



IT Systems Engineering | Universität Potsdam

# Crawling the Web

Felix Elliger & Markus Steiner

# Übersicht

2

- HTML Parser
  - Anforderungsanalyse
  - Probleme
  - Externe Open Source Bibliotheken
  - Jericho HTML Parser
  - Vorstellung unserer Bibliothek
  
- Wikipedia
  - Möglichkeiten um an die Personendaten zu kommen
  - Auswahl der Daten für unsere Wikipedia-Tabelle

# Anforderungsanalyse für den Parser

3

- **Extraktion von Texten aus dem HTML-String**
  - Body-Text
  - Titel-Text
  
- **Umwandlung aller HTML-Sonderzeichen in Unicodezeichen**
  - Zum Beispiel *&auml* → ä
  - Ersetzung aller *&nbsp;* durch gewöhnliche Leerzeichen
  
- **Zuverlässige Ergebnisse auch bei schlechtem HTML Code**
  
- **spezielle Hervorhebungen im Body oder Titel erhalten**
  - Fett geschriebenen Text
  - Kursiv geschriebenen Text

# Probleme beim Parsen

4

- **Sehr schlechter und nicht standardkonformer HTML-Code**

- Endtags fehlen z.B. `<body>foobar`
- Feine Unterschiede z.B. `<body >foobar</body >`

→ reguläre Ausdrücke könnten sehr komplex werden

- **Umwandlung der Sonderzeichen**

- Erster Ansatz:  
manuelle Definition und Umwandlung von Zeichen wie ÄÖÜ

**aber:** was passiert mit **Mario Gómez García** ?

**Folge:** Liste der zu definierenden Sonderzeichen wird sehr lang

# Externe HTML Parser Bibliotheken

5

## ■ Beispiele

- JTidy, NekoHTML, Java HTML Parser, TagSoup
- Weitere unter <http://java-source.net/open-source/html-parsers>

## ■ Vorteile

- Alle Bibliotheken sind Open Source
- Automatische Korrektur von schlechtem HTML-Code
- Definitionen durch RegEx entfallen fast vollständig
- Wesentlich robuster als unsere Definitionen

