verstanden werden:

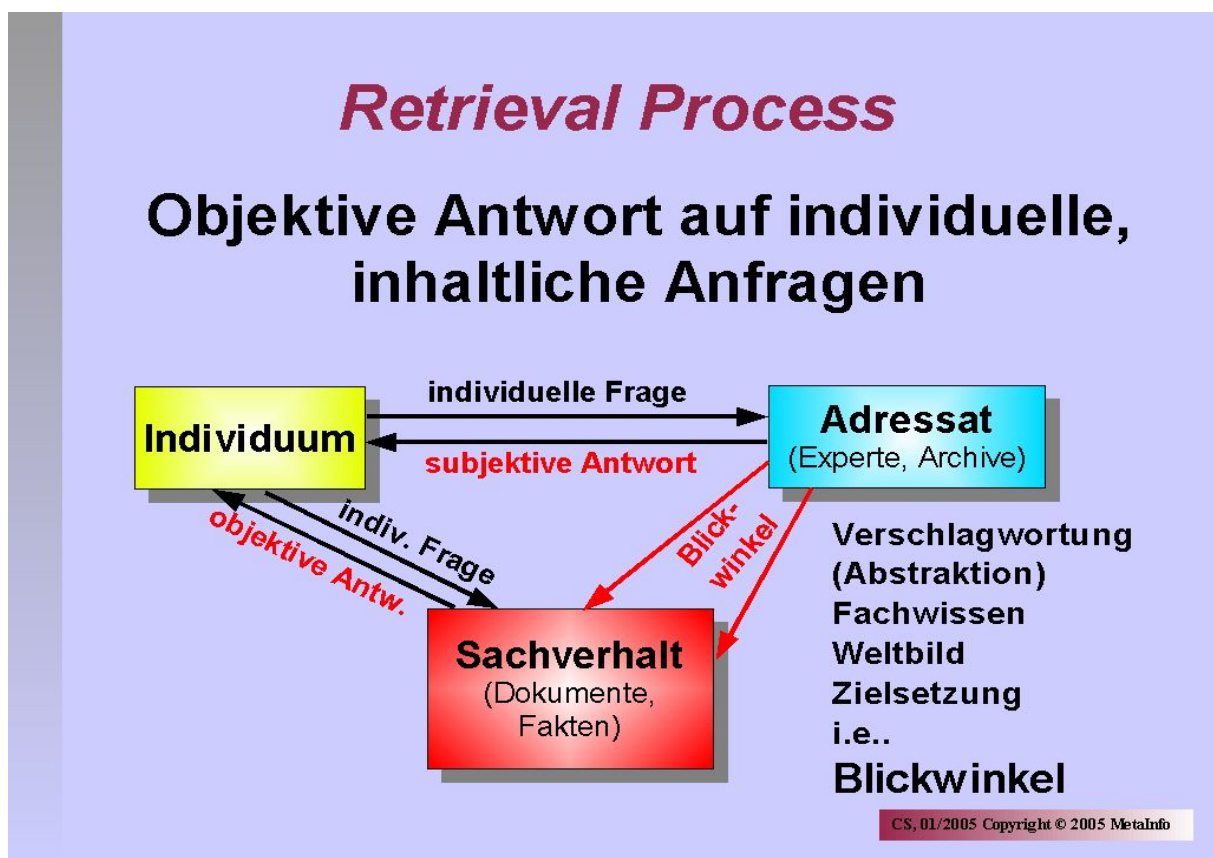
So wie wir unser Wissen sprachlich zu „Papier“ bringen, sollte es umgekehrt möglich sein, rein sprachlich dieses Wissen wieder zu erschließen.

Denken, Sprache, Dokumente können als Einheit aufgefasst werden: ohne Sprache kein Denken, ohne Sprache keine Dokumente (bzw. keine Sprachaufzeichnungen)  
Sprache ist somit der „Schlüssel“ für die Erschließung von Wissen

## Was man bräuchte...

**Objektive Antworten auf individuelle Fragen (im Sinne von inhaltlichen Anfragen) ...**

Über eine rein „sprachliche Schnittstelle“ soll mit dem System kommuniziert werden. Sprachverständnis, Fachkenntnisse bzw. Bildung werden über diese „sprachliche Schnittstelle“ umgesetzt. Das System soll sich intuitiv d.h. selbsterklärend erschließen.



**Bild 1:** Objektive Antworten auf individuelle Fragen

Dass dieses Problem nicht neu ist, zeigen sehr anschaulich zwei Zitate, die bis heute nichts von ihrer Aktualität eingebüßt haben:

**Eine kluge Frage ist die Hälfte der Weisheit**

Francis Bacon

**Wenn du eine weise Antwort willst,  
musst du vernünftig fragen**

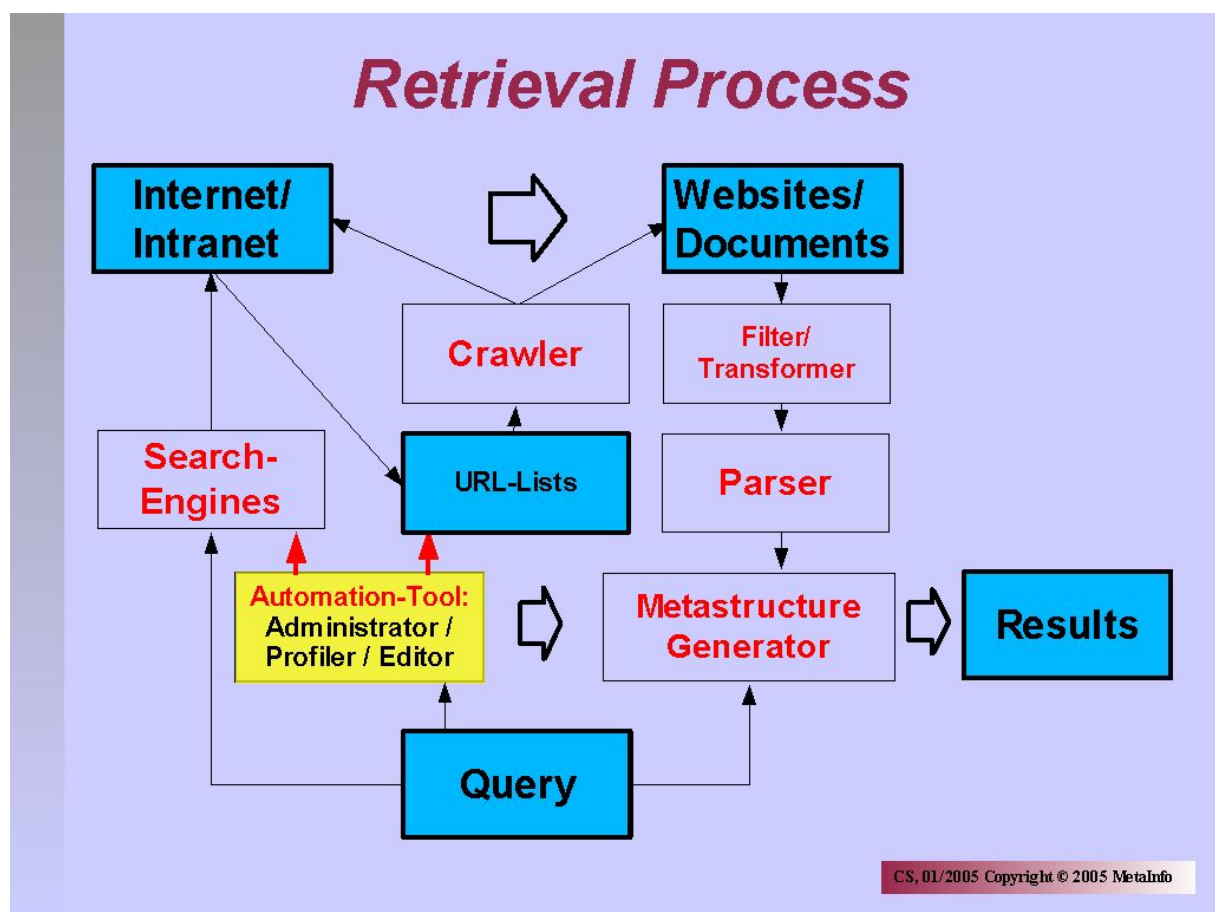
Johann Wolfgang von Goethe

An diesem Sachverhalt hat sich bis heute nichts geändert:

Menschliche Intelligenz umgesetzt mit sprachlichen Mitteln ist der entscheidende Schlüssel für die Erschließung von Wissen!

