Studienstiftung des deutschen Volkes

# Standards und Werkzeuge Überblick

Josef Schneeberger

Studienstiftung Greifswald 2008
„Kulturerbe digital“

Studienstiftung des deutschen Volkes

# Übersicht

- Das Projekt „Kulturerbe digital“
- Was ist ein digitales Dokument?
- Dimensionen eines digitalen Dokuments
- Standards: XML & co.
- Werkzeuge für das Projekt

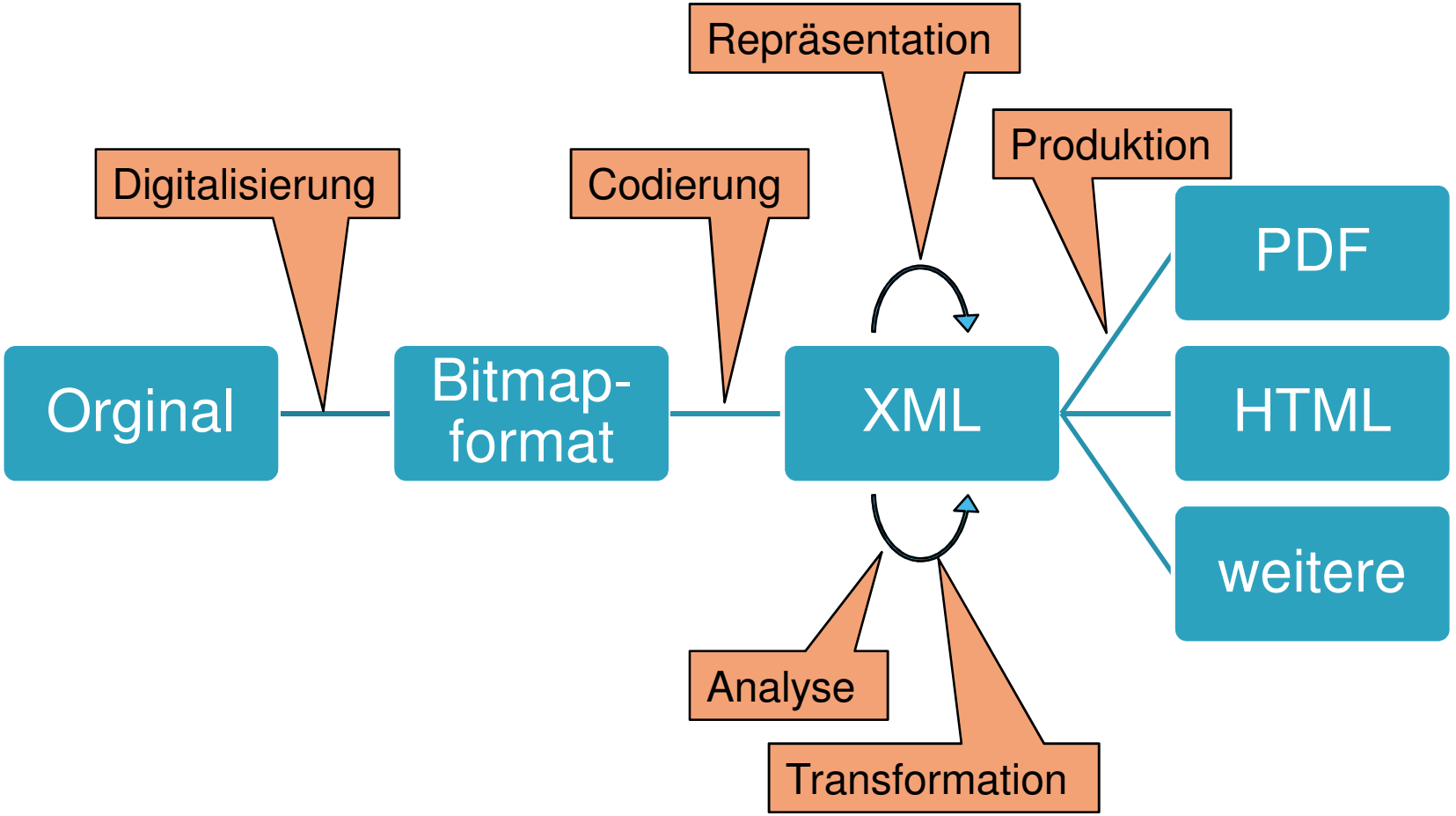
- Hinweise zur Erfassung des „deutschen Ptolemäus“