## ■ Nachteile

- Teilweise sehr schlechte Dokumentation

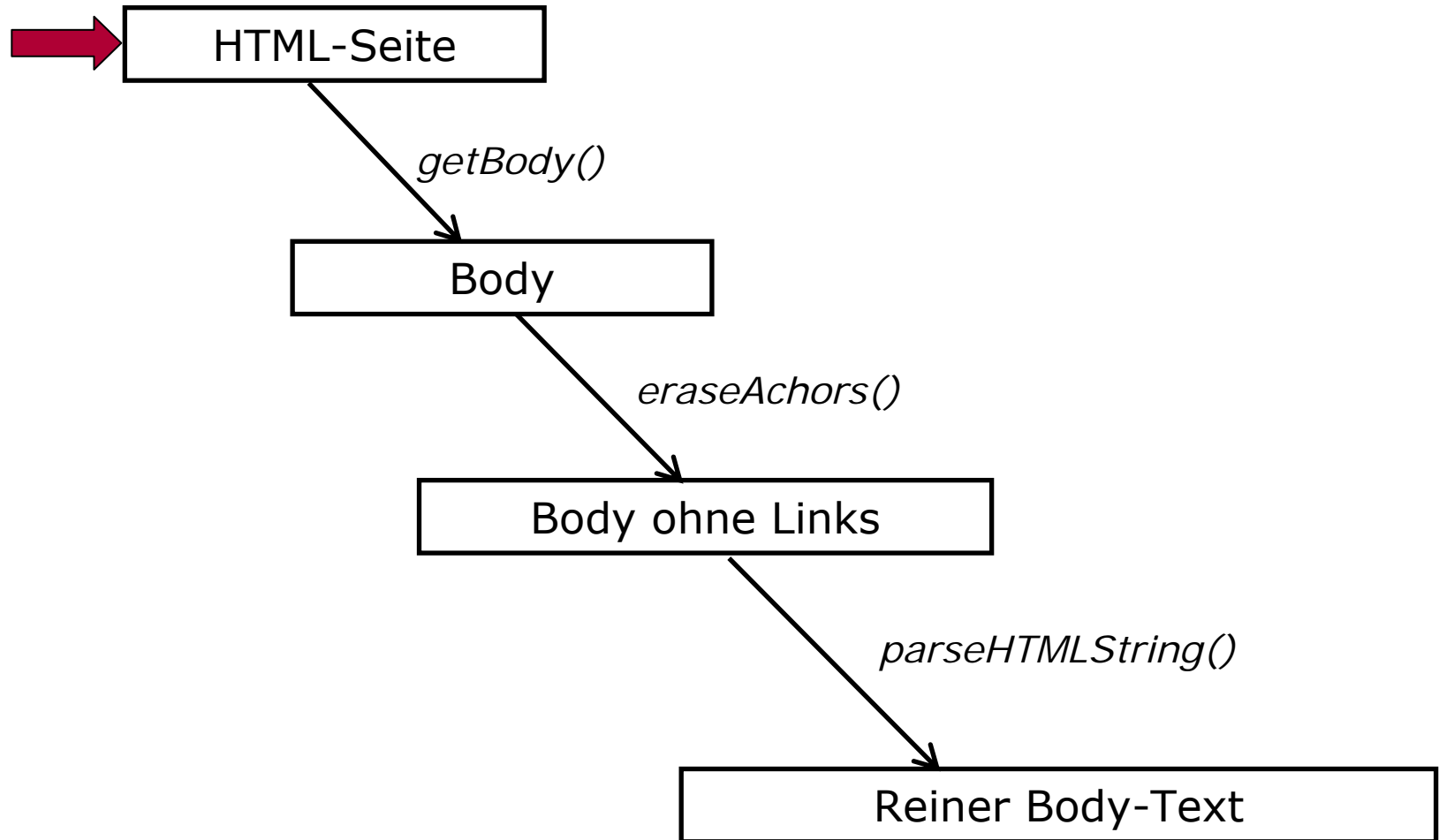
# Jericho HTML Parser 2.5

6

- Open Source Bibliothek unter GNU Lesser General Public License
- HTML-Dokumente parsen, analysieren und manipulieren
- **Vorteile**
  - sehr robust, d.h. auch bei schlechtem HTML-Code gute Ergebnisse
  - keine riesigen Definitionen von Sonderzeichen
  - gute Dokumentation unter
    - <http://jerichohtml.sourceforge.net>
    - <http://jerichohtml.sourceforge.net/doc/api/index.html>
- **Nachteil**
  - INFO Ausgaben bei schlechtem HTML-Code