## Das System

Informationen liegen in vielfältiger Form und an vielfältigen Orten vor. Wenn sie nicht in elektronischer Form vorliegen, lassen sie sich mit Standardtechnologien in eine elektronische Form überführen (z. B. Scannen plus OCR). Dank Internet lassen sie sich falls gewünscht weltweit sehr einfach publizieren und zugreifen.



**Bild 2:** Information Retrieval als komplexer Prozess

In Bild 2 ist der komplexe Retrieval Prozess dargestellt: Die einzelnen Utilities, Werkzeuge oder Applikationen sind rot beschriftet. Mit Pfeilen ist die Datenflussrichtung (von → nach) dargestellt. Besonders hervorzuheben ist, dass sich der gesamte Prozess vollständig automatisieren lässt. Damit ist diese Technologie auch für die Bearbeitung sehr großer Informationsmengen geeignet, zumal sich der Ablauf des Gesamtprozesses entsprechend parallelisieren und skalieren lässt. Internet oder Intranet lässt sich damit quasi als Supercomputer nutzen.

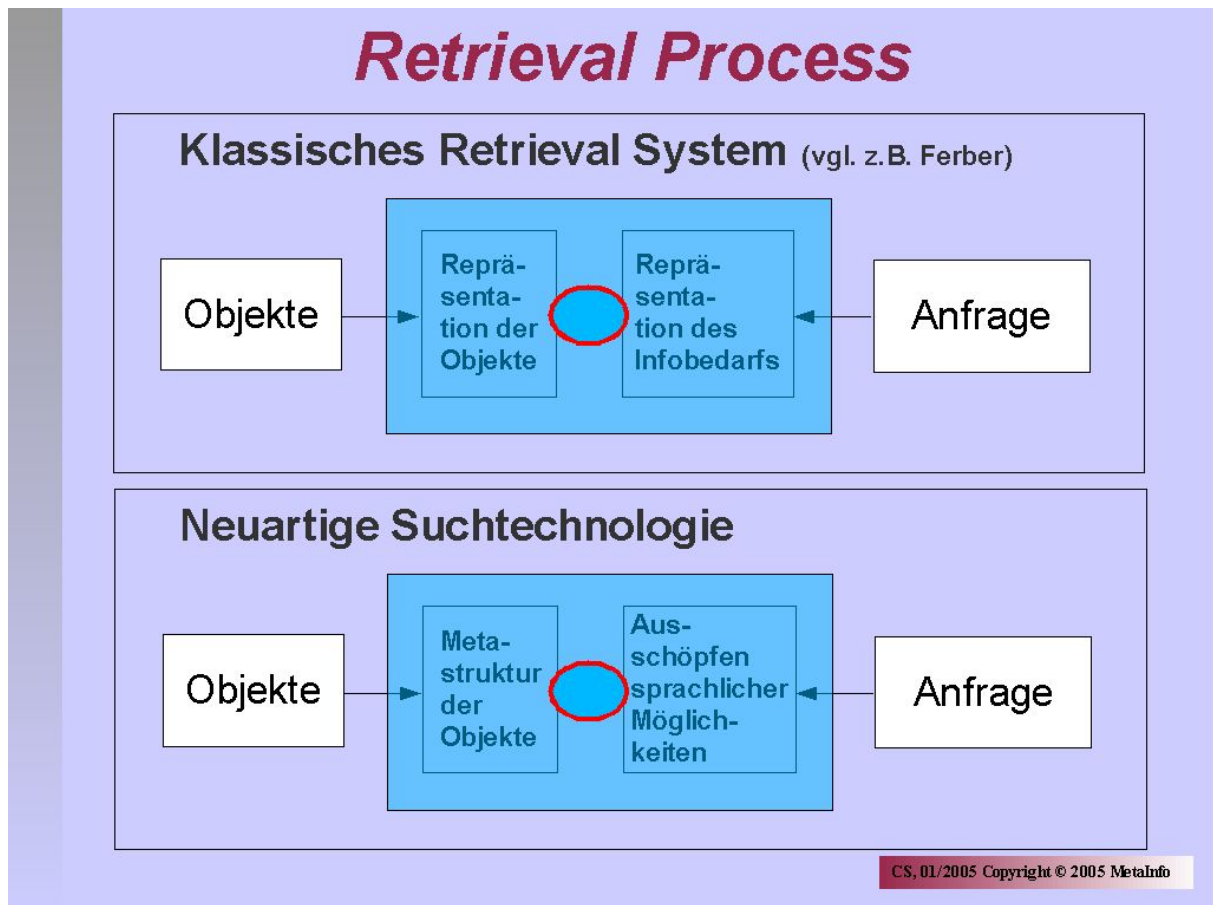
Vom Einsammeln der Information, der automatischen Aufbereitung, bis zur (evtl. automatischen) Abfrage oder Auswertung ist damit der ganze Information Retrieval Prozess abbildbar und für die unterschiedlichsten Anforderungen oder Anwendungen einfach adaptierbar.

## Klassisches Information Retrieval und neuartige Suchtechnologie

Klassische Retrieval Systeme repräsentieren jedes Dokument durch charakteristische Begriffe oder Indexterme. Jeder Indexterm wird evtl. noch mit einer Häufigkeit bzw. Gewichtung verknüpft. Mit Hilfe „normierter Sprache“ wird versucht, die Verwendung von Indextermen zu vereinheitlichen und in ihrer Grundform (Nominativ Singular oder Plural z.B.) zu verwenden.

Ergebnis: Jedes Dokument wird prinzipiell durch eine Liste von Indextermen repräsentiert. Das Retrieval System arbeitet also mit reduzierter Information.

**Aus der Liste von Indextermen kann aber das Originaldokument nicht mehr rekonstruiert werden.**



**Bild 3:** Unterschiedliche Repräsentationsformen der Systeme

Die neuartige Suchtechnologie verwendet kein mathematisch statistisches Modell wie z.B. das oben beschriebene Vektorraummodell. Statt Information zu reduzieren, werden alle Informationen, die ein Dokument enthält, redundanzfrei expliziert. Ergebnis sind die so genannten Metastrukturen. Diese bilden die

**Basis für die hochperformante Algorithmen dieser neuartigen Suchtechnologie.**

## **Bedeutung der Sprache**

Da Denken ohne Sprache nicht möglich ist, andererseits Dokumente als Sprache in geschriebener Form aufgefasst werden können, kann die Sprache als Schlüssel zur Wissenserschließung angesehen werden.



