Basisinformationstechnologie II

Sommersemester 2023. 5: Text I

Klassifikation & Strukturierung. *Basierend auf Jan Wieners' Folien*

Institut für Digital Humanities, Historisch-Kulturwissenschaftliche Informationsverarbeitung | Prof. Dr. Øyvind Eide | Slavina Stoyanova

Text: Aspekte

- Textklassifikation
  - Natürliche bzw. unstrukturierte Texte
  - Semistrukturierte Texte
  - Strukturierte Texte
- Textstrukturierung: XML
  - Ein XML-Standard: die Text Encoding Initiative
- Information Retrieval: Inhalte auffinden, clustern, etc.

Textklassifikation

HARUKI MURAKAMI

Der Himmel hat die Erde still geküsst

Der japanische Weltautor Haruki Murakami setzt mit dem zweiten Band von "1Q84" sein ironisches Weltverklärungswerk fort.

Das Jahr neigt sich seinem Ende zu. Es ist kalt in Tokyo. Der Verkehr auf den Stadtautobahnen fließt stockend und träge vor sich hin. Der Mond steht am Winterhimmel und sieht aus wie ein Häufchen weiße Asche. Das Liebespaar Tengo und Aomame hat sich endlich gefunden. Zwanzig Jahre haben sie nacheinander gesucht und aufeinander gewartet. Jetzt stehen sie nach ihrer ersten Liebesnacht in der Juniorsuite eines Tokyoer Hotels in weißen Hotelbademänteln, halten sich an der Hand und sehen in den Himmel. Sie wollen sich nie wieder trennen.

Liebe, die glückt, gibt es entweder im Leben oder im Trivialroman. Im anspruchsvollen literarischen Segment trifft man sie selten. Wer für einander bestimmt ist, verfehlt sich im gehobenen Liebesroman mit großer Wahrscheinlichkeit, scheitert, kommt zu früh, zu spät, verstrickt sich in den Schlingen gesellschaftlicher Zwänge und der eigenen Wünsche. Es gibt viele ergreifende Liebesromane, aber kaum eine Ausnahme von der Regel, dass sie unglücklich zu enden haben, um literarisch zu reüssieren.

VON Iris Radisch

DATUM 01.01.2012 - 19:31 Uhr

SEITE 1 | 2 | Auf einer Seite lesen

QUELLE DIE ZEIT, 29.12.2011 Nr. 01

KOMMENTARE 4

VERSENDEN E-Mail verschicken

EMPFEHLEN Facebook, Twitter, Google+

AUTOREN ABONNIEREN RSS-Feed

ARTIKEL DRUCKEN Druckversion | PDF

SCHLAGWORTE Haruki Murakami | Mond | New York

NEU AUF ZEIT ONLINE

1. **DAMMBRUCH IN FISCHBECK** ICE-Hauptstrecken lahmgelegt
2. **WAHLKAMPF** Steinbrück entlässt seinen Sprecher
3. **NSA-ÜBERWACHUNGSAFFÄRE** Bundesregierung fordert Aufklärung von Obama
4. **LITERATUR** Philologe Walter Jens ist tot
5. **ENTWICKLERKONFERENZ WWDC** Warten auf das nächste große Ding von Apple

NEU IM RESSORT

1. **LITERATUR** Philologe Walter Jens ist tot
2. **YORAM KANIUK** Er trotzte der Erniedrigung
3. **200 JAHRE RICHARD WAGNER** Der Anarchist vom Grünen Hügel

Haruki MurakamiDer Himmel hat die Erde still geküsst
Der japanische Weltautor Haruki Murakami setzt mit de
Band von "1Q84" sein ironisches Weltverklärungswerk f
Das Jahr neigt sich seinem Ende zu. Es ist kalt in To
Verkehr auf den Stadtautobahnen fließt stockend und t
sich hin. Der Mond steht am Winterhimmel und sieht au
Häufchen weiße Asche. Das Liebespaar Tengo und Aomame
endlich gefunden. Zwanzig Jahre haben sie nacheinande
und aufeinander gewartet. Jetzt stehen sie nach ihrer
Liebesnacht in der Juniorsuite eines Tokyoer Hotels i
Hotelbademänteln, halten sich an der Hand und sehen i
Himmel. Sie wollen sich nie wieder trennen.
Liebe, die glückt, gibt es entweder im Leben oder im
Trivialroman. Im anspruchsvollen literarischen Segmen
man sie selten. Wer für einander bestimmt ist, verfeh
im gehobenen Liebesroman mit großer Wahrscheinlichkei
scheitert, kommt zu früh, zu spät, verstrickt sich in
Schlingen gesellschaftlicher Zwänge und der eigenen W
Es gibt viele ergreifende Liebesromane, aber kaum ein
Ausnahme von der Regel, dass sie unglücklich zu enden
um literarisch zu reüssieren.

Haruki MurakamiDer Himmel hat die Erde still geküsst
Der japanische Weltautor Haruki Murakami setzt mit de
Band von "1Q84" sein ironisches Weltverklärungswerk f
Das Jahr neigt sich seinem Ende zu. Es ist kalt in To

```
<text>
    <autorin>Iris Radisch</autorin>
    <adresse>http://www.zeit.de/2012/01/L-Murakami</adresse>
    <ueberschrift>
        Der Himmel hat die Erde still geküsst
    </ueberschrift>
    <inhalt art="Rezension">
        <teaser>
            Der <ort>japanische</ort> Weltautor <autorname>Haruki Murakami</autorname>
            setzt mit dem zweiten Band von <werktitel>"1Q84"</werktitel>
            sein ironisches Weltverklärungswerk fort.
        </teaser>
        <haupttext>
            Das Jahr neigt sich seinem Ende zu. Es ist kalt in <ort>Tokyo</ort>.
            Der Verkehr auf den Stadtautobahnen fließt stockend und träge vor sich hin.
        </haupttext>
    </inhalt>
</text>
```

Schlingen gesellschaftlicher Zwänge und der eigenen W
Es gibt viele ergreifende Liebesromane, aber kaum ein
Ausnahme von der Regel, dass sie unglücklich zu enden
um literarisch zu reüssieren.

Textklassifikation

Die Strukturiertheit von Texten:
(Text von lat. textus: Gewebe, Geflecht)

- Natürliche und unstrukturierte Texte
  Beispiel: „Vor dieser Burleske frühkapitalistischen Übereifers flohen die coolen Kinder der Nachkriegsgeneration zu Beginn der achtziger Jahre in ein reptilienartiges Singledasein mit minimalen Ausschlägen.“
  (http://www.zeit.de/2012/01/L-Murakami)

- Strukturierte Texte
  Beispiel: MySQL-DB, XML

- Semistrukturierte Texte
  Beispiel: HTML → Was bezeichnet ein bestimmtes HTML-Tag? Werden Standards in der Auszeichnung eingehalten?

HARUKI MURAKAMI

Der Himmel hat die Erde still geküsst

Der japanische Weltautor Haruki Murakami setzt mit dem zweiten Band von "1Q84" sein ironisches Weltverklärungswerk fort.

Das Jahr neigt sich seinem Ende zu. Es ist kalt in Tokyo. Der Verkehr auf den Stadtautobahnen fließt stockend und träge vor sich hin. Der Mond steht am Winterhimmel und sieht aus wie ein Häufchen weiße Asche. Das Liebespaar Tengo und Aomame hat sich endlich gefunden. Zwanzig Jahre haben sie nacheinander gesucht und aufeinander gewartet. Jetzt stehen sie nach ihrer ersten Liebesnacht in der Juniorsuite eines Tokyoer Hotels in weißen Hotelbademänteln, halten sich an der Hand und sehen in den Himmel. Sie wollen sich nie wieder trennen.

Liebe, die glückt, gibt es entweder im Leben oder im Trivialroman. Im anspruchsvollen literarischen Segment trifft man sie selten. Wer für einander bestimmt ist, verfehlt sich im gehobenen Liebesroman mit großer Wahrscheinlichkeit, scheitert, kommt zu früh, zu spät, verstrickt sich in den Schlingen gesellschaftlicher Zwänge und der eigenen Wünsche. Es gibt viele ergreifende Liebesromane, aber kaum eine Ausnahme von der Regel, dass sie unglücklich zu enden haben, um literarisch zu reüssieren.

VON Iris Radisch

DATUM 01.01.2012 - 19:31 Uhr

| 2 | Auf einer Seite lesen

QUE E ZEIT, 29.12.2011 Nr. 01

KOMM E 4

VERS E-Mail verschicken

EMP N Facebook, Twitter, Google+

EN ABONNIEREN RSS-Feed

ARTIKEL DRUCKEN Druckversion | PDF

SCHLAGWORTE Haruki Murakami | Mond | New York

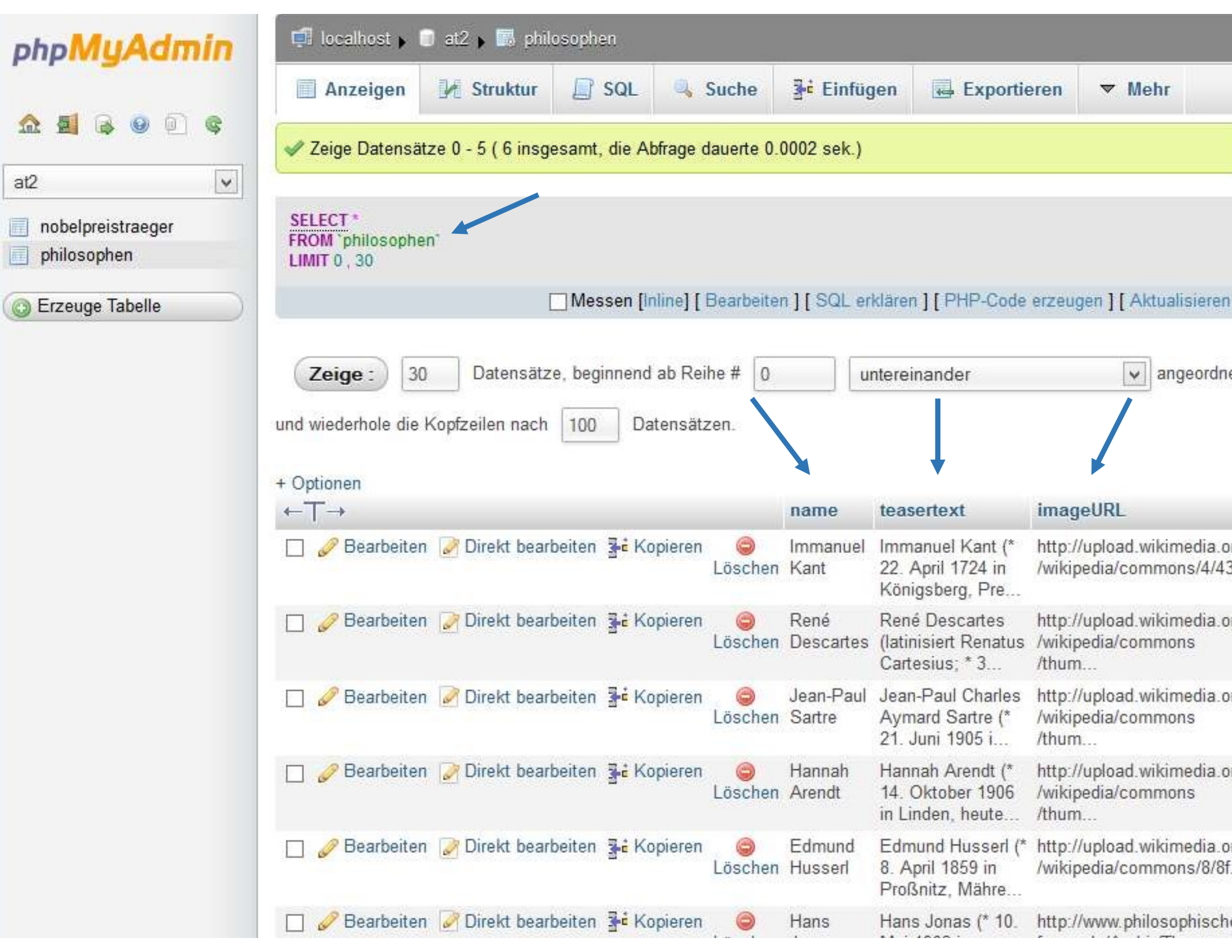
NEU AUF ZEIT ONLINE

AMMBRUCH IN FISCHBECK ICE-Hauptstrecken lahmgelegt

2. WAHLKAMPF Steinbrück entlässt seinen Sprecher
3. NSA-ÜBERWACHUNGSAFFÄRE Bundesregierung fordert Aufklärung von Obama
4. LITERATUR Philologe Walter Jens ist tot
5. ENTWICKLERKONFERENZ WWDC Warten auf das nächste große Ding von Apple