    - ➔ Fehlerausgabe in Datei

# Textextraktion aus dem Body

7



# getBody()

8

```
public static String getBody(String htmlPage) throws TagNotFoundException{
    String s = null;
    Source src = new Source(htmlPage);
    Element bodyElement = src.findNextElement(0, HTMLElementName.BODY);

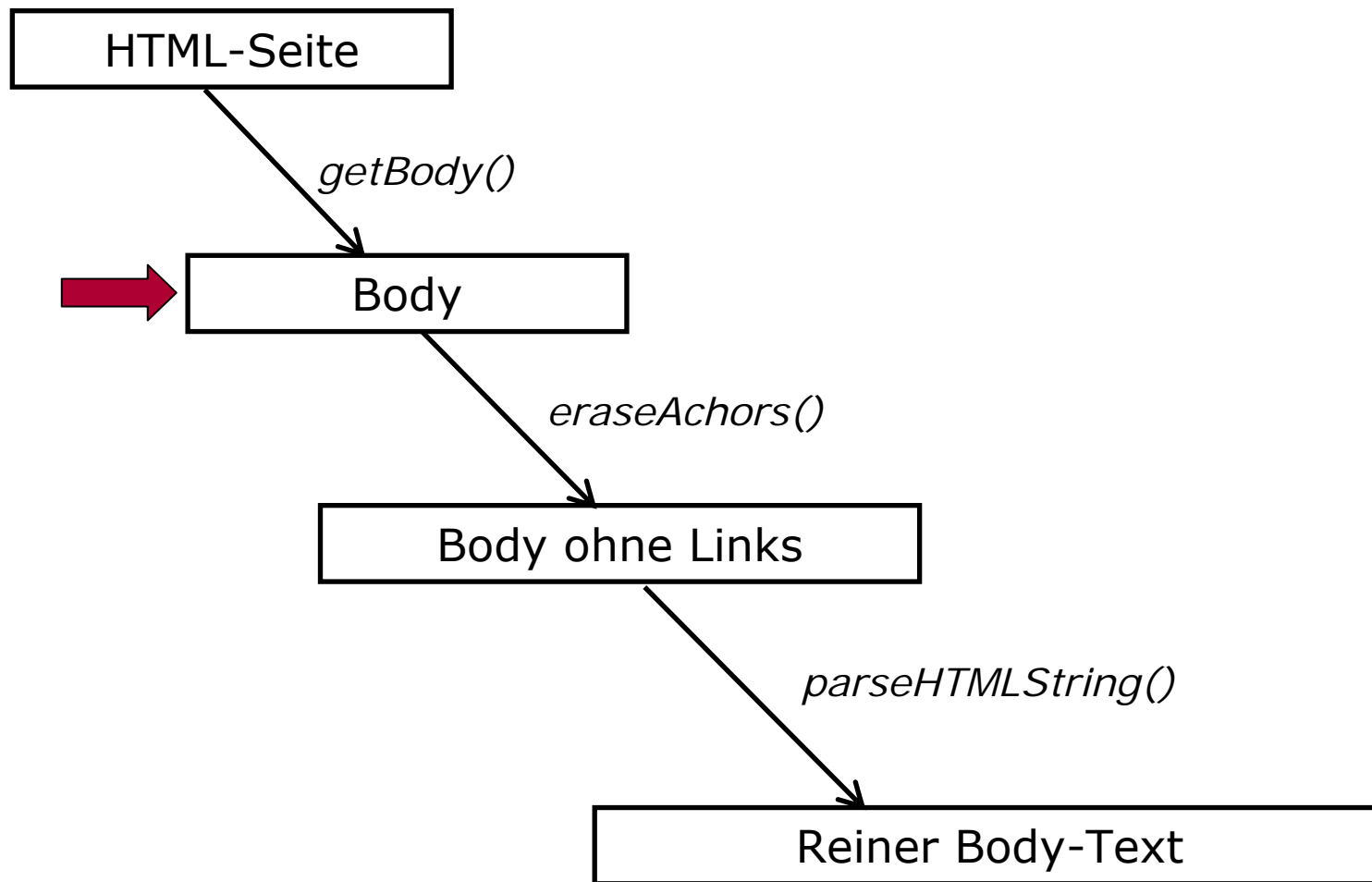
    if (bodyElement != null) {
        s = CharacterReference.decodeCollapseWhiteSpace
            (bodyElement.getContent());
    }
    else {
        throw new TagNotFoundException("No body-tag found");
    }

    return s;
}
```