**Bild 4:** Sprache als Schlüssel zur Wissenserschließung

Die Güte von Retrievalsystemen steht und fällt deshalb mit der Vielfalt sprachlicher Formulierungen und Ausdrucksweisen, die beim Retrievalprozess entsprechend manuell, halbautomatisch oder automatisch berücksichtigt werden können. Eine Systemunterstützung ist hierbei unabdingbar, da ansonsten eine umfassende bzw. vollständige Anfrage nur sehr schwer oder gar nicht formulierbar ist.

Im ersten Schritt erzeugt das System deshalb ein suchraumabhängiges, erweitertes Lexikon. Erweitert, weil alle Wörter oder Informationseinheiten des entsprechenden Suchraumes inkl. aller Flexionsformen, Schreibfehler oder Schreibweisen eingetragen werden. Dieses erweiterte Lexikon definiert damit den vollständigen Sprachraum des betrachteten Suchraums. Alle linguistischen Funktionen operieren auf diesem realitätskonformen Sprachraum.

**Im Gegensatz zu klassischen Retrieval Systemen wird der Sprachraum nicht reduziert, d.h. es werden z.B. nicht nur Begriffe betrachtet. Es werden auch**

**keine so genannte Stoppwörter eliminiert** (was natürlich auch möglich ist). **Selbst Satz- und Sonderzeichen können als eigenständige Informationseinheiten betrachtet werden.**

## Das Phänomen Sprache

Vielfältige Ausdrucksmöglichkeiten sowie enorme Fehlertoleranz zeichnen die Sprache aus. So haben wir keine Probleme z. B. gebrochenes bzw. fehlerhaftes Deutsch zu verstehen. Dabei kann der situationsbezogene Kontext oft maßgeblich zum Verständnis beitragen. Homonyme bereiten normalerweise keine Probleme, da der sprachliche Kontext für ihre Eindeutigkeit sorgt: Das Kind isst eine Birne – der Fakir isst eine Birne. Birne isst eine Birne (also unser Altbundeskanzler Dr. Kohl). Während der Fakir wohl eine Glühbirne isst, trifft das für das Kind hoffentlich nicht zu und unserem Altbundeskanzler wird es wohl wie immer schmecken.

Bis heute ist es äußerst schwierig, gute Übersetzungstools zu entwickeln. Perfekte Tools zu entwickeln, wird wohl nie möglich sein, da eine gute Übersetzung im Regelfall von sehr komplexen Kontextinformationen abhängig ist, die dem/der Übersetzer/in bekannt sind, aber einem Tool nur schwerlich (vollständig) übermittelt werden können. Aus diesem Dilemma, den Kontext nur unzureichend beschreiben zu können bzw. nur unzureichend aus vorhandener Information rekonstruieren zu können, resultieren die Schwächen eines noch so guten Tools.

Anders ausgedrückt ein realitätskonformes Sprachmodell für den Fall perfekter Übersetzung zu entwickeln, welches allgemein anwendbar ist, ist aus diesen Gründen wohl nie möglich, namentlich da sich Sprache und Sprachgebrauch verändern, weiterentwickeln und sehr stark von Fachgebiet und oder Personenkreis abhängig sein können.

**Im Falle des Information Retrieval gestaltet sich die Entwicklung eines realitätskonformen Sprachmodells dagegen relativ einfach, da das Ergebnis nicht erzeugt werden muss sondern vom Ergebnis ausgegangen werden kann:**

- Alle den Suchraum bildende Dokumente beschreiben immer einen vollständigen Sprachraum
- Der betrachtete (sinnvoll gewählte) Sprachraum enthält immer die vollständige Kontextinformation (d.h. immer die „fertige (quasi die perfekte) Übersetzung“). Individuelle Anfragen lassen sich mit Hilfe von Sprache formulieren und automatisch oder manuell parametrisieren (→ sprachliche Extensionen)
- Eine Vielzahl unterstützender Tools ist möglich: so lassen sich linguistische Regeln sprachspezifisch definieren (i. e. es lassen sich Regeln oder Grammatiken sprachspezifisch definieren), fachspezifische Thesauri aufbauen oder aus allgemeinen Quellen (z. B. auch aus dem Internet) verwenden
- Das Ergebnis muss nicht wie im Fall einer Übersetzung erzeugt werden, sondern liegt vor: Es steht als vielseitiger Fundus für unterstützende Funktionen dem/der Anwender/in zur Verfügung (z. B. in Form erweiterte lexikalischer Funktionen). Der Inhalt wird quasi aus Anwendersicht unter thematischem Blickwinkel aufbereitet oder zusammengestellt (Automatic Reader).

Ein Sprachmodell des Information Retrievals muss im Gegensatz zu einem Sprachmodell eines Übersetzungstools keine aktive Komponente sein. Es muss aber

in der Lage sein, „abgebildete sprachliche Kreativität“ zu explizieren und nutzbar zu machen, um auf diese Weise die Vielfalt sprachlicher Ausdrucksmöglichkeiten zu erschließen zu können.

## Sprachlicher Kontext

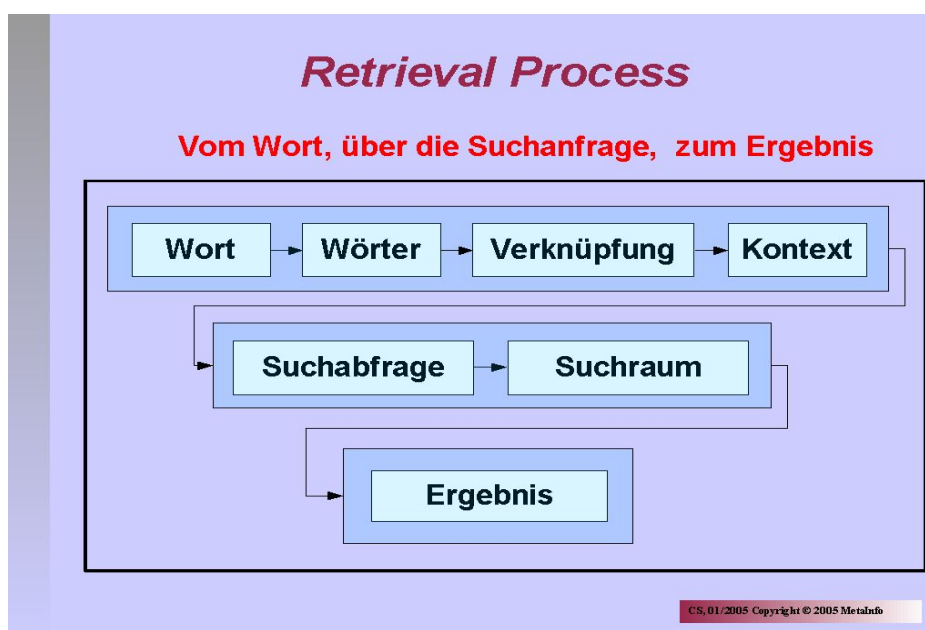
Wörter haben Synonyme oder Wörter ähnlicher Bedeutung. Zu Wörtern existieren Oberbegriffe, Unterbegriffe, Antonyme etc. d.h. Wörter sind allgemein ausgedrückt mit anderen Wörtern inhaltlich verbunden oder assoziiert. Auch individuelle oder persönliche Assoziationen von Wörtern sind denkbar. Während dieser Sachverhalt uns sehr vertraut ist, ist uns die „geometrische Dimension“ der Sprache weniger bewusst (oder vielleicht Dank klassischer Retrievalverfahren bzw. Dank Google gar nicht bewusst):

Werden Satzkonstruktionen zu kompliziert oder Sätze zu lang, sind wir nicht mehr in der Lage den Inhalt vollständig zu erfassen. Das menschliche Hirn verarbeitet Sprache quasi in kleineren Portionen oder „Paketen“.