NEU IM RESSORT

1. LITERATUR Philologe Walter Jens ist tot
2. YORAM KANIUK Er trotzte der Erniedrigung
3. 200 JAHRE RICHARD WAGNER Der Anarchist vom Grünen Hügel

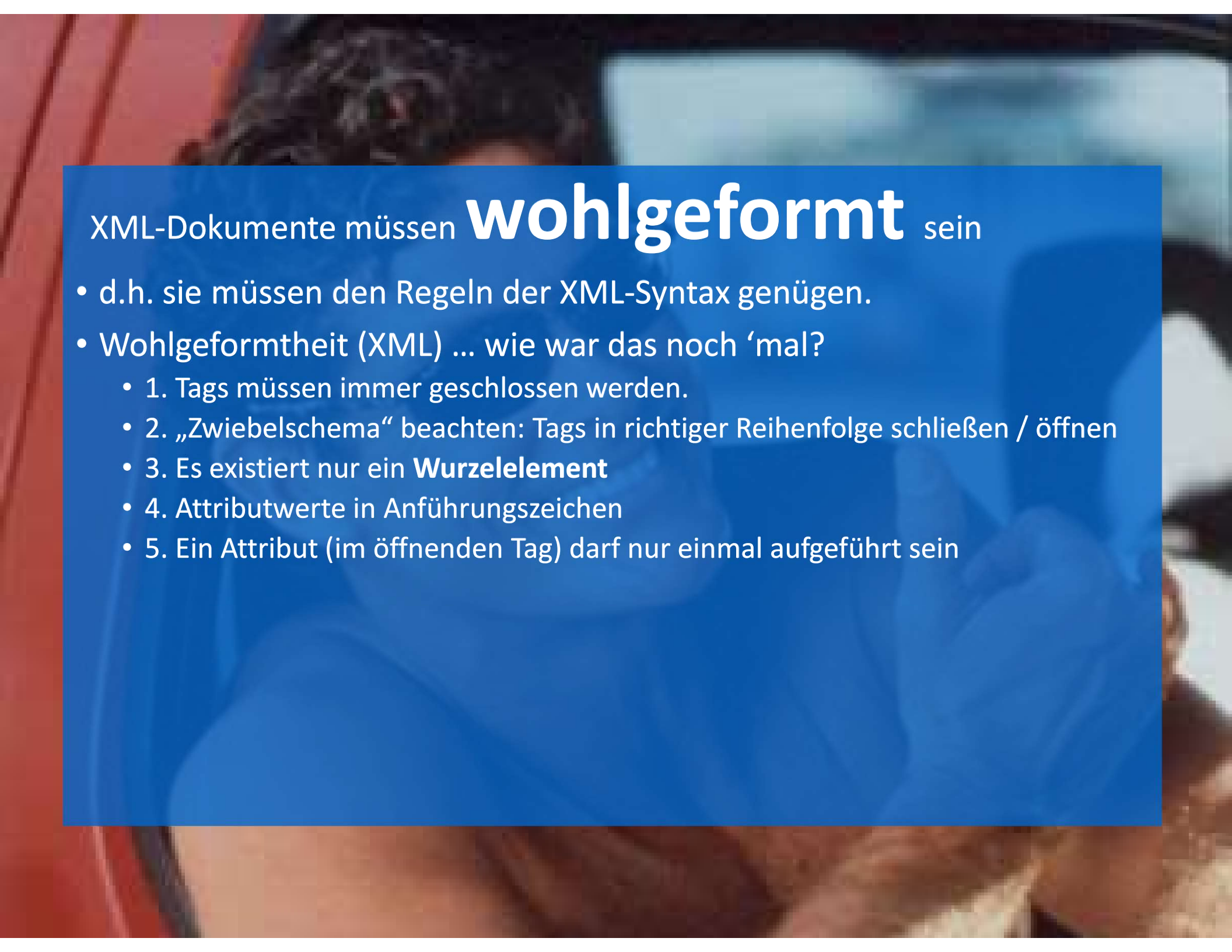

phpMyAdmin
at2
nobelpreistraeger
philosophen
Erzeuge Tabelle
localhost
at2
philosophen
Anzeigen
Struktur
SQL
Suche
Einfügen
Exportieren
Mehr
Zeige Datensätze 0 - 5 (6 insgesamt, die Abfrage dauerte 0.0002 sek.)
SELECT *
FROM `philosophen`
LIMIT 0 , 30
Messen [Inline] [Bearbeiten] [SQL erklären] [PHP-Code erzeugen] [Aktualisieren
Zeige :
30
Datensätze, beginnend ab Reihe #
0
untereinander
angeordne
und wiederhole die Kopfzeilen nach
100
Datensätzen.
+ Optionen
name
teasertext
imageURL
Bearbeiten
Direkt bearbeiten
Kopieren
Löschen
Immanuel Kant
Immanuel Kant (* 22. April 1724 in Königsberg, Pre...
http://upload.wikimedia.o
/wikipedia/commons/4/43
René Descartes
René Descartes (latinisiert Renatus Cartesius; * 3...
http://upload.wikimedia.o
/wikipedia/commons
/thum...
Jean-Paul Sartre
Jean-Paul Charles Aymard Sartre (* 21. Juni 1905 i...
http://upload.wikimedia.o
/wikipedia/commons
/thum...
Hannah Arendt
Hannah Arendt (* 14. Oktober 1906 in Linden, heute...
http://upload.wikimedia.o
/wikipedia/commons
/thum...
Edmund Husserl
Edmund Husserl (* 8. April 1859 in Proßnitz, Mähre...
http://upload.wikimedia.o
/wikipedia/commons/8/8f
Hans
Hans Jonas (* 10.
http://www.philosophisch

Strukturierte Texte

Extensible Markup Language (XML)

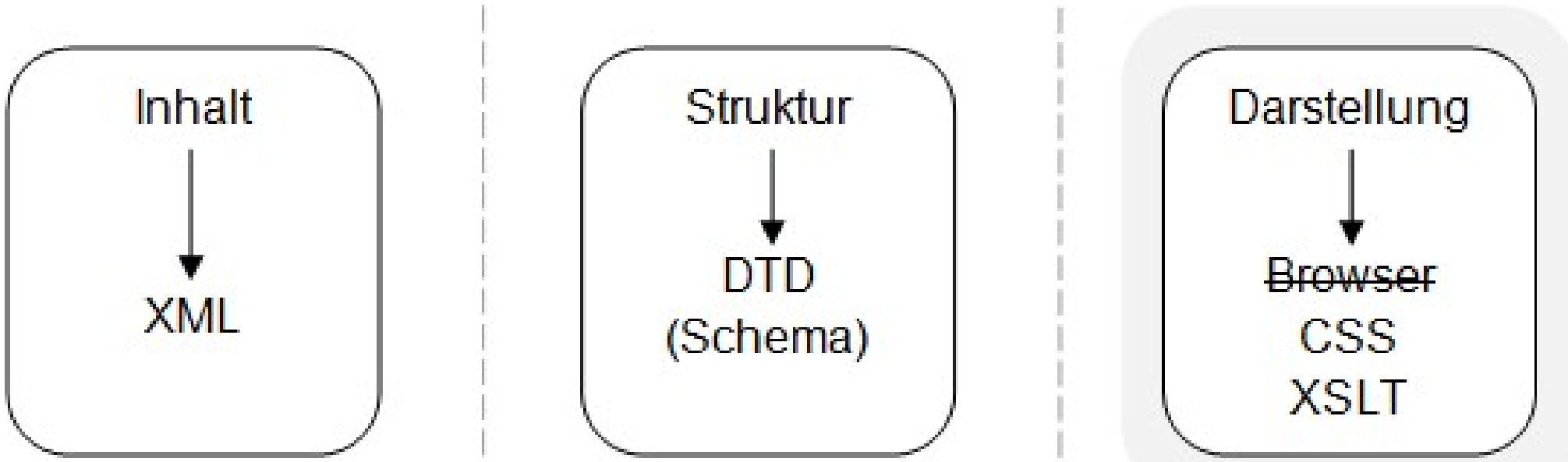
- Standard Generalized Markup Language (SGML)
- Tags
- Attribute und Attributwerte
- Wohlgeformtheit von XML-Dokumenten
- Validität / Gültigkeit von XML-Dokumenten
- Schemata
- Transformation von XML-Dokumenten


XML-Dokumente müssen **wohlgeformt** sein…

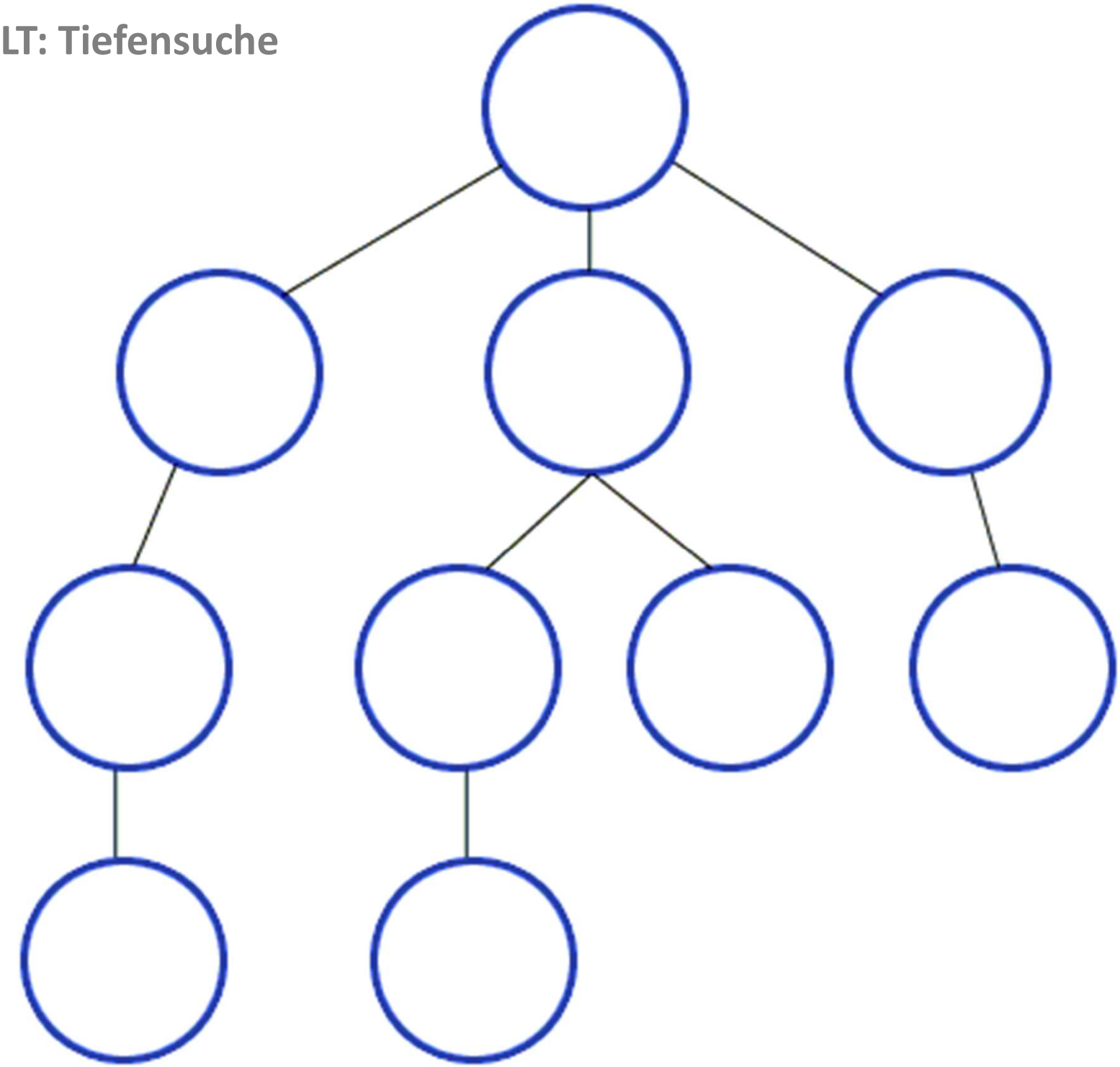


XML-Dokumente müssen **wohlgeformt** sein

- d.h. sie müssen den Regeln der XML-Syntax genügen.
- Wohlgeformtheit (XML) … wie war das noch ‘mal?
  - 1. Tags müssen immer geschlossen werden.
  - 2. „Zwiebelschema“ beachten: Tags in richtiger Reihenfolge schließen / öffnen
  - 3. Es existiert nur ein **Wurzelelement**
  - 4. Attributwerte in Anführungszeichen
  - 5. Ein Attribut (im öffnenden Tag) darf nur einmal aufgeführt sein