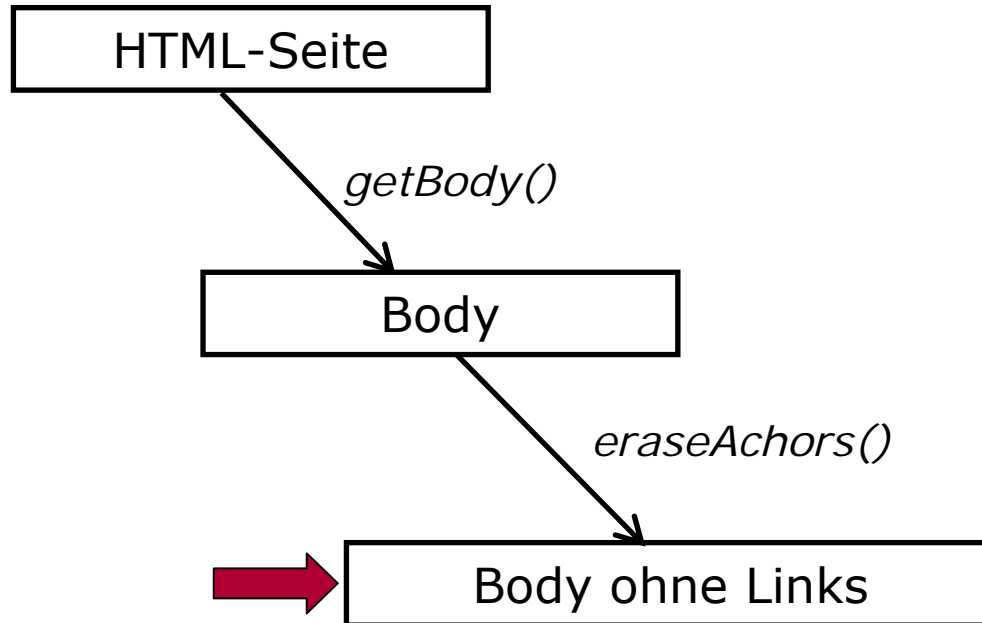
# Textextraktion aus dem Body

9



# parseHTMLString

10



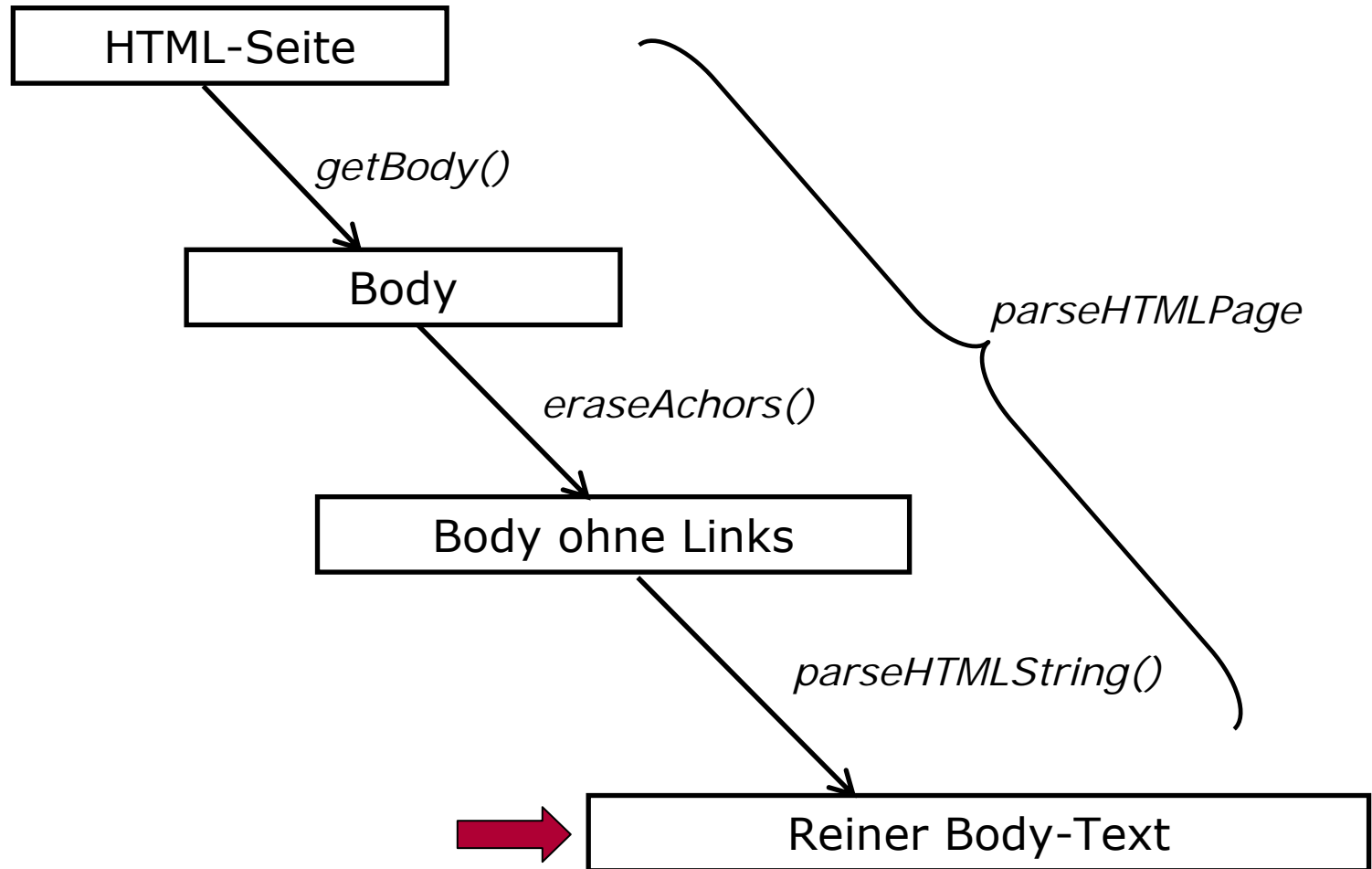
```

public static String parseHTMLString(String htmlString) {
    Source scr = new Source(htmlString);

    return scr.getTextExtractor().toString();
}
  
```

# Textextraktion aus dem Body

11



# Extraktion des Titels – getTitle()



```

public static String getTitle(String htmlPage) throws TagNotFoundException{
    String s = null;
    Source src = new Source(htmlPage);
    Element titleElement = src.findNextElement(0,HTMLElementName.TITLE);

    if (titleElement != null) {
        s = CharacterReference.decodeCollapseWhiteSpace
            (titleElement.getContent());
    }
    else {
        throw new TagNotFoundException("No title-tag found");
    }

    return s;
}
  
```

# Problem bei getTitle()

13

- **Quelleninformationen im Titel**
  - Zum Beispiel „ BUNTE.T-Online Newsline: Caroline Beil ... “
- **Diskussion: Stören uns diese Informationen später???**
- **Mögliche Lösungen**
  1. **Manuelle Schemata für alle Quellen entwickeln**
    - ◇ Sehr aufwändig
    - ◇ Nicht robust
      - ➔ kleine Änderungen haben große Auswirkungen
      - ➔ ständige Wartung nötig
  2. **Vergleichen aller News eines Tages und entfernen gleicher Titelemente**
    - ◇ Automatische Anpassung an Formatänderungen
    - ◇ Sehr robust

# Diskussion weiterer Methoden

14

## ■ **Tokenize**

- zerlegt den String in ein Array einzelner Worte
- Trennung erfolgt an Leerzeichen

## ■ **ModifyPunctuation**

- Entfernt Satzzeichen
- Ersetzt ` durch ``

## ■ **getUpperCaseWords**

- Liefert alle großgeschriebenen Worte

## ■ **getAllTextsIn**

- Liefert alle Textelemente zwischen bestimmten Tags

## **3 Möglichkeiten um an die Daten zu kommen**

- 1. Tabelle**
- 2. CSV**
- 3. Direkter Zugriff auf Toolserver**

# Parsen der Tabelle

16

TT-Table Personendaten - Mozilla Firefox

http://tools.wikimedia.de/~kolossos/templateiger/tt-table4.php?lang=de&template=Personendaten&where=&is=&offset=120&limit=40

```
SELECT DISTINCT `name_id`,`name` FROM `pub_tt1_de` WHERE `tp_name` LIKE 'Personendaten' LIMIT 120,40
```