Die Sprache bietet bezüglich der Anordnung der Wörter d.h. bezüglich der Satzstellung viel Freiheit. Der Inhalt einer (exakten) Phrase muss sich nicht verändern, wenn die Wortstellung sich ändert. Volltextsuche wird diesem Sachverhalt nicht gerecht.

Für das Information Retrieval bedeutet das, dass sprachlicher Kontext in 2 Dimensionen beschrieben werden muss:

- Der uns geläufige Kontext von Wörtern und ihren Synonymen, Oberbegriffen, Unterbegriffen etc. oder allgemein ihren assoziierten Wörtern (d.h. der linguistische Kontext)
- Der „Länge“ bzw. Größe der Sprachpakete sowie die freie Wortstellung (d.h. die „geometrische Dimension“ der Sprache oder ihr „räumlicher Kontext“)



**Bild 5:** Vom Wort über die Suchanfrage zum Ergebnis

## Ein Text besteht nicht nur aus Wörtern ...

Im zweiten Schritt erzeugt deshalb das System eine Darstellung der Dokumente, die nicht nur die Wörter, sondern auch die Strukturinformation des Textes erfasst. Selbst reine Texte enthalten implizite und explizite Strukturinformationen ohne die sie nicht lesbar wären: Schneidet man die Wörter eines Textes aus, mischt sie und versucht den Text zu rekonstruieren, steht man selbst bei kurzen Texten vor einer im Regelfall unlösbaren Aufgabe.

**Der klassische Ansatz des Indexierens von Texten ist praktisch mit dem Ausschneiden von Wörtern vergleichbar, in Wirklichkeit aber noch schlechter: nur ein Teil aller Wörter wird zur Indexierung herangezogen, normalerweise nur Begrifflichkeiten.**

## Die Dimensionen eines Wortes ...

Betrachtet man ein Wort als sprachliche Einheit um Anfragen an ein Retrievalsystem zu richten ist man mit folgendem Problem konfrontiert:

- Unterschiedliche Synonyme können benutzt werden
- Ein Wort kann in unterschiedlichen Deklinations- bzw. Konjugationsformen (i.e. Flexionsformen) existieren

(Klassische Retrieval Systeme versuchen das Problem zu umgehen, indem nicht die Dokumente, sondern eine Annotation standardisierter, lemmatisierter Indexterme betrachtet wird, was andere gravierende Nachteile nach sich zieht)

Eine weitere Dimension, die nicht unmittelbar erkennbar ist, ist die Dimension des Wortabstandes oder des Wortbereiches:

Die Sprache ist weitgehend tolerant, was die Anordnung von Wörtern d.h. ihre Wortstellung betrifft, d.h. es gibt zwei Arten von „exakten Phrasen“:

- Phrasen, bei denen die Anordnung der Wörter exakt der angegebenen Reihenfolge entspricht (vgl. so genannte Volltextsuche)
- Phrasen, die exakt die angegebenen Wörter enthalten, aber eine beliebige Anordnung dieser Wörter zulassen

In Wirklichkeit ist das Problem noch diffiziler:

Eine Phrase kann durch „Einschübe“ unterbrochen sein, so dass letztendlich

- Wörter
- ihr Kontextbereich
- sowie ihre Extension wie z. B. Erweiterung um Flexionsformen oder Erweiterung um Synonyme

betrachtet bzw. ausgewertet werden müssen.

Was relativ harmlos klingt, führt jedoch auch schon bei einfachen Abfragen zu Komplexitäten von solch erheblichem Ausmaß, dass klassische Retrievalverfahren nicht mehr angewandt werden können.

## Was wird eigentlich gesucht?

Die Systemschnittstelle für den/die Nutzerin ist die Sprache. Sprache bietet aber eine Vielzahl von Formulierungsmöglichkeiten.

Das System unterstützt die Formulierung

- sinngleicher
- sinnähnlicher

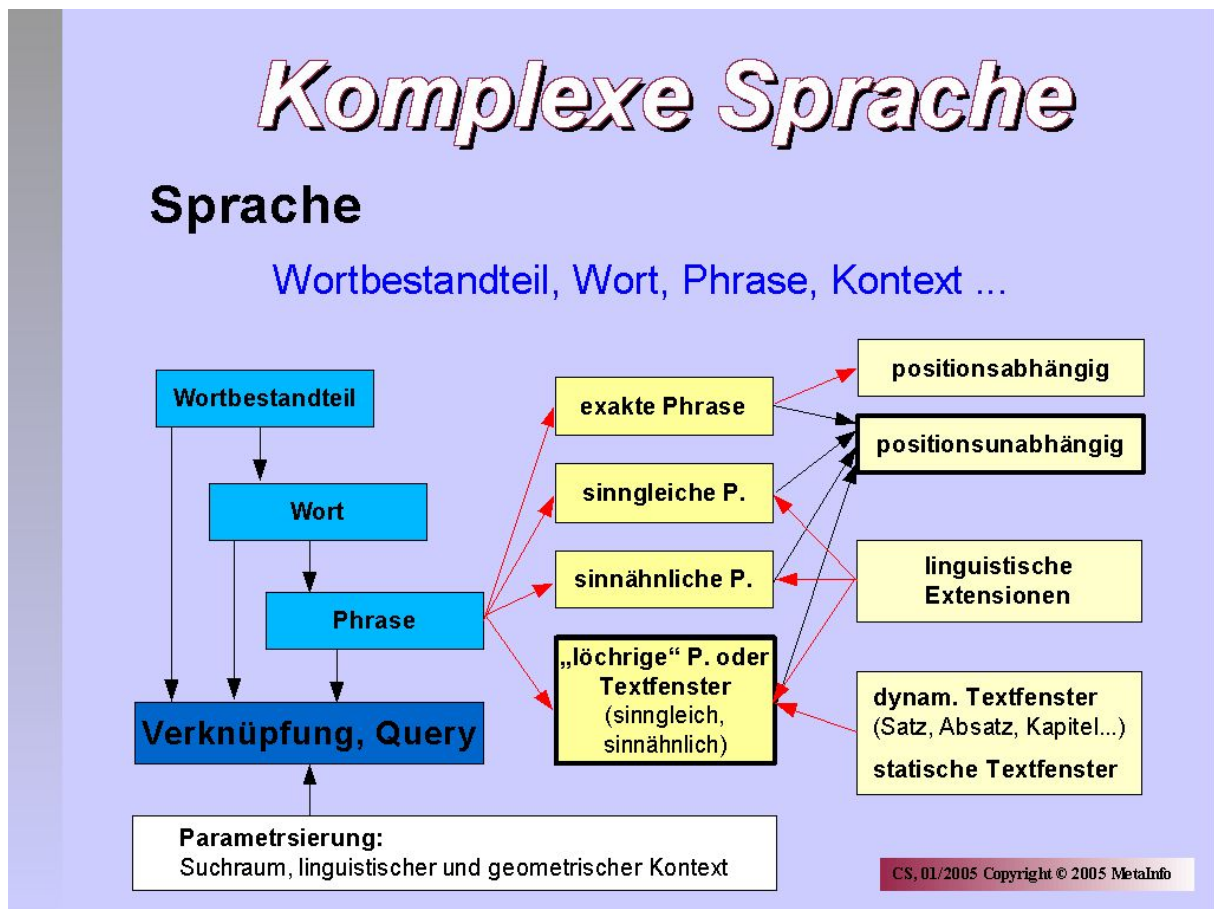
Anfragen in Form zusammenhängender Texte oder in Form logischer Verknüpfungen.