Inhalt
XML
Struktur
DTD
(Schema)
Darstellung
~~Browser~~
CSS
XSLT

XSLT: Tiefensuche



```
<text>
    <autorin>Iris Radisch</autorin>
    <adresse>http://www.zeit.de/2012/01/L-Murakami</adresse>
    <ueberschrift>
        Der Himmel hat die Erde still geküsst
    </ueberschrift>
    <inhalt art="Rezension">
        <teaser>
            Der <ort>japanische</ort> Weltautor <autorname>Haruki Murakami<
            setzt mit dem zweiten Band von <werktitel>"1Q84"</werktitel>
            sein ironisches Weltverklärungswerk fort.
        </teaser>
        <haupttext>
            Das Jahr neigt sich seinem Ende zu. Es ist kalt in <ort>Tokyo</
            Der Verkehr auf den Stadtautobahnen fließt stockend und träge v
        </haupttext>
    </inhalt>
</text>
```

bestimmt ist, verfehlt sich im gehobenen Liebesroman mit großer Wahrscheinlichkeit, scheitert, kommt zu früh, zu spät, verstrickt sich in den Schlingen gesellschaftlicher Zwänge und der eigenen Wünsche. Es gibt viele ergreifende Liebesromane, aber kaum eine Ausnahme von der Regel, dass sie unglücklich zu enden haben, um literarisch zu reüssieren.

grüne Ding von Apple

NEU IM RESSORT

1. **LITERATUR** Philologe Walter Jens ist tot
2. **YORAM KANIUK** Er trotzte der Erniedrigung
3. **200 JAHRE RICHARD WAGNER** Der Anarchist vom Grünen Hügel

Einheitliches Strukturieren: Standards

Hugo von Hofmannsthal – Die Beiden

Sie trug den Becher in der Hand
– Ihr Kinn und Mund glich seinem Rand –,
So leicht und sicher war ihr Gang,
Kein Tropfen aus dem Becher sprang.

So leicht und fest war seine Hand:
Er ritt auf einem jungen Pferde,
Und mit nachlässiger Gebärde
Erzwang er, daß es zitternd stand.

Jedoch, wenn er aus ihrer Hand
Den leichten Becher nehmen sollte,
So war es beiden allzu schwer:
Denn beide bebten sie so sehr,
Daß keine Hand die andre fand
Und dunkler Wein am Boden rollte.

Hugo von Hofmannsthal – Die Beiden

Vers

Sie trug den Becher in der Hand
– Ihr Kinn und Mund glich seinem Rand –,
So leicht und sicher war ihr Gang,
Kein Tropfen aus dem Becher sprang.

So leicht und fest war seine Hand:
Er ritt auf einem jungen Pferde,
Und mit nachlässiger Gebärde
Erzwang er, daß es zitternd stand.

Strophe

Jedoch, wenn er aus ihrer Hand
Den leichten Becher nehmen sollte,
So war es beiden allzu schwer:
Denn beide bebten sie so sehr,
Daß keine Hand die andre fand
Und dunkler Wein am Boden rollte.

330 Broome Oct. 24

Wednesdays at Ⓐ's

p r e s e n t s

performances by Kirsten Hawthorne, Claire Fergusson, Joe Lewis and special Guest Clowns

color xerox works by Cleveland, Higgins, Hawthorne, Jorgensen, Astrom, Miller, Evans, Buczak, Smagma, Wilson, Schneemann, Dr. MS / Richwood, Avery, Gordon, Keith, Chait, SAMO© Oistreanu, et al

music by Mania D.

BYO

8 P.M. $3

Ein Standard von dem man unbedingt ‘mal gehört haben muss: Die Text Encoding Initiative (TEI)

< Text Encoding Initiative >

Geschichte:

- 1987 entstanden als internationale Initiative von Philologinnen und Philologen
- Dokumentenformat zur Repräsentation von Texten in digitaler Form
- Vielseitigkeit & Praxisnähe

Differenzierung: TEI bezeichnet sowohl

- das **Konsortium** (TEI-C), 2000 gegründet
- als auch **Richtlinien** und **Empfehlungen** zur Kodierung und zum Austausch von Textdokumenten.
  Intention: Geisteswissenschaftlerinnen und Geisteswissenschaftler sollen über größtmögliche Freiheit verfügen, textuell vorliegende Information nach eigenem Textbegriff in XML zu codieren.

Versionsgeschichte

- 1990: TEI P1 (P => Proposal, Entwurf / Plan)
  Basiert auf SGML (Standard Generalized Markup Language)
- 1992 / 1993: TEI P2
- 1994: TEI P3 ("Green Books")
- 2002: TEI P4 (XML-basiert)
- 2002: TEI Lite
- 2007 TEI P5

TEI P5:
Guidelines for Electronic Text Encoding and Interchange

by the TEI Consortium

Originally edited by C.M. Sperberg-McQueen and Lou Burnard for the ACH-ALLC-ACL Text Encoding Initiative
Now entirely revised and expanded under the supervision of the Technical Council of the TEI Consortium

http://www.tei-c.org/release/doc/tei-p5-doc/en/Guidelines.pdf (**1646** Seiten)

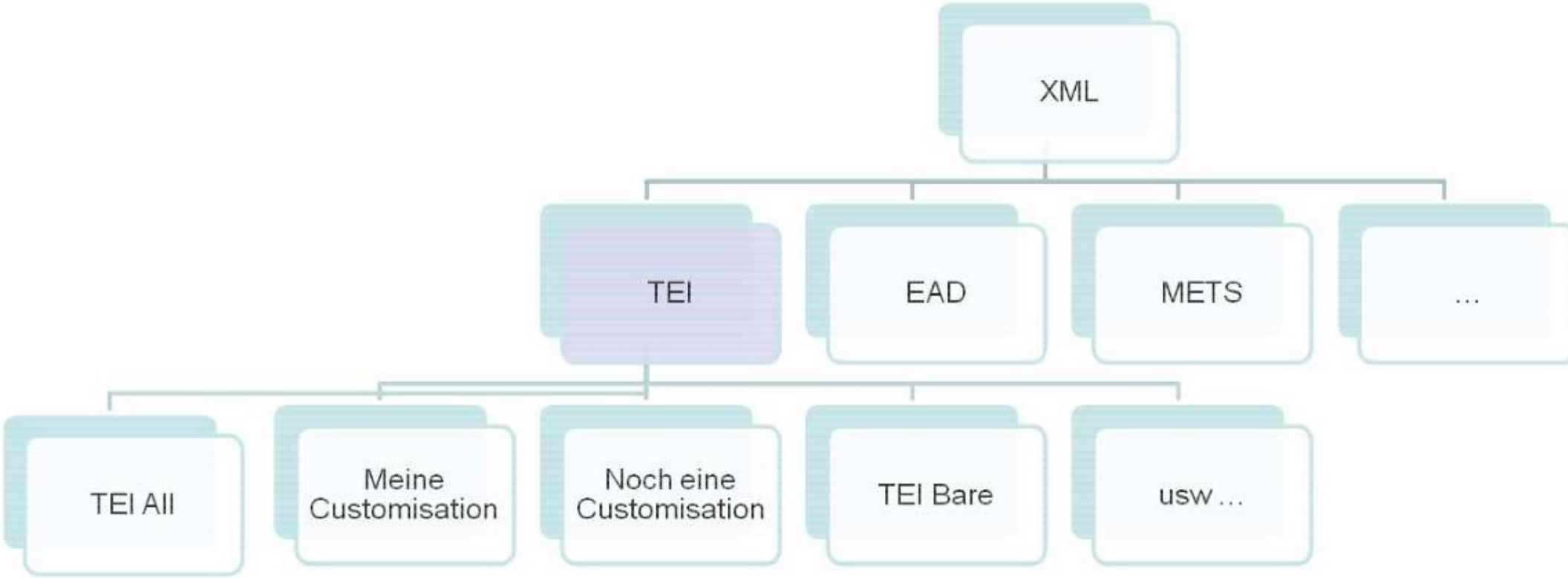

MakeAGIF.com

Modularisierung

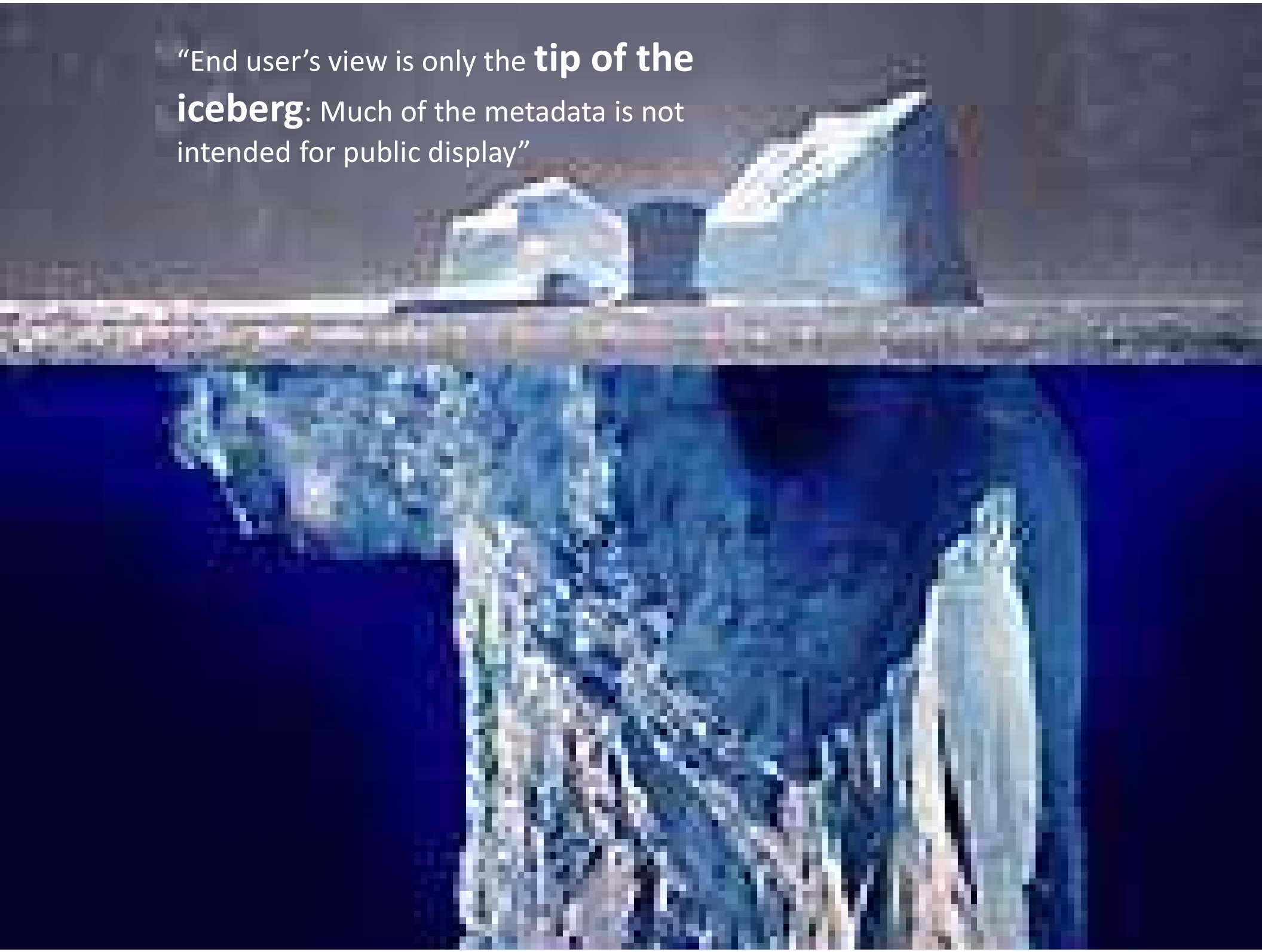
- Flexible Auswahl von TEI-Elementen aufgrund des modularen Charakters der TEI. So muss ein eigenes Schema nicht alle Elemente und Attribute der TEI enthalten.