<<<>> 120 - 160 of 163212

NAME	ALTERNATIVNAMEN	KURZBESCHREIBUNG	GEBURTS DATUM
Enik Hubble	Hubble, Edwin Powell	[[Vereinigte Staaten US-amerikanische]] [[Schauspieler]]in und [[Playmate]]	September]] [[1969]]
Edwin Hubble	Hubble, Edwin Powell	[[Vereinigte Staaten US-amerikanischer]] [[Astronom]]	September]] [[1889]]
Enrico Fermi	Fermi, Enrico	italienischer [[Kernphysik]]er	September]] [[1901]]
Ernst Thälmann	Thälmann, Ernst	[[Hamburg]]er Politiker der Kommunistischen Partei Deutschlands [[KPD]]	1886
Edmund Stoiber	Stoiber, Edmund	deutscher Politiker und seit 1993 [[Ministerpräsident]] des [[Freistaat]]es [[Bayern]]	September]] [[1941]]
Euklid	Euklid von Alexandria	griechischer [[Mathematiker]]	um [[365 v. Chr.]]
Ernesto Che Guevara	Guevara, Che	argentinischer Arzt und kubanischer Revolutionär	[[14. Mai]] [[1928]]
Ernst Jünger	Jünger, Ernst	deutscher [[Schriftsteller]] und [[Publizist]]	[[29. März]] [[1895]]
Esther Friesner	Friesner-Stutzman, Esther Mona (verheiratet)	US-amerikanische Schriftstellerin	16. Juli 1951
Ennio Morricone	Savo, Dan (Pseudonym), Nichols, Leo (Pseudonym)	[[Italien italienischer]] Filmmusik-Komponist und Dirigent	10. November 1928
Emil Nolde	Nolde, Emil	Emil Hansen	17. August 1867
Évariste Galois	Galois, Évariste	Maler des [[Expressionismus]]	
Erasmus Reinhold	Reinhold, Erasmus	[[Frankreich französischer]] [[Mathematiker]]	
Euripides	Euripides	deutscher [[Astronom]] und [[Mathematiker]]	
Erika Fuchs	Fuchs, Erika	[[Theater der griechischen Antike klassischer griechischer Dichter]]	5 v. Chr.[485
Émile Durkheim	Durkheim, Émile	deutsche Übersetzerin	
Emmy Noether	Noether, Emmy	französischer [[Soziologe]]	
Erich Kästner	Kästner, Erich	[[Deutschland deutsche]] [[Mathematiker]]in	
Ernst Röhm	Röhm, Ernst Julius	[[Deutschland deutsche]] [[Schriftsteller]], [[Drehbuchautor]] und [[Kabarettist]]	
Edsger Wybe Dijkstra	Dijkstra, Edsger Wybe	führender [[Nationalsozialist]] und langjähriger Führer der [[Sturmabteilung]]	
Edelgard Bulmahn	Bulmahn, Edelgard	[[Niederlande niederländischer]] [[Informatiker]]	
Eric Hoffer	Hoffer, Eric	[[Deutschland deutsche]] [[Politiker]]in ([[Sozialdemokratische Partei Deutschlands]])	
Fritz Lang	Lang, Fritz	US-amerikanischer [[Sozialkritik sozialkritischer]] [[Philosophie Philosoph]]	
Francis Ford Coppola	Coppola, Francis Ford	österreichisch-US-amerikanischer Filmregisseur, Drehbuchautor und Schauspieler	5. Dezember 1890
Friedrich Wilhelm Murnau	Murnau, Friedrich Wilhelm	[[Vereinigte Staaten US-amerikanischer]] [[Regisseur]] und [[Filmproduzent Produzent]]	[[7. April]] [[1939]]
Frederic Volz	Volz, Frederic	deutscher [[Regisseur]]	[[28. Dezember]] [[1888]]
Fred Zinnemann	Zinnemann, Fred	[[Biochemiker]], Umweltexperte und populärwissenschaftlicher [[Autor]]	23. November 1925
Frank Capra	Capra, Frank		
Friedrich Nietzsche	Nietzsche, Friedrich Wilhelm		
Fidel Castro	Castro, Fidel		
Franz Münterberg	Münterberg, Franz		
Fritz Kuhn	Kuhn, Fritz		
Franz Kafka	Kafka, Franz		
Florianus	Florianus		
Francis William Aston	Aston, Francis William		
Ferdinand de Saussure	Saussure, Ferdinand de		
Friedrich Dürrenmatt	Dürrenmatt, Friedrich		
Fritz Walter	Walter, Fritz		
Franz Werfel	Werfel, Franz		
Farkas Wolfgang Bolyai	Bolyai, Farkas		