Hierbei ist es besonders wichtig, dass komplexe Logik einfach formuliert und gegebenenfalls angepasst oder nachgeprüft werden kann.

Logische Anfragen können in Form einer UND-Verknüpfung von Alternativen sehr einfach formuliert und gegebenenfalls ebenso wie Texte manuell, halbautomatisch oder automatisch extended werden.

Welcher Suchraum bearbeitet werden soll, ist ebenfalls parametrisierbar.

Die Arbeitsweise d.h. die „Suchstrategien“ des Systems können „on the fly“ eingestellt bzw. verändert werden.



**Bild 6:** Die Komplexität der Sprache, linguistischer und geometrischer Kontext

## Was wird eigentlich „gefunden“?

Von klassischen Retrieval Systemen sind wir gewohnt, einen Artikel, ein Buch oder eine HTML-Seite eben ein Dokument als Ergebnis zu erhalten. **Was wir aber suchen, sind eigentlich nicht die Dokumente, sondern bestimmte durch die Suchanfrage spezifizierte Inhalte aus diesen Dokumenten.** Die Dokumente sind also lediglich die genauer zu untersuchende Menge für die eigentlich gesuchten Textstellen. Daran ändert auch die Tatsache nichts, wenn eine „Leseprobe“, die die Wörter der Anfrage enthält, uns präsentiert wird (vgl. z.B. Ergebnisse von Internetsuchmaschinen).

**Ganz wichtig ist jedoch, dass eine Suchanfrage alle sinngleichen bzw. sinnähnlichen Suchanfragen mit einschliessen kann und deshalb auch dementsprechende Ergebnisse zu liefern vermag.** Der/die Anwenderin könnte glauben, dass das System mitdenkt, was in diesem Fall nicht einmal ganz verkehrt wäre, eben wegen der besonderen Bedeutung der Sprache für unser Denken.

## Bewertung von Ergebnissen

Bewertung von Ergebnissen oder Ranking ist eine vielschichtige Thematik. In erster Linie entscheiden die Anfrage und ihre Parametrisierung über die Qualität der Ergebnisse (Vgl. Bacon, Goethe). Im Falle mehrerer oder vieler gleichwertiger Ergebnisse kann der Anwender weitere Kriterien zur Festlegung der Ergebnisreihenfolge bzw. zur Modifizierung von Suchraum, Anfrage und Kontext definieren.

Das System verfügt über Verfahren die

- immer Relevanz im mathematischen Sinne gewährleisten
- die Ermittlung von relevanten Ergebnissen im Anwendersinn unterstützen (entsprechendes Sprachmodell und seine Anwendung auf Suchraum, Anfrage sowie Kontext der Anfrage)

Diese Anforderungen werden von klassischen Retrieval Systemen oder Suchmaschinen (inkl. Google) nicht erfüllt.

**Anmerkung:** der Begriff der Relevanz ist nicht ganz unproblematisch, da er in doppeltem Sinne gebraucht wird:

- mathematische Relevanz
- Relevanz in den Augen des Nutzers

Die Relation Anfrage und zugeordnete Ergebnisse kann als Relevanzrelation aufgefasst werden. Wenn zu einer beliebigen (noch so komplizierten) Anfrage alle Ergebnisse immer vollständig und reproduzierbar zugeordnet sind, kann man von Relevanz im mathematischen Sinne sprechen. Die Ergebnisse müssen deshalb nicht immer dem individuellen, persönlichen Empfinden von „relevantem Ergebnis“ entsprechen. Je präziser und schärfer die Anfrage (inkl. ihres sprachlichen Kontextes) desto präziser die Übereinstimmung von objektiver (mathematischer) Relevanz und subjektiv empfundener relevanter Ergebnismenge (siehe Zitate von Bacon: Eine kluge Frage ist die Hälfte der Weisheit und Goethe: Wenn du eine weise Antwort willst, musst du vernünftig fragen).

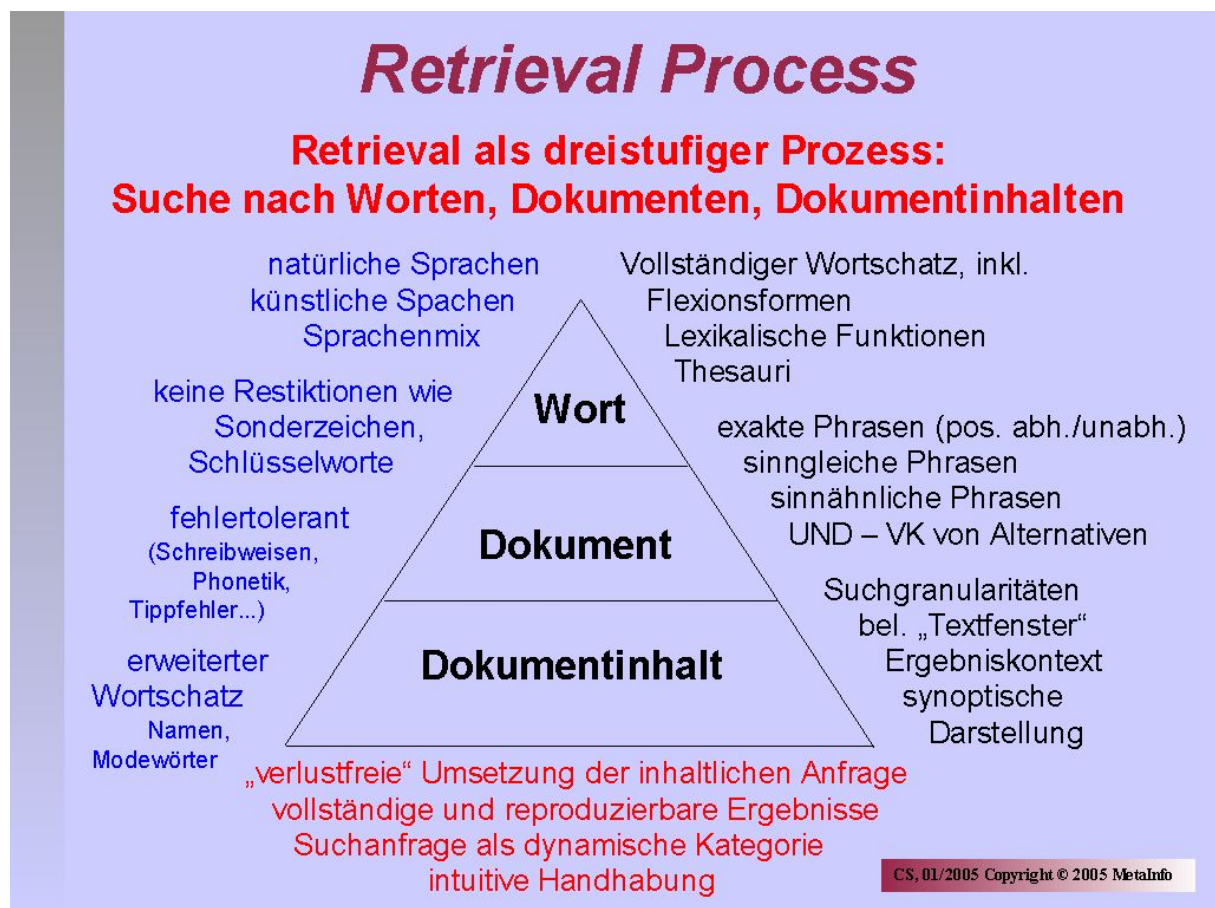
## Drei Schichten-Modell des Systems

Der Retrieval Prozess lässt sich als dreistufiger Prozess darstellen.