- Module, u.a.:
  - core für Basiselemente
  - header für Metadaten
  - textstructure für grundlegende Textstrukturen
  - drama für Dramen
  - prose, poetry, etc.

Verortung:



EAD: Encoded Archival Description
METS: Metadata Encoding and Transmission Standard


"End user's view is only the **tip of the iceberg**: Much of the metadata is not intended for public display"

Arten von Metadaten (vgl. Witten, Bainbridge, Nichols (2010): How to Build a Digital Library):

- **Administrative** metadata for managing resources, such as rights information
- **Descriptive** metadata for describing resources (Beispiel: Zettel des Zettelkataloges)
- **Preservation** metadata for describing resources, such as recording preservation actions
- **Technical** metadata related to low-level system information, such as data formats and any data compression used
- **Usage** metadata related to system use, such as tracking user behavior

→ "End user's view is only the **tip of the iceberg**: Much of the metadata is not intended for public display"


Hf – Hog
Jak –
Jo –
Hoj – Hz

MARC: MAchine Readable Cataloging

- Vorgestellt in den späten 1960er Jahren von Henriette Avram (Library of Congress)

- !Grundproblem/-intention: Migration von Zettelkatalogen zu computerbasierter Repräsentation von Datensätzen (Records)

- MARC-Datensätze gespeichert als Sammlung von Feldern in einem „ziemlich komplexen Format“
  [Witten, Bainbridge, Nichols (2010): How to Build a Digital Library]

- “Producing a MARC record for a particular publication is an onerous undertaking that is governed by a detailed set of (highly detailed) rules and guidelines called the Anglo-American Cataloging Rules (AACR2R, 2R → final revised 2nd edition).”
  [Witten, Bainbridge, Nichols (2010): How to Build a Digital Library]

Maschinelles Austauschformat für Bibliotheken (MAB),

- MAB → 1970er, Deutsche Nationalbibliothek
- MAB2 → 1990er
- Verwendung mit RAK (Regeln zur Alphabetischen Katalogisierung)

DEUTSCHE NATIONAL BIBLIOTHEK

English | Kontakt | A-Z | Förderer | Datenschutz | Impressum | Hilfe | Mein Konto

LEIPZIG FRANKFURT AM MAIN

Home // Standardisierung // Formate und Schnittstellen // MAB

- Home
- Wir über uns
- Aktuell
- Kataloge
- Erwerbung
- Netzpublikationen
- Service
- Standardisierung
  - Arbeitsstelle für Standardisierung
  - Internationalisierung
  - Regelwerke und Arbeitshilfen
  - Formate und Schnittstellen

MAB

MAB ist das Maschinelle Austauschformat für Bibliotheken. Mit MAB können alle im Bibliotheksbereich erzeugten Daten ausgetauscht werden: bibliografische Daten, Norm- und Lokaldaten.

Die Entwicklung und Pflege von MAB (seit 1995 MAB2) ist 2006 abgeschlossen worden, das Format wurde »eingefroren«. Änderungen oder Erweiterungen sind damit in MAB nicht mehr möglich. Dies gilt sowohl für das Datendienstformat (früher »Bandformat«) als auch für das Diskettendienstformat (oder »Diskettenformat«). Als Nachfolger dient das MARC-21-Format, der Zielstandard des Projektes »Umstieg auf MARC 21«.

Die Auslieferung von MAB2-Daten durch die Deutsche Nationalbibliothek endet am 30. Juni 2013.

Im Jahr 2012 wurden die zuvor bestehenden Normdateien Personennamendatei (PND), Schlagwortnormdatei (SWD) und Gemeinsame Körperschaftsdatei (GKD) sowie die Einheitssachtitel-Datei des Deutschen Musikarchivs in einer Gemeinsamen Normdatei (GND) zusammengeführt. Das GND-Format ist für den Austausch ausschließlich in MARC 21 definiert.

Das Format MAB2 besteht aus fünf einzelnen Datenformaten, die auf einer einheitlichen, integrierten und für alle Formate gültigen Feldstruktur aufsetzen, den Segmenten 0-- (txt, 34KB, Datei ist nicht barrierefrei). Diese Einzelformate sind:

- MAB-Format für bibliografische Daten - MAB2-TITEL Online-Kurzreferenz-Version (November 2001) (txt, 44KB, Datei ist nicht barrierefrei)
- MAB-Format für Personennamen - MAB2-PND Online-Kurzreferenz-Version (November 2001) (txt, 8KB, Datei ist nicht barrierefrei)
- MAB-Format für Körperschaftsnamen - MAB2-GKD Online-Kurzreferenz-Version (November 1998) (txt, 5KB, Datei ist nicht barrierefrei)

- Variable Control Fields (00x)
- Variable Data Fields
  - Numbers and Codes (0xx)
  - Main Entries (1xx)
  - Titles (2xx)
  - Edition, Imprint, etc. (2xx)
  - Physical Description, etc. (3xx)
  - Series Statements (4xx)
  - Notes (5xx)
  - […]

→ Vgl. http://catalog2.loc.gov
sowie die Referenz unter http://www.loc.gov/ma

MARCXML-Darstellung

```
- <record>
    <leader>00693cam a2200217u 4500</leader>
    <controlfield tag="001">10462695</controlfield>
    <controlfield tag="005">20030718130135.0</controlfield>
    <controlfield tag="008">841013r19681788be 000 0 ger </controlfield>
  - <datafield tag="035" ind1=" " ind2=" ">
      <subfield code="9">(DLC) 79459272</subfield>
    </datafield>
  - <datafield tag="906" ind1=" " ind2=" ">
      <subfield code="a">0</subfield>
      <subfield code="b">cbc</subfield>
      <subfield code="c">premunv</subfield>
      <subfield code="d">u</subfield>
      <subfield code="e">ncip</subfield>
      <subfield code="f">19</subfield>
      <subfield code="g">y-gencatlg</subfield>
    </datafield>
  - <datafield tag="010" ind1=" " ind2=" ">
      <subfield code="a"> 79459272 </subfield>
    </datafield>
  - <datafield tag="040" ind1=" " ind2=" ">
      <subfield code="a">DLC</subfield>
      <subfield code="c">CarP</subfield>
      <subfield code="d">DLC</subfield>
    </datafield>
```

http://lccn.loc.gov/79459272/dc

```
- <record>
    <leader>00693cam a2200217u 4500</leader>
    <controlfield tag="001">10462695</controlfield>
    <controlfield tag="005">20030718130135.0</controlfield>
    <controlfield tag="008">841013r19681788be 000 0 ger </controlfield>
  - <datafield tag="035" ind1=" " ind2=" ">
      <subfield code="9">(DLC) 79459272</subfield>
    </datafield>
  - <datafield tag="906" ind1=" " ind2=" ">
      <subfield code="a">0</subfield>
      <subfield code="b">cbc</subfield>
      <subfield code="c">premunv</subfield>
      <subfield code="d">u</subfield>
      <subfield code="e">ncip</subfield>
      <subfield code="f">19</subfield>
      <subfield code="g">y-gencatlg</subfield>
    </datafield>
  - <datafield tag="010" ind1=" " ind2=" ">
      <subfield code="a"> 79459272 </subfield>
    </datafield>
  - <datafield tag="040" ind1=" " ind2=" ">
      <subfield code="a">DLC</subfield>
      <subfield code="c">CarP</subfield>
      <subfield code="d">DLC</subfield>
    </datafield>
  - <datafield tag="050" ind1="0" ind2="0">
      <subfield code="a">B2799.S7</subfield>
      <subfield code="b">W4 1968</subfield>
    </datafield>
  - <datafield tag="100" ind1="1" ind2=" ">
      <subfield code="a">Weishaupt, Adam,</subfield>
      <subfield code="d">1747-1830.</subfield>
    </datafield>
  - <datafield tag="245" ind1="1" ind2="0">
    - <subfield code="a">
        Zweifel über die Kantischen Begriffe von Zeit und Raum.
      </subfield>
    </datafield>
  - <datafield tag="250" ind1=" " ind2=" ">
```

```
- <srw_dc:dc xsi:schemaLocation="info:srw/schema/1/dc-schema
  http://www.loc.gov/standards/sru/resources/dc-schema.xsd">
  - <title>
      Zweifel über die Kantischen Begriffe von Zeit und Raum.
    </title>
    <creator>Weishaupt, Adam, 1747-1830.</creator>
    <type>text</type>
    <publisher>[Bruxelles, Culture et Civilisation,</publisher>
    <date>1968]</date>
    <language>ger</language>
    <subject>Kant, Immanuel, 1724-1804.</subject>
    <subject>Space and time.</subject>
  </srw_dc:dc>
```

- Benannt nach Dublin, Ohio, wo 1995 das erste Treffen der Gruppe / Initiative veranstaltet wurde.
- Dublin Core (DC): Satz von vordefinierten Metadatenelementen, intendiert für
  - Nutzung durch Nicht-Spezialisten
  - die Beschreibung digitaler Ressourcen (i.e. Websites), die häufig keinen eigenen MARC Katalog-Eintrag erhalten würden
- Verglichen mit MARC: Sehr einfach
- Designziel: Allgemeinheit, Einfachheit

Dublin Core® Metadata Initiative
Making it easier to find information.

Satz von 15 Elementen zur Beschreibung von Ressourcen:

- Title
- Creator
- Subject
- Description
- Publisher
- Contributor
- Date
- Type
- Format
- Identifier
- Source
- Language
- Relation
- Coverage
- Rights

```
- <dc>
    <title>Arithmetic / </title>
    <creator> Sandburg, Carl, 1878-1967. </creator>
    <creator> Rand, Ted, ill. </creator>
    <type/>
    <publisher>San Diego :Harcourt Brace Jovanovich,</publisher>
    <date>c1993.</date>
    <language>eng</language>
  - <description>
      A poem about numbers and their characteristics. Features anamorphic,
      or distorted, drawings which can be restored to normal by viewing
      from a particular angle or by viewing the image's reflection in the
      provided Mylar cone.
    </description>
    <description>One Mylar sheet included in pocket.</description>
    <subject>Arithmetic</subject>
    <subject>Children's poetry, American.</subject>
    <subject>Arithmetic</subject>
    <subject>American poetry.</subject>
    <subject>Visual perception.</subject>
  </dc>
```

→ Alle Elemente sind optional und wiederholbar, die Reihenfolge ist beliebig

Fokus: Unstrukturierte und schwach strukturierte Texte

Buzzwords: Text Mining, Data Mining, Information Retrieval, Machinelles Lernen, Textklassifikation, Web Mining

- **Data Mining**: Einsatz auf stark strukturierten Daten

- **Text Mining**: Informationsextraktion aus (u.a. semistrukturierten) Texten; Verwendung von Verfahren / Algorithmen des Data Minings → Automatisierte Strukturierung von Texten (insbes. sehr großen Mengen von Texten)

- **Information Retrieval**: Suchanfragen an einen Textcorpus → Wie finde ich die von mir gesuchte Information?

Die Sache mit der Bedeutung…


Bild
Sport
Autn
am Sonn
Bild
der
Fra
EXPRESS
0,70 €
FC bestätigt nun offiziell:
Podolski geht zu Arsenal
Druck auf die Ukraine
UEFA droht mit Verlegung der EM 2012
Mangelnde Koordination
Köln eine einzige Baustelle
Mann (36) putzte sein Klo: Jetzt liegt er im Krankenhaus
STAR WARS
TOPPS

Quelle: http://www.wzl.rwth-aachen.de/de/ebecb2e7d199a686c125736f00454c10/11_v_deu.pdf

Termfrequenz $tf_{i,j}$: Wie häufig findet sich die Wortform / der Term i im Dokument j?