Mögliche Spalten

Wikimedia-Titel / URL

Name

Alternativnamen

Kurzbeschreibung

Geburtsdatum

Geburtsort

Sterbedatum

Sterbeort

Problem: &limit=3000

Lösung  
Offsetwert hoch zählen  
&offset=3000

Steuerung der Tabelle über die URL

- &columns= NAME,ALTERNATIVNAME,...
- &order=article
- &format=csv



# Parsen der CSV-Datei

17

article NAME ALTERNATIVNAMEN KURZBESCHREIBUNG GEBURTSDATUM GEBURTSORT STERBEDATUM STERBEORT Lee Ang Lee, Ang [[Regisseur]] und Drehbuchautor [[23. Oktober]]  
 Warhol, Andy Warhola, Andrew Begründer und bedeutendster Künstler der [[Pop-Art]] [[6. August]] [[1928]] [[Pittsburgh]] [[22. Februar]] [[1987]] [[New York City|New York]], [[New York (Bundesstaat)|New York]]  
 [[Regisseur]] [[6. Januar]] [[1954]] [[Isle of Wight]], [[Vereinigtes Königreich Großbritannien]] Alfred Hitchcock Hitchcock, Alfred [[Vereinigtes Königreich britischer]] [[Filmregisseur]] und [[Filmproduzent|Produzent]]