- Im 1. Schritt gilt es die geeigneten Wörter der Anfrage und ihren sprachlichen Kontext zu bestimmen, falls nicht Standardeinstellungen benützt werden sollen. Ebenso ist zu prüfen, ob evtl. eine Eingrenzung des Suchraums Sinn macht.
- Im 2. Schritt ermittelt das System automatisch alle Dokumente, die unabhängig vom geometrischen Kontext die Anfrage erfüllen können.
- Im 3. Schritt werden automatisch die kontextsensitiven Textpassagen ermittelt.

### Anmerkung:

Klassische Retrievalsysteme unterstützen Schritt 1 nicht bzw. nicht ausreichend. Schritt 3 wird gar nicht unterstützt.



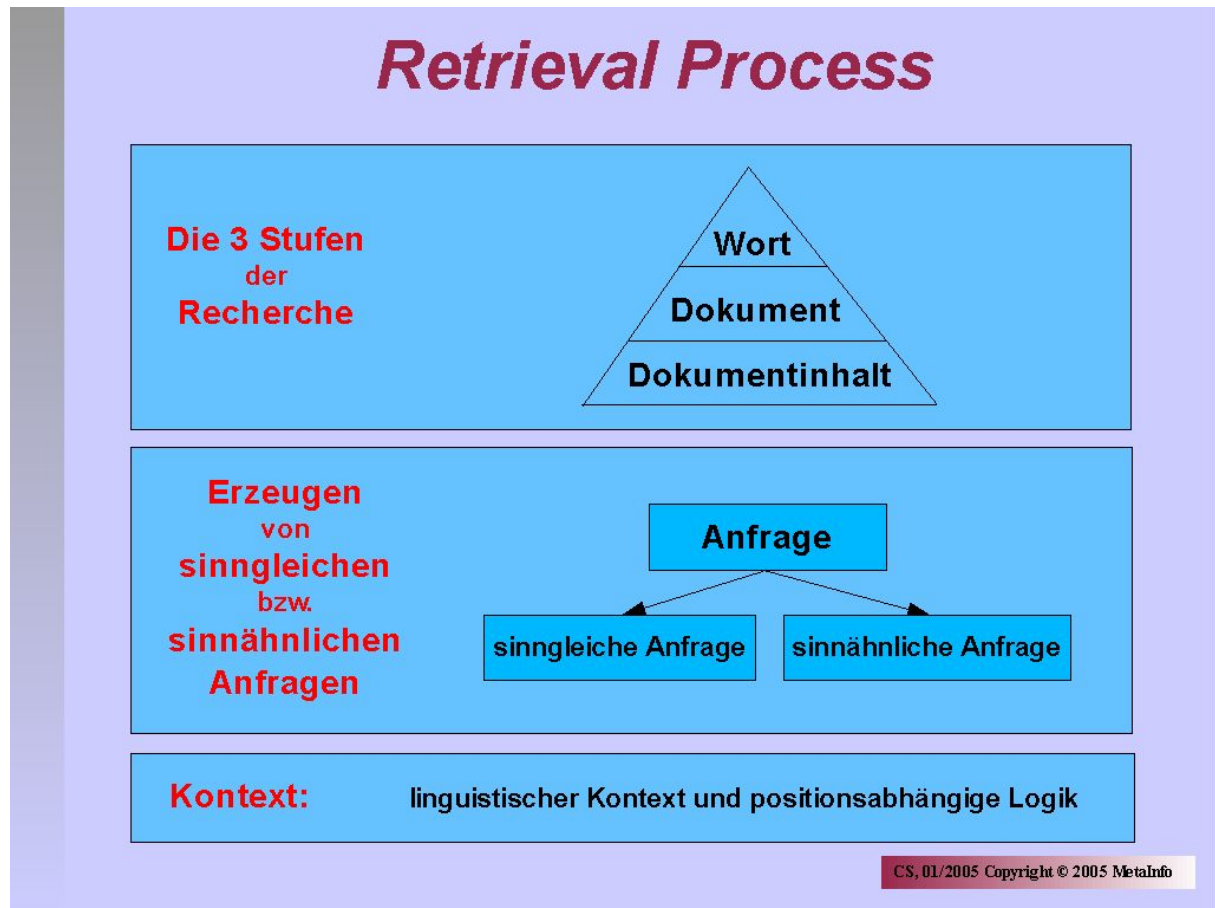
**Bild 7:** Retrieval als dreistufiger Prozess

### Pfeiler des Verfahrens

Das Verfahren besteht im Kern aus 3 wesentlichen Pfeilern:

- dreistufiger Retrieval Prozess: **Wörter** → **Dokumente** → **Dokumentinhalt**

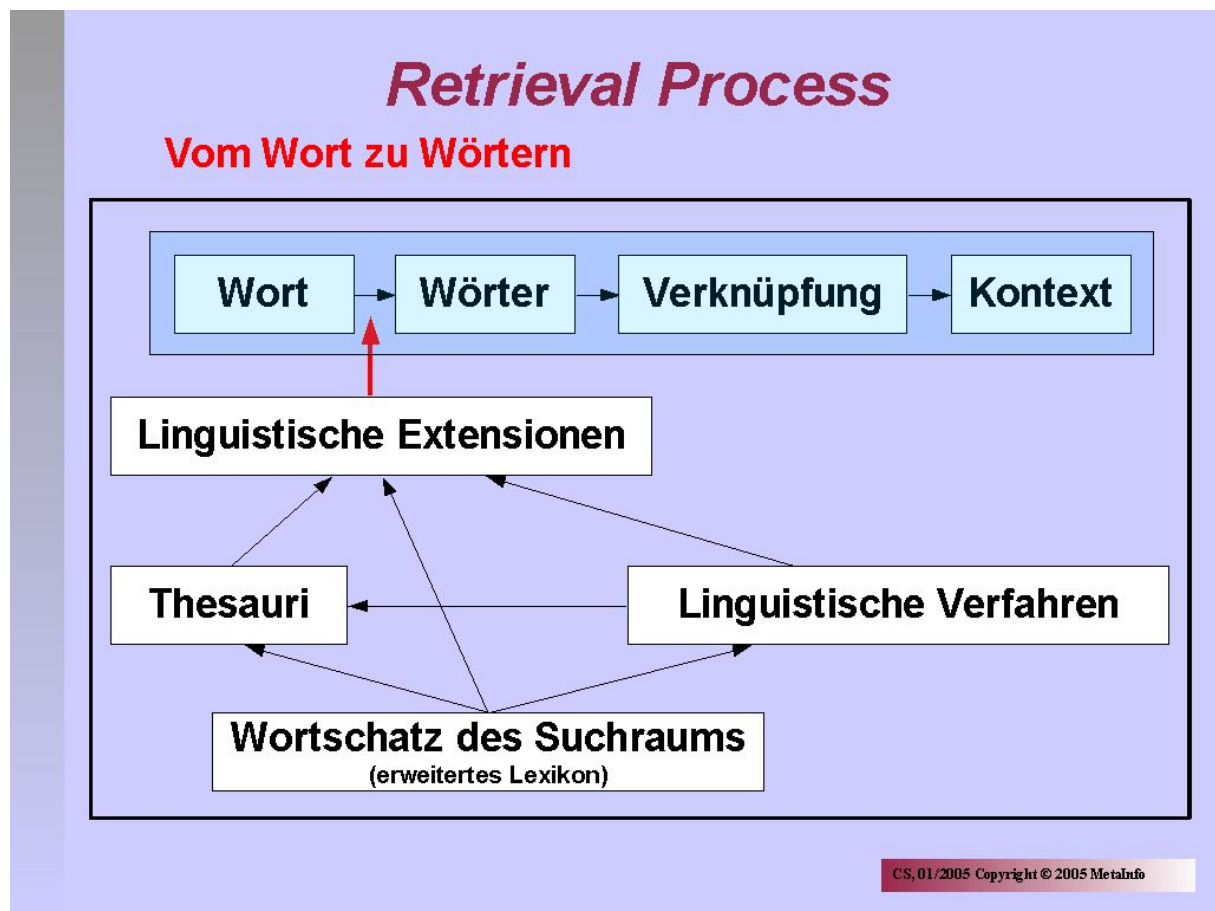
- linguistische Extension von Suchanfragen: Extension von Wörtern mit Hilfe interner und externer Hilfsmittel
- Berücksichtigung des sprachlichen Kontexts: Ausdrucksvielfalt und Sprachzusammenhang