Beispiel-Dokument *dok1*; betrachtete Wortform: *der*

„Es gibt zwei Hauptgründe dafür, dass der akademische Grad für den Beweis der Kompetenz langsam an Bedeutung verliert, während früher die meisten Berufsprogrammierer Universitätsabschlüsse in Informatik, Mathematik oder ähnlichen Disziplinen vorzuweisen hatten. Zum einen ist es durch den Mangel an Bewerbern gerade für kleine und mittelständische Softwareunternehmen, die nicht wie die deutschen Marktführer Microsoft oder SAP über einen internationalen Ruf verfügen, nicht mehr möglich, ihren Bedarf ausschließlich durch Uniabsolventen zu decken - das zeigen 43.000 offene Stellen in der IT. Zum anderen sind gerade in der sich schnell verändernden Webprogrammierung praktische Fertigkeiten mehr vonnöten als Theorie - Universitäten können mit solch einer Aktualität im Lehrstoff nicht mehr mithalten. Per Fragemann leitet das Berliner Startup Small Improvements. In den Stellenanzeigen des kleinen Unternehmens steht ausdrücklich, dass keine Lebensläufe oder ausgefeilte Anschreiben gewünscht sind. "Es kommt nicht auf den Titel an. Wichtiger ist: Der Bewerber kann coden und er kann es auch zeigen." Ein Github-Repository, die Beteiligung an Open-Source-Projekten oder das Spiel, das jemand in der Freizeit programmiert hat, zählen weit mehr als die Bestnote in der Klausur über theoretische Informatik.“ (www.golem.de/news/programmieren-programming-motherfucker-do-you-speak-it-1405-106106-3.html)

$$tf_{der,dok1} = ?$$

Termfrequenz $tf_{i,j}$: Wie häufig findet sich die Wortform / der Term i im Dokument j?

Beispiel-Dokument *dok1*; betrachtete Wortform: *der*

„Es gibt zwei Hauptgründe dafür, dass **der** akademische Grad für den Beweis **der** Kompetenz langsam an Bedeutung verliert, während früher die meisten Berufsprogrammierer Universitätsabschlüsse in Informatik, Mathematik oder ähnlichen Disziplinen vorzuweisen hatten. Zum einen ist es durch den Mangel an Bewerbern gerade für kleine und mittelständische Softwareunternehmen, die nicht wie die deutschen Marktführer Microsoft oder SAP über einen internationalen Ruf verfügen, nicht mehr möglich, ihren Bedarf ausschließlich durch Uniabsolventen zu decken - das zeigen 43.000 offene Stellen in **der** IT. Zum anderen sind gerade in der sich schnell verändernden Webprogrammierung praktische Fertigkeiten mehr vonnöten als Theorie - Universitäten können mit solch einer Aktualität im Lehrstoff nicht mehr mithalten. Per Fragemann leitet das Berliner Startup Small Improvements. In den Stellenanzeigen des kleinen Unternehmens steht ausdrücklich, dass keine Lebensläufe oder ausgefeilte Anschreiben gewünscht sind. "Es kommt nicht auf den Titel an. Wichtiger ist: **Der** Bewerber kann coden und er kann es auch zeigen." Ein Github-Repository, die Beteiligung an Open-Source-Projekten oder das Spiel, das jemand in **der** Freizeit programmiert hat, zählen weit mehr als die Bestnote in **der** Klausur über theoretische Informatik.“ (www.golem.de/news/programmieren-programming-motherfucker-do-you-speak-it-1405-106106-3.html)

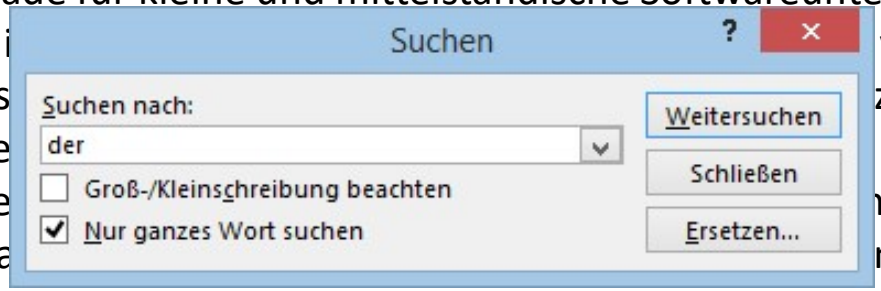
$$tf_{de\ ,dok1} = 6$$

Tf-idf-Maß

Termfrequenz $tf_{i,j}$: Wie häufig findet sich die Wortform / der Term i im Dokument j?

Beispiel-Dokument *dok1*; betrachtete Wortform: *der*

„Es gibt zwei Hauptgründe dafür, dass **der** akademische Grad für den Beweis **der** Kompetenz langsam an Bedeutung verliert, während früher die meisten Berufsprogrammierer Universitätsabschlüsse in Informatik, Mathematik oder ähnlichen Disziplinen vorzuweisen hatten. Zum einen ist es durch den Mangel an Bewerbern gerade für kleine und mittelständische Softwareunternehmen, die nicht wie die deutschen Marktführer Mi



verfügen, nicht mehr möglich, ihren Bedarf auss zeigen 43.000 offene Stellen in **der** IT. Zum ande Webprogrammierung praktische Fertigkeiten me mit solch einer Aktualität im Lehrstoff nicht mehr mitha mall Improvements. In den Stellenanzeigen des kleinen Unternehmens steht ausdrücklich, dass keine Lebensläufe oder ausgefeilte Anschreiben gewünscht sind. "Es kommt nicht auf den Titel an. Wichtiger ist: **Der** Bewerber kann coden und er kann es auch zeigen." Ein Github-Repository, die Beteiligung an Open-Source-Projekten oder das Spiel, das jemand in **der** Freizeit programmiert hat, zählen weit mehr als die Bestnote in **der** Klausur über theoretische Informatik.“ (www.golem.de/news/programmieren-programming-motherfucker-do-you-speak-it-1405-106106-3.html)

$$tf_{de\ ,dok1} = 67$$

Tf-idf-Maß

Inverse Dokumentfrequenz idf_i: Wie häufig findet sich die Wortform / der Term i im Gesamtkorpus?

Annahme: Eine Wortform, die nur in wenigen Titelaufnahmen des Gesamtbestandes anzutreffen ist, verfügt über eine höhere Trennschärfe als eine Wortform, die sich in zahlreichen Titelaufnahmen findet.

$$idf_i = \log\left(\frac{\sum Titel\ in\ Korpus}{\sum Treffer\ Suchterm\ in\ Korpus}\right)$$

Termgewichtung $w_{i,j}$:

$$w_{i,j} = tf_{i,j} \times idf_i = tf_{i,j} \times \log\left(\frac{\sum Titel\ in\ Korpus}{\sum Treffer\ Suchterm\ in\ Korpus}\right)$$

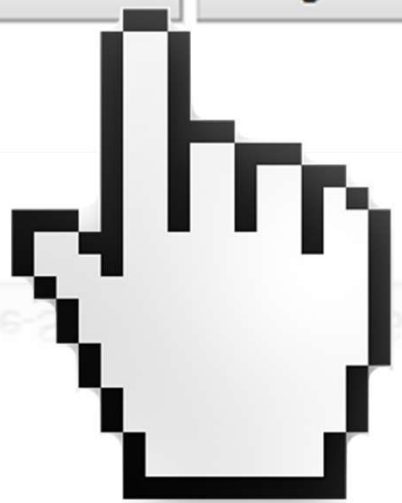
Suche: Inhalte auffinden

Google
Deutschland

Wacky Waving Inflatable Arm Flaling Tube Man

Erweiterte Suche
Sprachtools

Google-Suche | Auf gut Glück!



Google Deutschland

Wacky Waving Inflatable Arm Flaling Tube Man

Erweiterte Suche
Sprachtools

Google-Suche | Auf gut Glück!

Google

Wacky Waving Inflatable Arm Flaling Tube Man | Suche

Ungefähr 87.500 Ergebnisse (0,08 Sekunden) — Erweiterte Suche

Alles
Bilder
Videos
News
Shopping
Mehr

Köln
Standort ändern

Das W
Seiten a
Seit
ersetzte Seiten

Alle
Neueste
Letzte 24 Std.
Letzte Woche
Letzter Monat
Letztes Jahr
Zeitraum festlegen...

Alle Ergebnisse
Websites mit Bildern

Mehr Optionen

Ergebnisse für Wacky Waving Inflatable Arm ***Flailing*** Tube Man.
Stattdessen suchen nach: Wacky Waving Inflatable Arm ***Flaling*** Tube Man

Videos zu **Wacky Waving Inflatable Arm Flailing Tube Man** - Videos melden

WACKY WAVING INFLATABLE ARM FLAILING TUBE MEN!
43 Sek. - 29. März 2010
Hochgeladen von shortendnoodle
youtube.com

wacky waving inflatable arm flailing tube man
43 Sek. - 11. Aug. 2010
ebaumsworld.com

Wacky Waving Inflatable Arm Flaling Tube Man ...
2 Min. - 12. Aug. 2006
Hochgeladen von SUPERLEGEND87
youtube.com

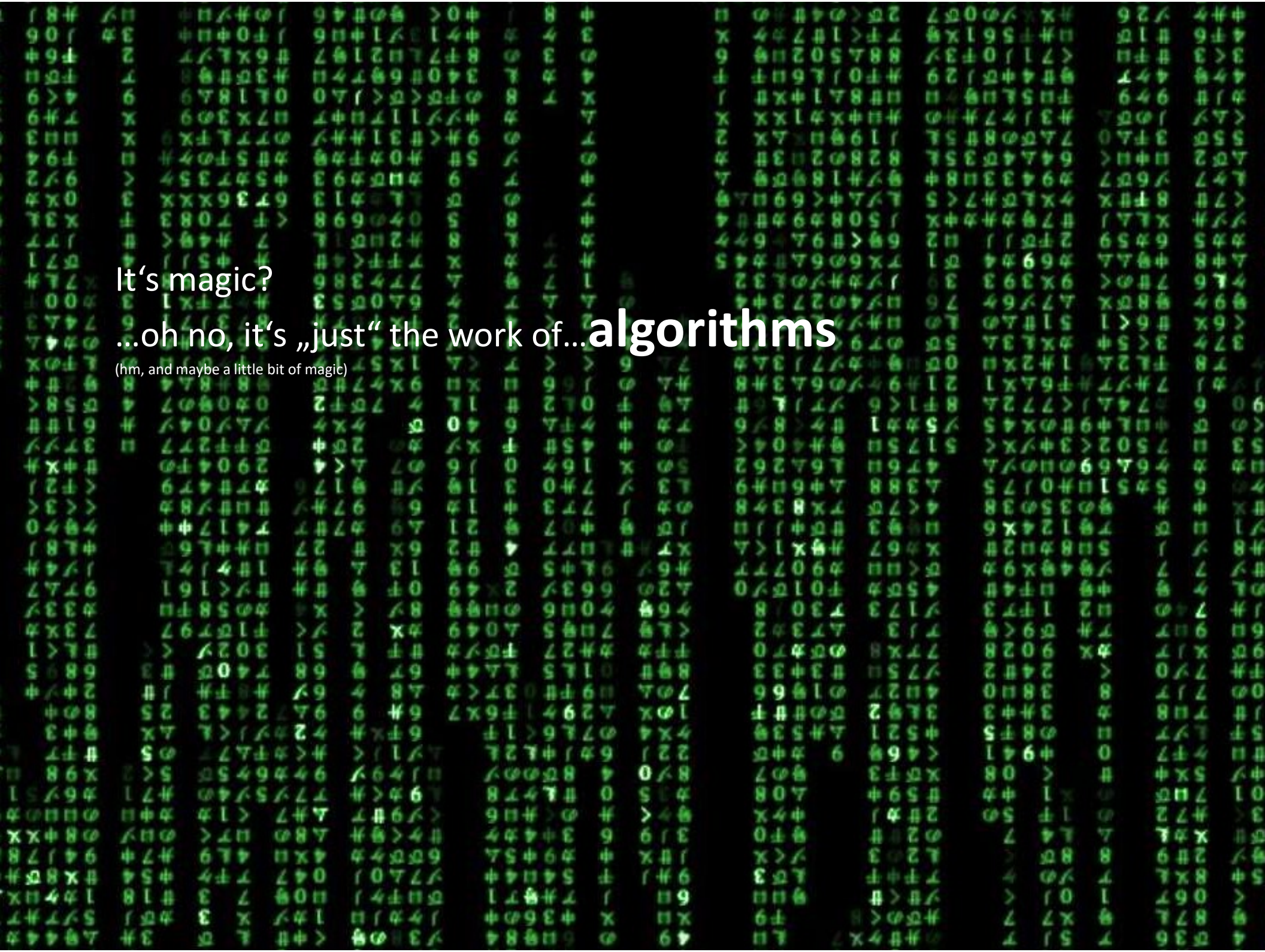
Wackywavinginflatablearmflailingtubeman.com ... - [Diese Seite übersetzen]
Wackywavinginflatablearmflailingtubeman.com offers **Waving Arm Inflatable**, Electron **Tube**, Under **Arm** Wax, and **Flailing Tube**.
www.**wackywavinginflatablearmflailingtubeman**.com/ - Im Cache - Ähnliche Seiten