## Probleme

Kooperativ (Schauspieler) [[28. April]] [[1980]] [[Los Angeles]], [[Kalifornien]], [[Vereinigte Staaten|USA]] Aki Kaurismäki Kaurismäki, Aki [[Finland|finnischer]] [[Regisseur]] [[4. April]] [[1957]] [[Orimattila]]  
 Al amerikanischer [[Schauspieler]], [[Regisseur]] und [[Filmproduzent|Produzent]] 25. April 1940 [[New York City|New York]], [[New York (Bundesstaat)|NY]] Alexander der Große Alexander der Große A  
 20. Juli und 30. Juli 356 v. Chr [[Pella (Makedonien)|Pella]] [[10. Juni]] [[323 v. Chr.]] [[Babylon]] Anthony Hope Hope, Anthony Hawkins, Anthony Hope englischer [[Rechtsanwalt]] und [[Autor]] [[9. Februar]]  
 Turing, Alan britischer Logiker, Mathematiker und Kryptanalytiker]] 23. Juni 1912 [[London]] [[7. Juni]] 1954 [[Wilmslow]] Arthur Harris Harris, Arthur Travers im [[Zweiter Weltkrieg|Zweiten Weltkrieg]]  
 [[Luftmarschall]] der [[Vereinigtes Königreich|britischen]] [[Royal Air Force]] [[13. April]] [[1892]] [[Cheltenham]] [[5. April]] [[1984]] Arthur Wellesley, 1. Herzog von Wellington Wellington, Arthur Wellesle  
 Feldmarschall und Staatsminister [[30. April]] [[18. Mai]] [[1769]] "Mornington House", "24 Upper Merrion Street", [[Dublin]], Irland [[14. September]] [[1852]] [[Walmer Castle]] bei [[Dover]], Grafschaft  
 [[Vereinigte Staaten|US-amerikanische]] [[Schauspieler]] in [[4. Juni]] [[1975]] [[Los Angeles]], [[Kalifornien]], [[Vereinigte Staaten|USA]] Archimedes Archimedes Archimedes von Syrakus griechischer Math  
 [[Syrakus]] auf [[Sizilien]] [[212 v. Chr.]] [[Syrakus]] auf [[Sizilien]] Aristoteles Aristoteles [[Der Stagirit]], [[Der Philosoph]] [[Griechenland|griechischer]] [[Philosoph]] und [[Naturforscher]] [[384 v. Chr.]] [[S  
 [[Chalkis]] [[Euböa]] Abraham Lincoln Lincoln, Abraham 16. Präsident der Vereinigten Staaten von Amerika [[12. Februar]] [[1809]] bei [[Hodgenville]], Hardin County "(heute: LaRue County)", [[Kentucky]]  
 [[Washington (District of Columbia)]] [[Vereinigte Staaten|USA]] Angela Merkel Merkel, Angela Bundeskanzlerin der Bundesrepublik Deutschland (CDU) 17. Juli 1954 [[Hamburg]] Alicia Silverstone Silverst  
 [[Schauspieler]] in [[4. Oktober]] [[1976]] [[San Francisco]], [[Kalifornien]], [[Vereinigte Staaten|USA]] Al-Biruni Al-Biruni al-Biruni, Abu al-Raihan Muhammad ibn Ahmad persischer Mathematiker, Wissens  
 [[Kath]] in [[Zentralasien]] (damals Teil der persischen Provinz [[Chorasani]]) [[9. Dezember]] [[1048]] [[Ghazni]] (damals in der persischen Provinz [[Chorasani]], heute in [[Afghanistan]]) Anna Seghers Segher  
 Schriftstellerin 19. November 1900 [[Mann]] 1. Juni 1983 [[Berlin]] Arvo Pärt Pärt, Arvo Estnischer [[Komponist]] [[11. September]] [[1935]] [[Paide]] (Weissenstein), [[Estland]] Aleister Crowley Crowley,  
 [[Okkultismus|Okkultist]], [[Mystik]]er, Poet, Künstler, Künstler-Manager und [[Sozialkritik]]er [[12. Oktober]] [[1875]] [[Leamington Spa]], England [[1947]] [[Hastings]], England Anne Haigis Haigis, Anne  
 [[1955]] [[Rom]] → auch hier muss der Offsetwert hoch gezählt werden]] Augustin Augustin, Augustinus des Turin Augustinus des Turin als Erster genannt Augustin 23. August 354 [[Nico]] [[Rom]] oder [[Velitiae]] 19. August 430  
 Schopenhauer, Arthur [[Deutschland|deutscher]] [[Philosoph]] [[22. Februar]] [[1788]] Stutthof bei [[Danzig]] [[21. September]] [[1860]] [[Frankfurt am Main]] Aemilianus (Kaiser) Aemilianus Aemilius Aemil  
 253 Aurelian Aurelianus, Lucius Domitius Aurelian römischer Kaiser von 270 bis 275 9. September 214 [[Moesia]] oder [[Sirmium]], [[Pannonien]] 275 bei [[Konstantinopel]] (heute [[Istanbul]]) Andrei Dmitri  
 [[Russland|russischer]] [[Kosmologe]] [[2. März]] [[1948]] [[Moskau]], [[Russland]] Adolf Hitler Hitler, Adolf deutscher Reichskanzler, Diktator und Kriegsverbrecher [[20. April]] [[1889]] [[Braunau am Inn]]  
 Arcadius Arcadius, Flavius Arcadius Kaiser des oströmischen Reiches zwischen 395 und 408 um 377 Spanien 408 [[Konstantinopel]] Avitus Avitus Eparchius Avitus weströmischer Kaiser um 400 Auvergne 457 Agatha  
 (Pseudonym) englische Krimi-Schriftstellerin [[15. September]] [[1890]] [[Torquay]], England [[12. Januar]] [[1976]] [[Wallingford (Oxfordshire)|Wallingford]], England Alfred Nobel Nobel, Alfred Bernhard  
 [[Stockholm]] [[10. Dezember]] [[1896]] [[San Remo]] Adam Ries Ries, Adam Riese, Adam Deutscher Rechenmeister [[1492]] [[Staffelstein]], [[Oberfranken]] [[30. März]] [[1559]] [[Annaberg-Buchholz|Annaberg-Buchholz]]  
 Anders Jahre Französischer [[Physiker]] und [[Mediziner]] [[29. Januar]] [[1735]] [[Lindenberg]] [[10. Juni]] [[1804]] [[Paris]] Anders Celsius Celsius, An  
 [[Mathematiker]] und [[Physiker]] [[27. November]] [[1701]] [[Orsa|Åker]] [[25. April]] [[1744]] [[Uppsala]] Annette von Droste-Hülshoff Droste-Hülshoff, Annette Elisabeth [[17. Dezember]] [[1797]] [[Münster]]  
 [[Burg Hülshoff]] bei [[Münster (Westfalen)|Münster]] 24. Mai 1848 [[Meyersburg]] Albrecht (Preußen) Albrecht von Brandenburg-Ansbach Albrecht von Hohenhausen Reichsmeister des Deutschen  
 [[Ansbach]] [[17. März]] [[1780]] [[Königsberg]] Alexander Fleming Fleming, Alexander [[Schottland|schottischer]] Bakteriologe [[1928]] [[New York City|New York]] Alexander Fleming Entdecker des [[Penicillin]]  
 [[London]] Ada Lovelace Lovelace, Ada Augusta Ada King Byron, Countess of Lovelace britische [[Mathematiker]] in 10. Dezember 1815 [[London]] 27. November 1852 London Alexandre Dumas der Jüng  
 [[Schriftsteller]] [[21. Juli]] [[1802]] [[Paris]] [[18. Februar]] [[1745]] [[Como]], [[Italien]] [[5. März]] [[1827]] [[Comnago]] bei Como Antoni van Leeuwenhoek Leeuwenhoek, Antoni van [[Niederl  
 Denkens in der [[Medizin]] [[1. Januar]] [[1514]] [[Brüssel]] [[15. Oktober]] [[1564]] auf [[Zakynthos]] Ariel Scharon Scharon, Ariel Scheinermann, Ariel israelischer [[Politiker]] [[27. Februar]] [[1928]] [[K  
 Achim von Arnim, Ludwig Achim von deutscher [[Dichter]] der [[Heidelberger Romantik]] [[26. Januar]] [[1781]] [[Berlin]] [[21. Januar]] [[1831]] auf [[Schloss Wiepersdorf]], Kreis [[Jüterbog]] Alessandro  
 Anastasio Graf von Begründer der [[Elektrizitätslehre]] [[18. Februar]] [[1745]] [[Como]], [[Italien]] [[5. März]] [[1827]] [[Comnago]] bei Como Antoni van Leeuwenhoek Leeuwenhoek, Antoni van [[Niederl