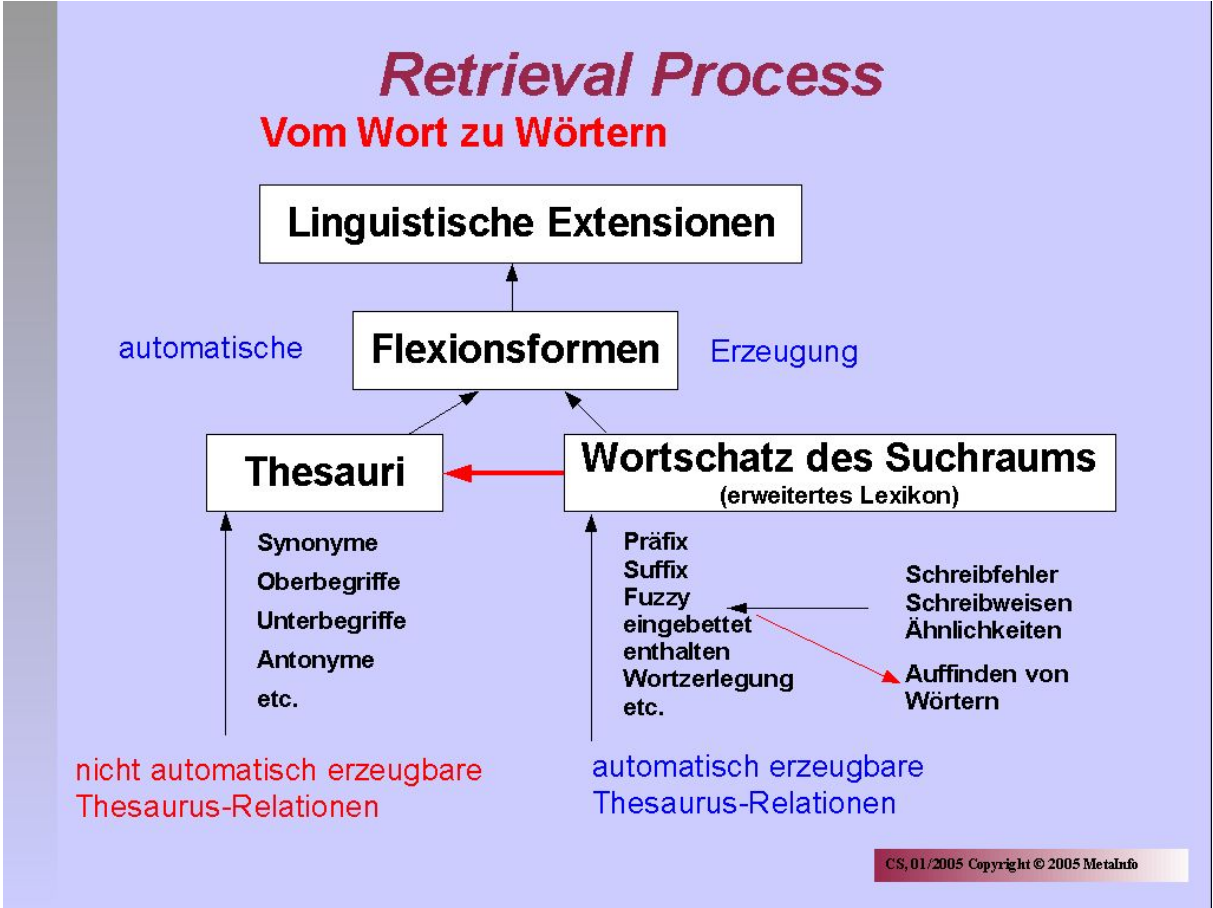
**Bild 8:** die drei Pfeiler des Retrieval Prozesses

Diese 3 Pfeiler ermöglichen es, der Vielfalt der sprachlichen Ausdrucksmöglichkeiten, auch wenn sie nicht der Schriftsprache entsprechen, d.h. fehlerhaft sind oder Modewörter, Namen etc. enthalten, vollständig Rechnung zu tragen. Das gilt insbesondere auch für multilinguale Texte.

## Erweiterte Lexika



**Bild 9:** Extension von Wörtern: Suchraum, Thesauri und linguistische Verfahren



**Bild 10:** Extension von Wörtern: Detaildarstellung

# *Retrieval Process*

## **„Wort“-Fuzzy**

- 1. Bedeutung/ Ausdruck:** Wörter ähnlicher Bedeutung
- 2. Schreibweise**
  - 2.1 Wortlänge, Anzahl Buchstaben  
(Buchstabe h, dehnende Vokale, Buchstabendopplung)
  - 2.2 Buchstabenfuzzy  
(z.B. b oder p)
  - 2.3 Buchstabenanordnung  
(Schreibfehler, Buchstabendreher)
- 3. Klang**
  - 3.1 ähnlich klingende Wörter  
(z.B. Rat / Tat)
  - 3.2 Aussprache von Fremdwörtern, Wörtern anderer Sprachen

CS, 01/2005 Copyright © 2005 MetaInfo

**Bild 11:** Wort Fuzzy: Unterschiedliche Einflüsse und Beiträge

## Bewertung bzw. Vergleich von Systemen

<i>Systemvergleich (Teil 1)</i>	
konventionelle IR-Systeme	Neuartige Suchtechnologie
<b>Modell:</b> Dokumente als reduzierte Kollektion von Attributen	Metastrukturen, viewunabhängige Repräsentationsformen
<b>Attribute:</b> Indexterme	alle Textelemente (Wörter inkl. Flexionen), Satzzeichen und Sonderzeichen
<b>Suche:</b> nach Dokumenten, eingeschränkte Suche in Dokus (Volltextsuche, NEAR-Operator)	nach Wörtern, nach Dokumenten und in Dokumenten (kontextsensitive Suche in Dokumenten)
<b>inh. Anfrage:</b> via Masken, boolsche Verknüpfung von Indextermen, spezielle Syntax, Keywords	Phrasen, UND-Verknüpfungen von Alternativen (intuitive Formulierung komplexer Logik), Ausschöpfung sprachlicher Möglichkeiten, automatische Extensionen
<b>Ergebnis:</b> relevante und nicht relevante Dokumente (Vollständigkeit der Ergebnisse ist nicht sichergestellt)	wissenschaftliche Recherche, d.h. vollständige und reproduzierbare Ergebnisse, (Dokumente und ihre entsprechenden Textstellen, nur relevante Ergebnisse im mathematischen Sinn), synoptische Ergebnisdarstellung, einstellbarer Ergebniskontext (direkte Auswertung)