Bilder zu **Wacky Waving Inflatable Arm Flailing Tube Man** - Bilder melden

Wacky Waving Inflatable Arm-Flailing Tubeman - Family Guy Wiki - [Diese Seite übersetzen]
Wacky Waving Inflatable Arm-Flailing Tubemen are products sold by Al Harrington at Al Harrington's **Wacky Waving Inflatable Arm-Flailing Tubeman** Emporium and **...**
familyguy.wikia.com/.../**Wacky_Waving_Inflatable_Arm-Flailing_Tubeman** -
Im Cache - Ähnliche Seiten

YTMND - **WACKY WAVING INFLATABLE TUBE MAN** - [Diese Seite übersetzen]
BACK TO YTMND. AUTHOR: - SITE PROFILE COMMENTS. SCORE: FOREGROUND: BACKGROUND: SOUND: DESCRIPTION: KEYWORDS: Login or register to hide ads.
wwitm.ytmnd.com/ - Im Cache - Ähnliche Seiten

YTMND - **Whacky Waving Inflatable Arm Flailing Tube Man** - [Diese Seite übersetzen]
BACK TO YTMND. AUTHOR: - SITE PROFILE COMMENTS. SCORE: FOREGROUND **...**
tubeman.ytmnd.com/ - Im Cache - Ähnliche Seiten


It‘s magic?
...oh no, it‘s „just“ the work of...algorithms
(hm, and maybe a little bit of magic)

Ein Praxisbeispiel: Das VD18 Projekt

Neuer Zeitungen von Gelehrten Sachen auf das Jahr ... Theil /
1742 : Neuer Zeitungen von Gelehrten Sachen auf das Jahr MDCCXLII Erster Theil

1742

Neuer Zeitungen von Gelehrten Sachen auf das Jahr ... Theil /
1741 : Neuer Zeitungen von Gelehrten Sachen auf das Jahr MDCCXLI Erster Theil

1741

Observationes Miscellaneae, oder Vermischte Gedancken über allerhand Theologische, Politische, Historische, auch andere zur Antiquität und Ausführung der [...] /
1 : Observationes Miscellaneae, oder Vermischte Gedancken über allerhand Theologische, Politische, Historische, auch andere zur Antiquität und Ausführung der Historie der Gelehrsamkeit dienende curieuse Materien

1713

Hessisches Heb-Opfer Theologischer und Philologischer Anmerckungen /
3 : Hessisches Heb-Opfer Theologischer und Philologischer Anmerckungen

1739

Neu-vermehrtes Gesang- Gebet- Und Com[m]union-Buch

Darinnen nebst denen vom Hrn. Luthero sel. schon Anno 1596. alhier zu Magdeburg heraus gegebenen Evangelischen Liedern, auch anderer gottseligen Männer geistreiche Gesänge enthalten.

Magdeburg : Müller und Faber, 1735

Magdeburgisch Gebet- und Communion-Büchlein

VD18a

Intention: Digitalisierung und Erschließung der im deutschen Sprachraum veröffentlichten Drucke des 18. Jahrhunderts

Kontext

VD 16	VD 17	VD 18
~100 000 erfasste Titel	~255 000 Titel	Ziel: Mehr als 600 000 Titel
Förderzeitraum: 1969-1999	Förderzeitraum: Seit Juli 1996	Förderzeitraum: Ab 2009

Ein Praxisbeispiel: Das VD18 Projekt

- Förderzeitraum Pilotphase: 2009-2011
- Aufgaben Bibliotheken:
  - Digitalisierung → http://digitale.bibliothek.uni-halle.de/vd18
- Aufgaben IDH (ehem. HKI), Köln:
  In der sehr großen Datenbank (*kleio*) mit mehr als ~1 Million Titeln:
  - Einzigartige, im Fundus nur einmal vorhandene, Werke identifizieren
  - Sets von **gleichen** Werken ausfindigmachen

Darinnen nebst denen vom Hrn. Luthero sel. schon Anno 1596. alhier zu Magdeburg heraus gegebenen Evangelischen Liedern, auch anderer gottseligen Männer geistreiche Gesänge enthalten.

Magdeburg : Müller und Faber, 1735

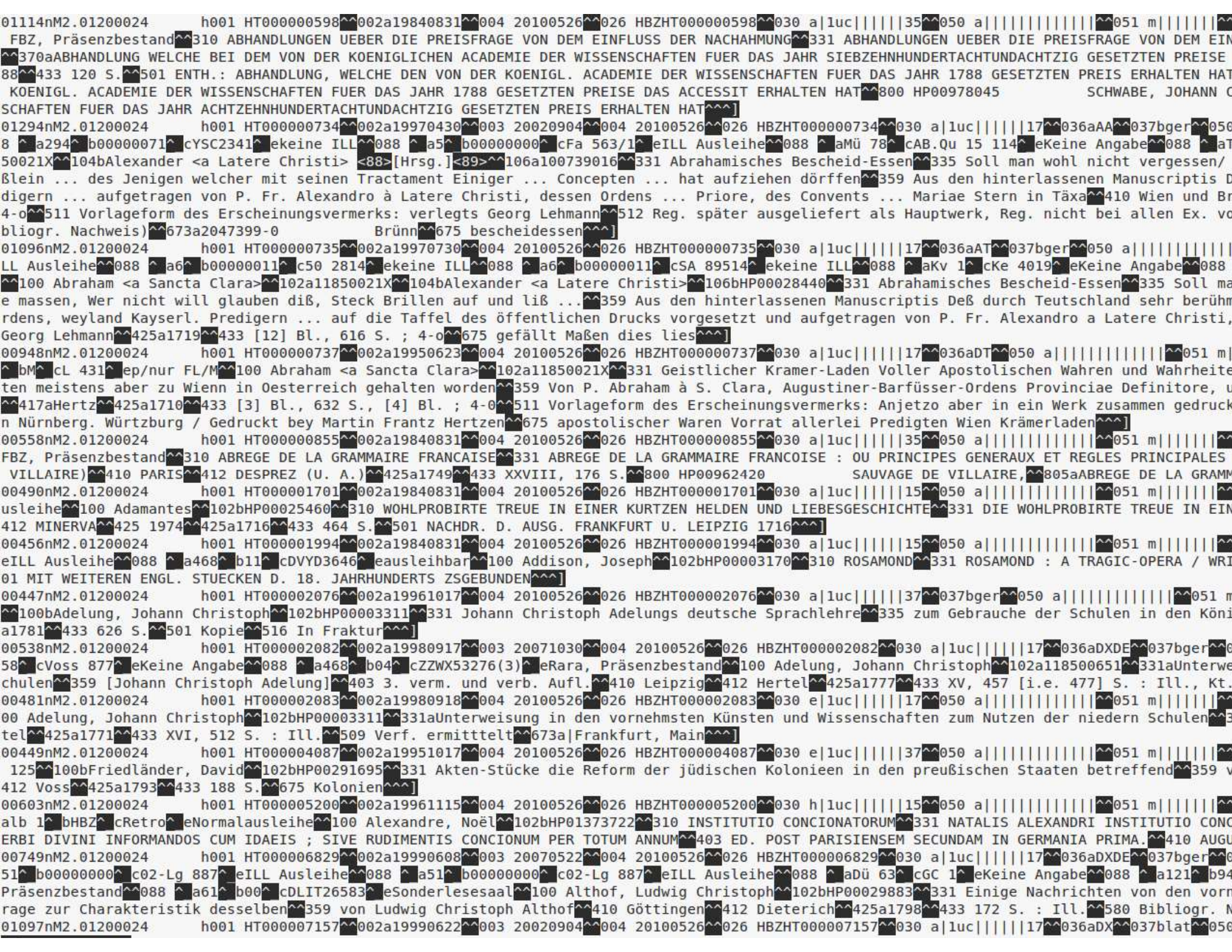
Magdeburgisch Gebet- und Communion-Büchlein


hbz
Wissen. Information. Innovation.


GBV


HKI


BVB BibliotheksVerbund
Bayern



Neuer Zeitungen von Gelehrten Sachen auf das Jahr ... Theil /
1742 : Neuer Zeitungen von Gelehrten Sachen auf das Jahr MDCCXLII Erster Theil

1742

Neuer Zeitungen von Gelehrten Sachen auf das Jahr ... Theil /
1741 : Neuer Zeitungen von Gelehrten Sachen auf das Jahr MDCCXLI Erster Theil

1741

Observationes Miscellaneae, oder Vermischte Gedancken über allerhand Theologische, Politische, Historische, auch andere zur Antiquität und Ausführung der [...] /
1 : Observationes Miscellaneae, oder Vermischte Gedancken über allerhand Theologische, Politische, Historische, auch andere zur Antiquität und Ausführung der Historie der Gelehrsamkeit dienende curieuse Materien

1713

Hessisches Heb-Opfer Theologischer und Philologischer Anmerckungen /
3 : Hessisches Heb-Opfer Theologischer und Philologischer Anmerckungen

1739

Neu-vermehrtes Gesang- Gebet- Und Com[m]union-Buch

Darinnen nebst denen vom Hrn. Luthero sel. schon Anno 1596. alhier zu Magdeburg heraus gegebenen Evangelischen Liedern, auch anderer gottseligen Männer geistreiche Gesänge enthalten.

Magdeburg : Müller und Faber, 1735

Magdeburgisch Gebet- und Communion-Büchlein

In welchem enthalten Wie sich ein Christ täglich durch Morgen- und Abend-Segen, auch in allerley Noth und Anliegen, durch andächtige Gebeter geistreicher Männer, Gott ergeben , und Zur Beichte und Hochwürdigem Abendmahl würdiglich bereiten soll.

Text Mining Tool:

- Termfrequenz: Häufigkeit des (Such)Terms / der Wortform im jeweiligen Dokument
- Bestimmung der Trennschärfe einer Wortform: Inverse Document Frequency (IDF), Inverse Dokumenthäufigkeit
- Annahme: Eine Wortform, die nur in wenigen Titelaufnahmen des Gesamtbestandes anzutreffen ist, verfügt über eine höhere Trennschärfe als eine Wortform, die sich in zahlreichen Titelaufnahmen findet.

$$IDF = \log\left(\frac{\sum Titel\ in\ DB}{\sum Treffer\ Suchterm\ in\ DB}\right)$$

IDF – Beispiel:
„Griechische Anthologie – aus den besten Dichtern gesammlet, nach den Dichtungsarten geordnet und mit literarischen Notizen begleitet; für Gymnasien und Akademien“

Wortform	Weight (IDF)
griechische	8.087948
anthologie	9.728360
aus	3.555348
den	2.302585
besten	6.173786
dichtern	9.064968
gesammlet	8.407825
nach	3.663562
den	2.302585
dichtungsarten	11.462957
geordnet	9.194211
und	0.693147
mit	3.367296
literarischen	9.971287
notizen	11.714273
begleitet	7.990238
für	3.663562
gymnasien	8.554104
und	0.693147
akademien	8.099251

Gewichtung

(2) Gewichtung der Suchphrase bestimmen, i.e.:
„Griechische Anthologie – aus den besten Dichtern gesammlet, nach den Dichtungsarten geordnet und mit literarischen Notizen begleitet; für