→ Keine Kommas

■ Limit bei 3000

→ auch hier muss der Offsetwert hoch gezählt werden

## Ausgabe:

Andy Warhol, Warhol, Andy Warhola, Andrew Begründer und bedeutendster Künstler der [[Pop-Art]] [[6. August]] [[1928]] [[Pittsburgh]] [[22. Februar]] [[1987]] [[New York City|New York]], [[New York (Bundesstaat)|N.Y.]]

# Direkter Zugriff auf Toolserver

18

## Wikimedia Toolserver

- Besteht aus aktuellem Datenbank-Dump
- Theoretisch besteht Möglichkeit direkt auf Toolserver zuzugreifen

### Vorteil

- Sehr elegante Lösung
- Relativ einfacher und schneller Zugriff auf Personendaten

### Nachteil

- Bisher nur sporadische Reaktion von Wikimedia-Admin auf die Zugriffsanfrage

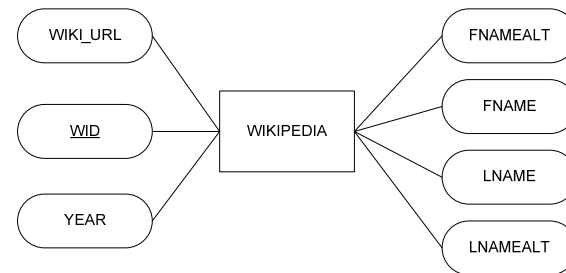
# Auswahl der Spalten

19

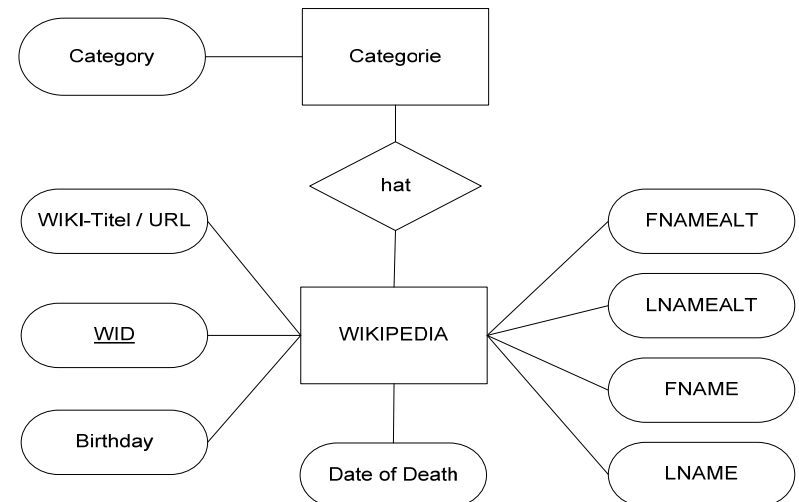
## Mögliche Spalten

- Wikipedia-Titel / URL
- Name
- Alternativnamen
- Kurzbeschreibung
- Geburtsdatum
- Geburtsort
- Sterbedatum
- Sterbeort

## Unser bisheriger Datenbankentwurf



## Unser Vorschlag



# Auswahl der Daten

20

- **Bisher ca. 92.000 Personennamen in unserer Wikipediatabelle**
- **Theoretisch 163.212 Personennamen möglich**
  - **Aber:**
    - Brauchen wir
      - ◇ Römische Kaiser
      - ◇ den achten Kalif der Umayyaden usw.
      - ◇ **Gegenbeispiel:** Was passiert, wenn Goethe Geburtstag hat
  - Ab welchem Geburtsdatum / Sterbedatum nehmen wir Personen auf?

# Quellen

21

- Mining the Web
- Java ist auch eine Insel
- <http://jerichohtml.sourceforge.net>
- <http://jerichohtml.sourceforge.net/doc/api/index.html>
- [http://de.wikipedia.org/wiki/Wikipedia:WikiProjekt\\_Vorlagenauswertung](http://de.wikipedia.org/wiki/Wikipedia:WikiProjekt_Vorlagenauswertung)