CS, 01/2005 Copyright © 2005 Metabto

**Bild 12:** Vergleich konventionelle IR-Systeme und Neuartige Suchtechnologie

## *Systemvergleich (Teil 2)*

### Erweiterte Van-Rijsbergen-Matrix

	Data Retrieval	Information Retrieval	NST
<b>Matching</b>	<b>Exact match</b>	Partial match, best match	<b>complete and exact match</b>
<b>Inference Model</b>	Deduction	Induction	based on MS
<b>Classification</b>	Deterministic Monothetic	Probabilistic Polythetic	Metastructures <b>query as dynamic Classification</b>
<b>Query language</b>	Artificial	<b>Natural</b>	<b>Natural</b>
<b>Query specification</b>	<b>Complete</b>	Incomplete	<b>Complete, autom. extendable</b>
<b>Items wanted</b>	Matching	<b>Relevant</b>	<b>Relevant Matching</b>
<b>Error response</b>	Sensitive	Insensitive	Sensitive

**Legende:**

**NST: New Search Technology**  
**MS: Metastructures**

Die Neuartige Suchtechnologie (NST) vereinigt und erweitert die positiven Eigenschaften von Data Retrieval und Information Retrieval

CS, 01/2005 Copyright © 2005 MetaInfo

**Bild 13:** Erweiterte Van-Rijsbergen-Matrix: Vergleich von Data Retrieval und Information Retrieval mit Neuartiger Suchtechnologie

# Eigenschaften der Neuartigen Suchtechnologie

## Suchanfrage:

- sprachliches Interface
- Textfeld (Eingabematrix) statt Editfeld
- quasi keine Restriktionen für die Eingabe, keine reservierten Wörter
- UND-VK von Alternativen (intuitive Logik) oder Textinterpretation der Eingabe
- Eingabeunterstützende Funktionen (Eingabehilfen, linguistische Extensionen)
- erstellen von Abfrageprofilen und ihrer Parametrisierung: Suchstrategien, kaskadierende Suche, Batchverarbeitung, automatic readers
- Textmining

## Suchraum:

- parametrierbarer Suchraum
- Datenbanken und Datenbankgruppen
- dynamische Kategorien zur Definition von abgeleiteten Suchräumen

## Herzstück:

### Kernel:

- Infoexplizierung (Metastrukturen)
- sprachbasiertes Modell
- Schnittstellen zu internen bzw. externen Ressourcen wie z.B. Lexika oder Thesauri

### Suchalgorithmus:

- Granularität: Phrase, Wort, Wortbestandteil
- parametrisierbare Extensionsstrategien
- sprachlicher Kontext: linguistischer (z.B. Thesauri) und „örtlicher“ Kontext (positionsabhängige Logik, „Textfenster“)
- inhaltliches Ranking als **primäre Methode**

## Ergebnisaufbereitung:

- Ergebnis- oder Trefferliste: Parametrisierung der Einbettung von Fundstellen (Lesbarkeit ohne auf das Original zurückgehen zu müssen, synoptische Auswertungen)
- inhaltliches Ranking als primäre Methode („fälschungssicheres“ Ranking), verknüpfbar mit Häufigkeitsanalysen, Metadaten basierten Verfahren bzw. Sortierungen

- Verlinkung zur originären Information: Momentaufnahme des Originals (d.h. lokale Abbildung des Originals, inkl. Umsetzung von Linkstrukturen, optionalem download von Bildern, Markierung von Suchwörtern (und ihren Erweiterungen), sowie externe Links), externe Referenzierung (Original-URL)

## Hauptstärken des Systems

- Nicht nur Nachweis von Dokumenten (mit Lesprobe), sondern Suche von entsprechenden Textstellen in Dokumenten
- automatische, halbautomatische oder händische Erzeugung von single-chen bzw. sinnähnlichen Abfragen
- Funktioniert auch für Informationsmengen, die nicht via Experten oder Fachleute aufbereitet werden können: Internet oder Intranetapplikationen, große Informationsmengen, schnelllebige Informationen
- Bewährte Verfahren können einfach eingebunden werden, Technologische Bausteine können auch als Add On in bestehende Systeme integriert werden
- Annotierte Information oder Information aus Datenbanken können ebenfalls expliziert werden. Diese zusätzlichen Informationen lassen sich auch für die Suche nutzen (erweitertes Knowledge Management)
- Anfragen können manuell, halbautomatisch oder automatische extendiert werden
- Auswertung der Anfrage im parametrisierbarem linguistischen Kontext: Textfenster und sprachliche Extensionen
- Skalierbarkeit: Internet- bzw. Intranet als Supercomputer
- Unterstützung unterschiedlicher Architekturen und Plattformen

## Grenzen herkömmlicher Systeme

Die Extension von Wörtern um Synonyme, Oberbegriffe, Unterbegriffe (und ihrer Flexionsformen) oder um Wörter, die Bestandteile eines Wortes enthalten führt sehr schnell auch bei einfachen Abfragen zu enormer inhaltlicher Vielfalt. Nachfolgendes Beispiel erläutert das an Hand der Zerlegung eines zusammengesetzten Wortes.

**Retrieval Process**

**Grenzen herkömmlicher Systeme:**

**Auswertung eines zusammengesetzten Wortes:**

Beispiel: Suchraum Wikipedia  
Wort: Kinder-arbeit:  
Extension: Kinder\* --> 728 (1. Zeile)  
\*arbeit\* --> 3825 (2. Zeile)  
Matrix: UND-Verknüpfung von Alternativen  
Kontext: 20 Wörter

Google-Anfragen:  $728 \times 3825 = 2\,784\,600$  Anfragen  
Eingabezeit:  $2\,784\,600 \times 2 \text{ sec} = 64,5$  Tage (a 24 Std)

**Schon die Formulierung der Anfrage wäre de facto nicht durchführbar!**

CS, 01/2005 Copyright © 2005 MetaInfo

**Bild 14:** Ein einfaches Textmining Beispiel

Zusammengesetzte Wörter haben gewissermaßen eine synergetische Bedeutung, die sich aus den einzelnen Wortbestandteilen ergibt. Dabei spezifizieren nachfolgende Begriffe meist den vorherigen Begriff. Diesen einfachen Sachverhalt kann man z.B. für Textmining-Algorithmen nutzen. Das ist mit herkömmlichen Systemen nicht oder nur sehr eingeschränkt möglich.

## **Beschreibung als Filtersystem**

Die neuartige Suchtechnologie besteht aus 3 fundamentalen Komponenten  
Das Herzstück der neuartigen Suchtechnologie lässt sich auch als Filtersystem beschreiben.

### **1. Beschaffungskomponente (Suchmaschinen, Crawler etc.)**

- Informationsbeschaffung
- Aufbereitung

(maximale Informationsmenge)

### **2. Filterkomponente**

- Suchraumdefinition
- (natürlich sprachliche) Anfrageformulierung
- Anfrageparametrisierung (Extension, Kontext)
- Suchstrategien (z.B. mehrfache Filterung, komplexe logische Verknüpfungen von Filterergebnissen, „Automatic Readers“)

### **3. Ergebnispräsentationskomponente**

- Ranking: immer inhaltliches Ranking d.h. nur relevante Resultate im Sinne der Anfrage, plus Sortieren bzw. Ordnen nach Suchräumen und Metadaten bzw. annotierten Informationen/Daten
- Ergebnisaufbereitung (Fundstellen, Einbettung von Fundstellen, Ergebnismetadaten, Navigation zum Original, Hervorhebungen im Original)
- Vorabergebnisse im Falle sehr großer Datenmengen