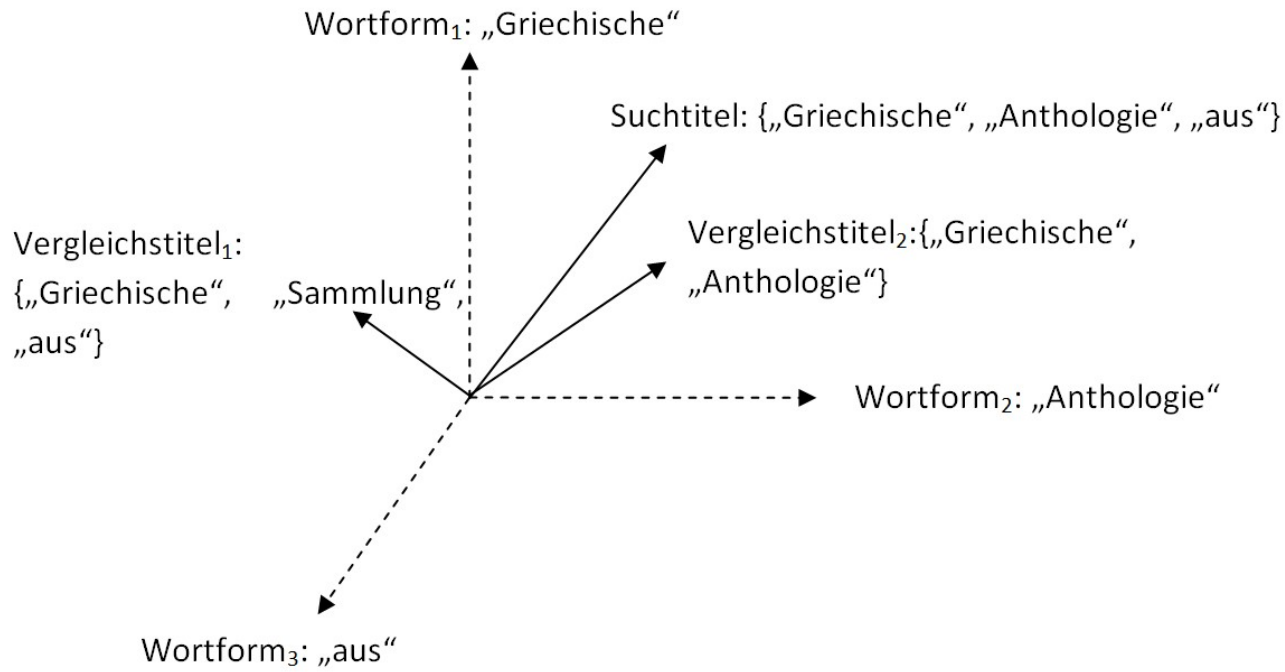
$$sumsearchtermweights = \sum relevantsearchterms$$

→ Summe der Gewichtungen relevanter Suchterme, die die Suchphrase charakterisieren

Vektorraum

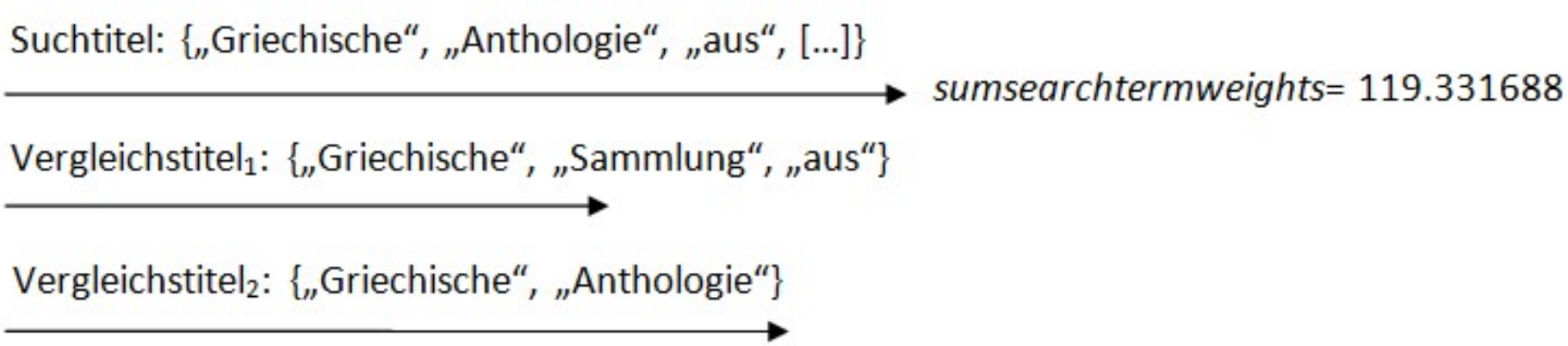
(3) Wie herausfinden, dass der Titel
„Griechische Anthologie – aus den besten Dichtern gesammlet, nach den Dichtungsarten geordnet und mit literarischen Notizen begleitet; für Gymnasien und Akademien“
gleich bzw. sehr ähnlich ist zu dem deutlich kürzeren Titel „Griechische Anthologie“ ?

- Eine Möglichkeit: Abbildung in einem n-dimensionalen Vektorraum



Vektorraum

(4) Komplexität verringern, retrieval erhöhen:



Ähnlichkeitsmaß

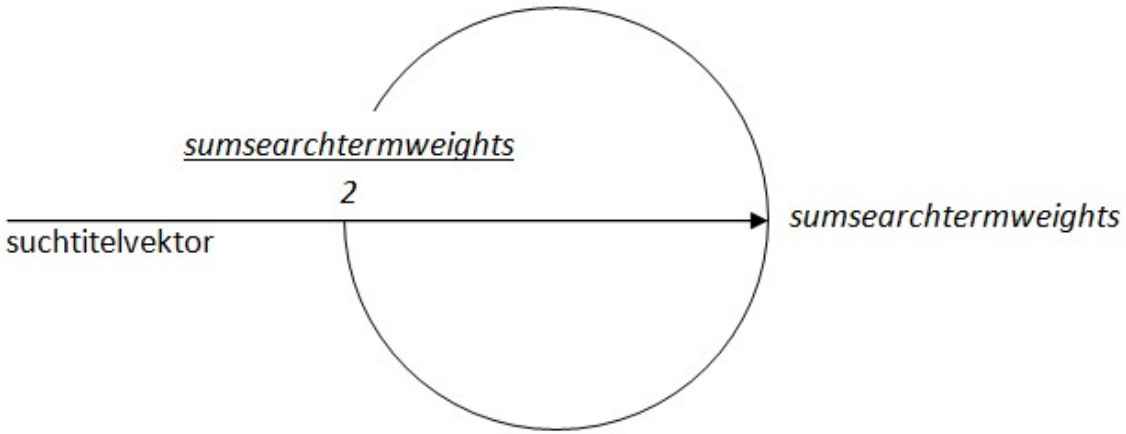
(5) Ähnlichkeit von Such- und Vergleichstitel bzw. der korrespondierenden Vektoren ermitteln: Ähnlichkeitsmaß

$$similiarmeasure = \sqrt{(x-y)^2 + \frac{(y-x)^2}{2}}$$

Hierbei:

- *x:* Summe der Suchtermgewichte
- *y*: Summe der Gewichtungen der im Suchtitel vorhandenen Wortformen des Vergleichstitels

Similiarmeasure: Distanz des Vergleichstitels zum Suchtitel Ein Vergleichstitel wird als potenziell relevant erachtet, wenn sein Abstand zum Suchvektor kleiner ist als $\frac{sumsearchtermweights}{2}$, der Vergleichstitelvektor sich also in räumlicher Nähe zum Suchtitelvektor befindet.



Cluster

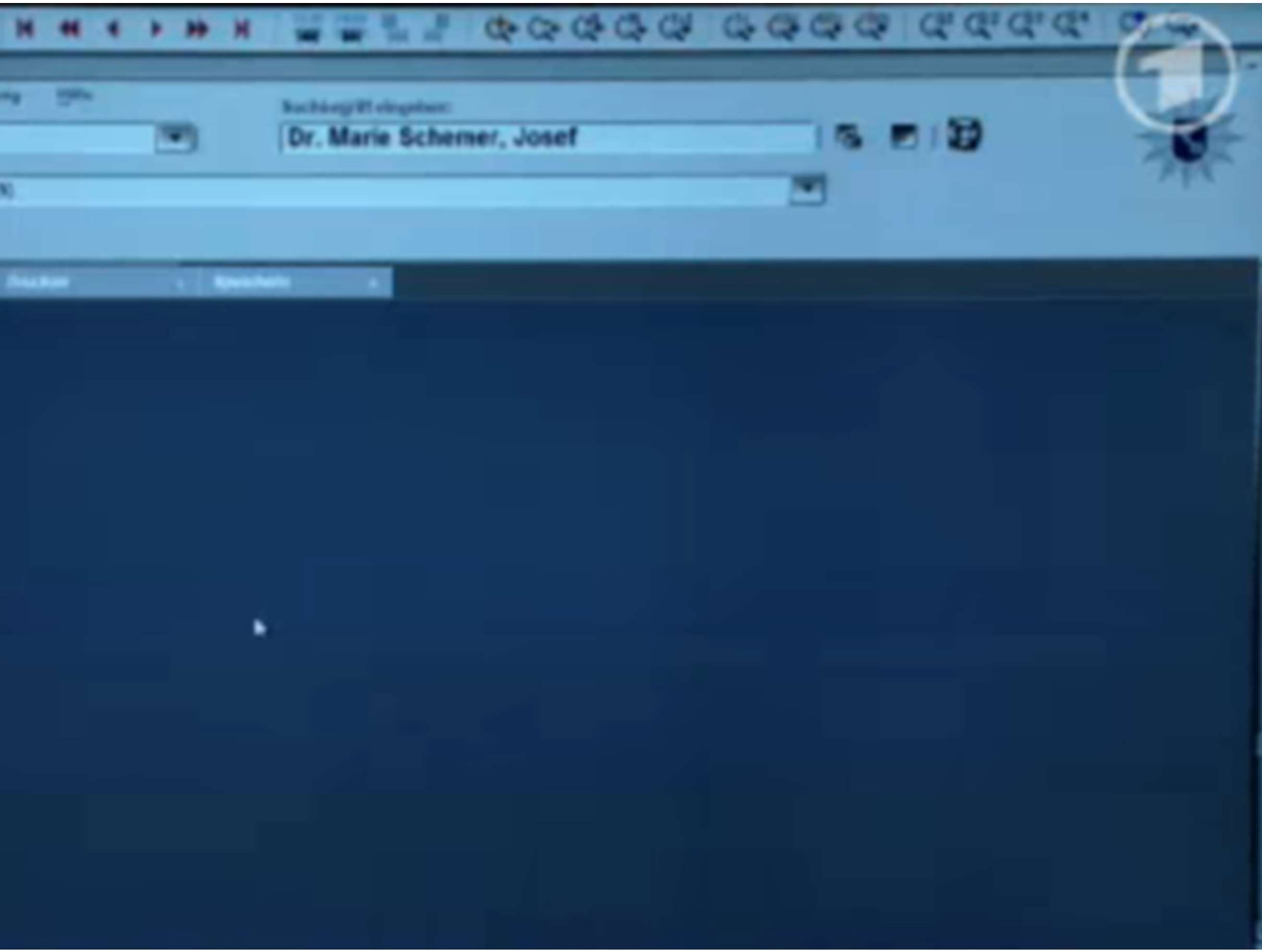
(6) Cluster ähnlicher Titel generieren:

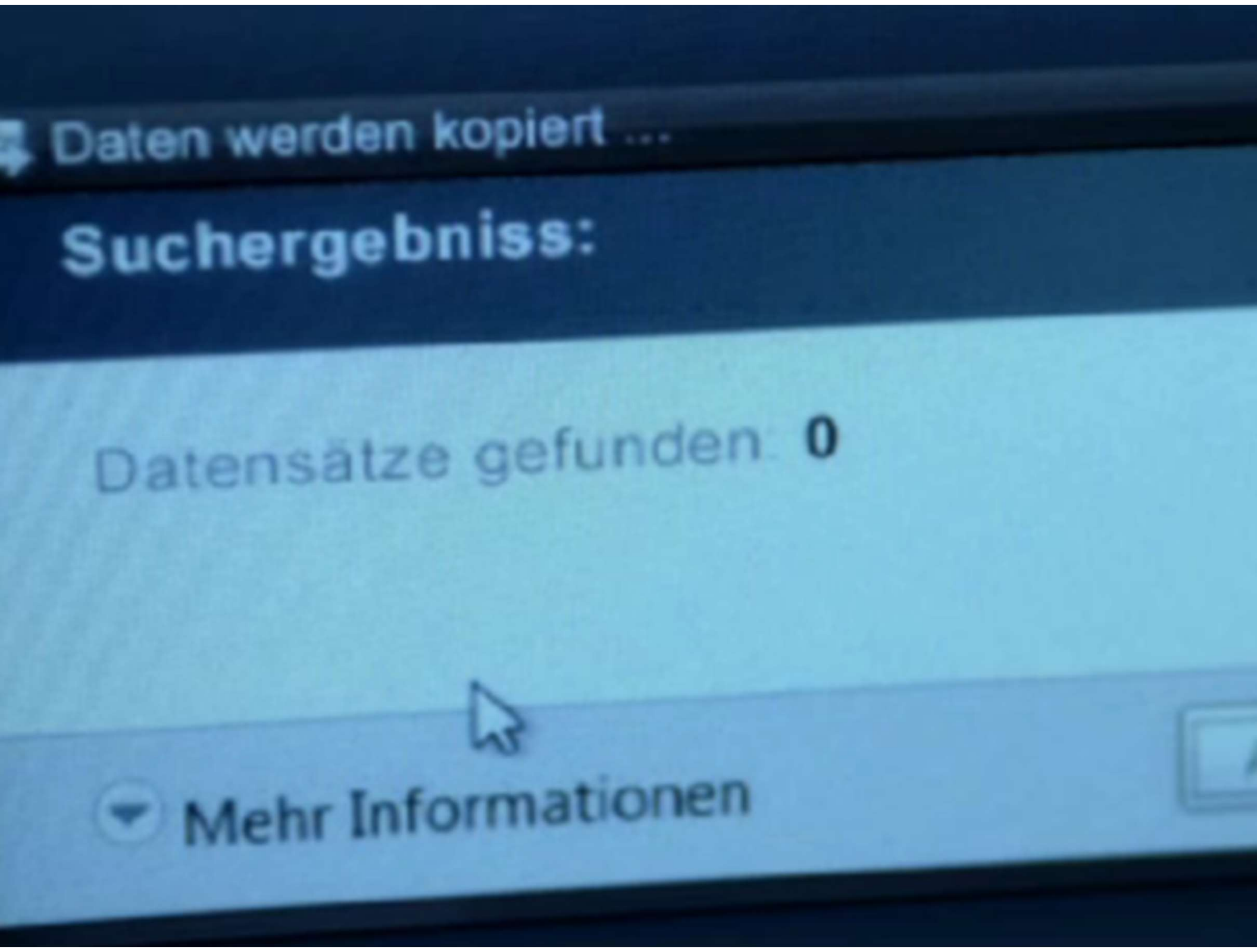
- Cluster I: Titel mit Gewichtung = 22.040516
  - Dissertatio jvridica inavgvralis de jvre consvetvdinario
  - Dissertatio Juridica Inauguralis De Jure Consuetudinario
  - […]
- Cluster II: Titel mit Gewichtung = 14.525173
  - <ns>Diss. iur. inaug.</ns> de iure consuetudinario
  - Dissertatio iuris Germanici de iure consuetudinario universali Germaniae Medii Aevi in speculis Saxonico et Suevico, eiusque cognoscendi ratione

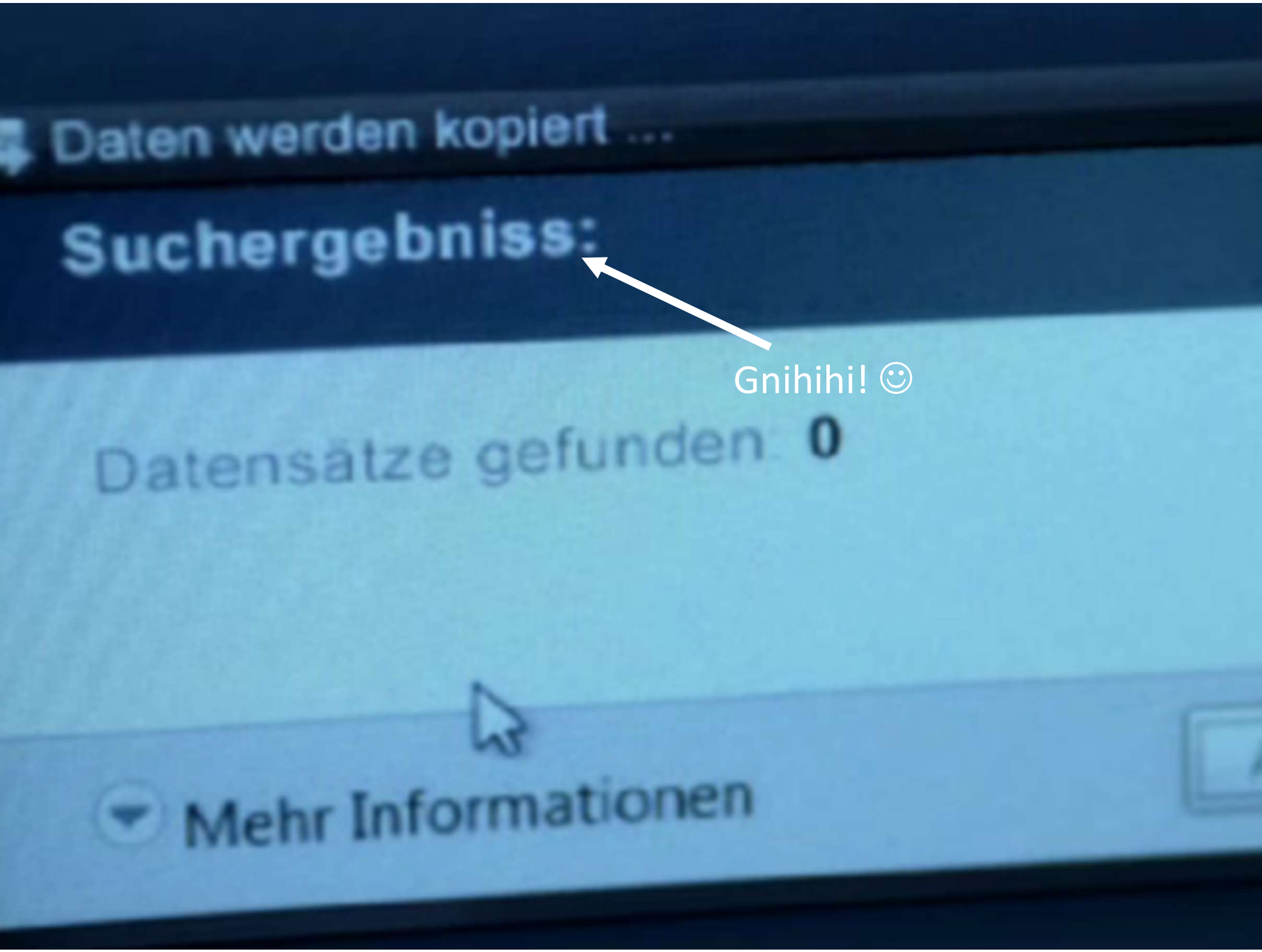
(7) MAB Einträge (Author Name, Place of Printing, etc.) unscharf (fuzzy) vergleichen

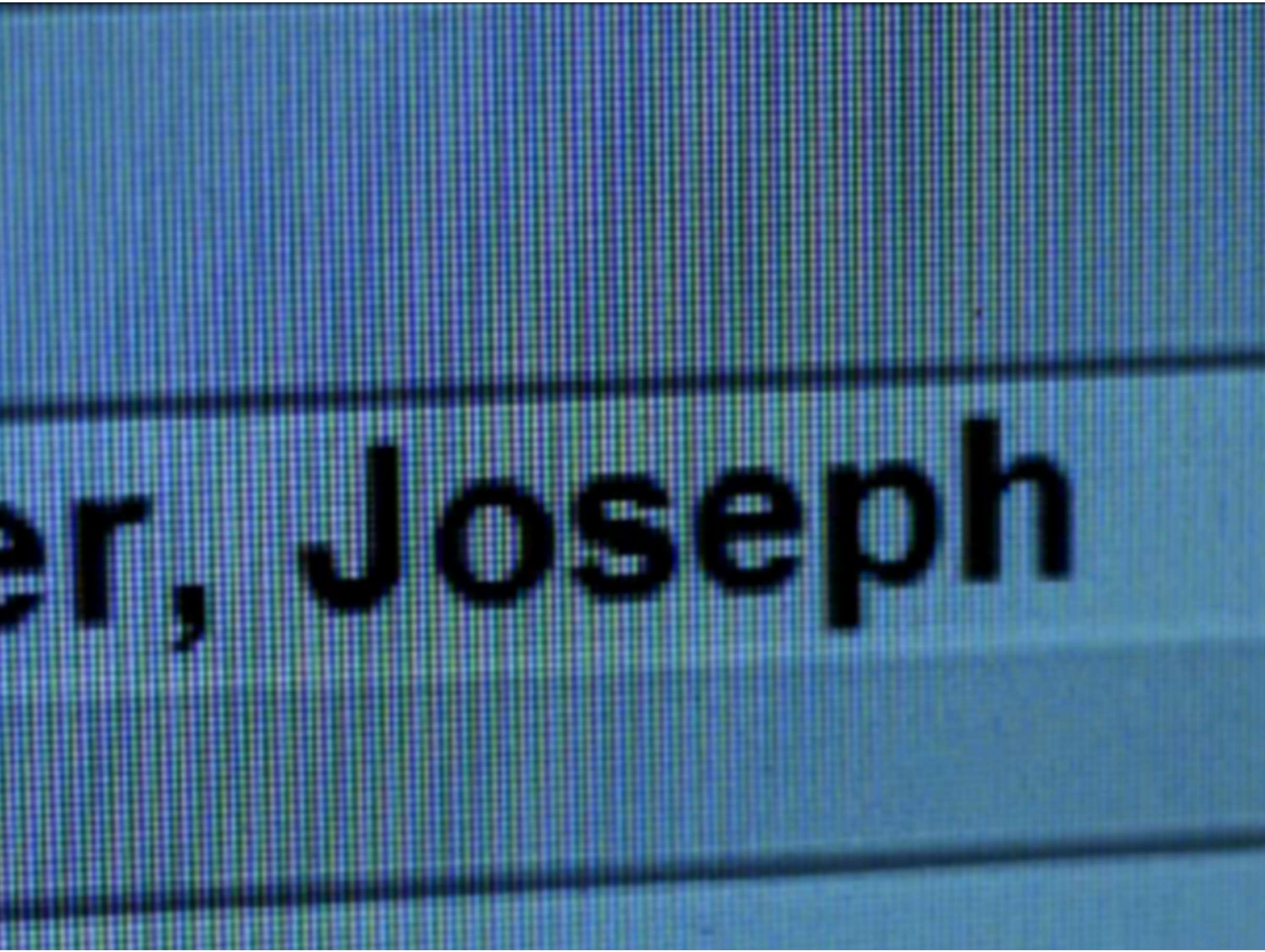
- Partial String Comparison
- Levenshtein Distance / Edit Distance
- …

„Tatort“-Folge „Er wird töten“ (09.06.2013)


Dr. Marie Scherner, Josef


Daten werden kopiert ...
Suchergebniss:
Datensätze gefunden 0
Mehr Informationen


Daten werden kopiert ...
Suchergebniss:
Gnihihi! ☺
Datensätze gefunden 0
Mehr Informationen


er, Joseph

Dr. Marie Schemer, Joseph

Bremen

Kontoname

23/24849199'2

Kartenberechtigung

VORSTRAFEN | LINKS | SONSTIGES

Dr. Marie Schemer
geboren: 14.2. 1974

Ehefrau Dr. Marie Schemer geb.: 14.2. 1974

Joseph Vegener wurde am 27.03.2004 vom Oberlandesgericht Bremen Aktz.: LG 34/7637/2004 zu lebenslanger Haft verurteilt, wg. Mord an seiner Tochter Christine.

Foto: Joseph Vegener

Levenshtein-Distanz

Levenshtein-Distanz, oder auch „Edit-Distance“:

- Geringste Anzahl der Bearbeitungsschritte, um eine Zeichenkette in eine andere Zeichenkette zu transformieren.
  Vorgestellt in Levenshtein, Vladimir I.: Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physics Doklady, Vol. 10, No. 8. (1966), pp. 707-710.

- Beispiel: „kleyer“ vs. „meyer“
  - Levenshtein-Distanz zwischen den beiden Zeichenketten beträgt zwei: Um „kleyer“ in „meyer“ umzuformen, muss das zweite Zeichen der Zeichenkette „kleyer“ gelöscht („k**l**eyer“ → „keyer“) und das erste Zeichen in den Buchstaben „m“ geändert werden („**k**eyer“ → „meyer“).

Bewertung der Suchergebnisse

- Recall: Liefert die Suchanfrage ein relevantes Ergebnis?
- Precision: Ist der gefundene / zurückgelieferte Treffer für die Suchanfrage relevant?


♥Text Mining