Studienstiftung
des deutschen Volkes

# Arbeitsplan

Studienstiftung des deutschen Volkes

- Digitalisierung
- Codierung
- Anreicherung / Repräsentation
- Interpretation
- Auswahl
- Publikation



Studienstiftung
des deutschen Volkes

# Digitale Dokumente

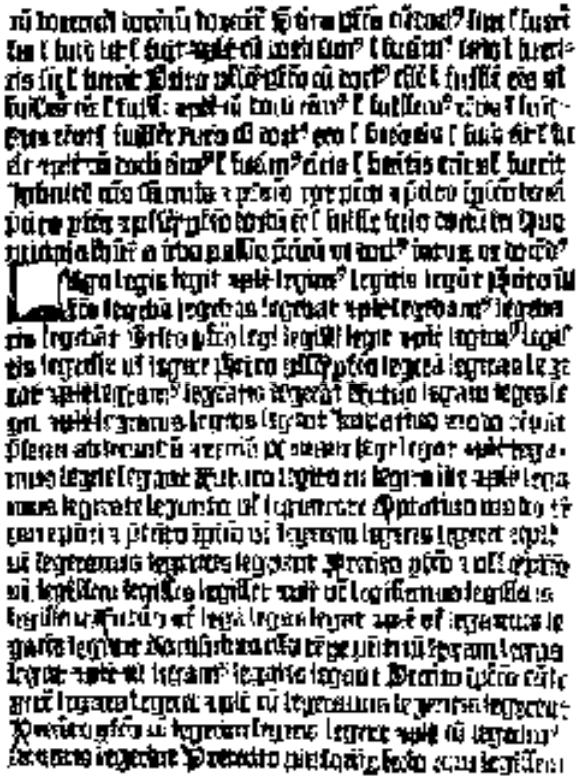
Studienstiftung des deutschen Volkes

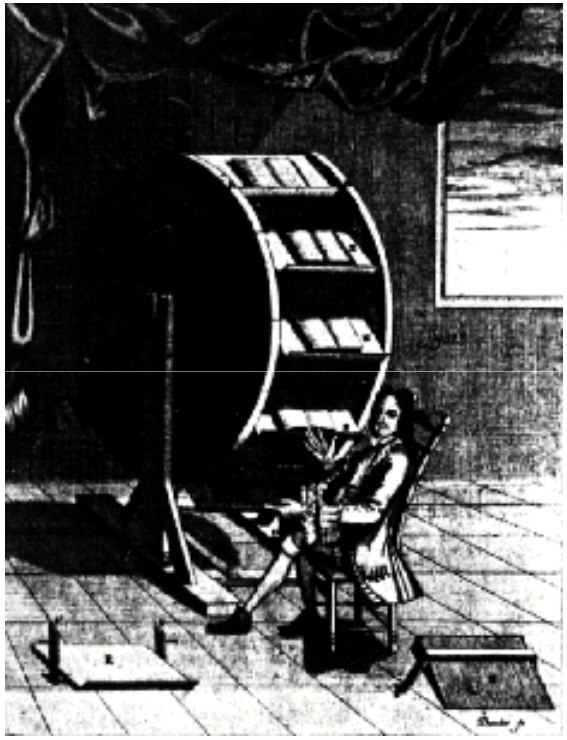
# Was ist eigentlich ein digitales Dokument?

- Ein Papierdokument zum Ausdrucken und Lesen auf einem Computer (PDF, PS, dvi, doc, odf, ...)
- ... das was ein Browser darstellt.
- Hypertext.
- Dokumente können auf dem Computer sehr unterschiedlich dargestellt werden. Ist das immer noch ein Dokument?

Studienstiftung des deutschen Volkes

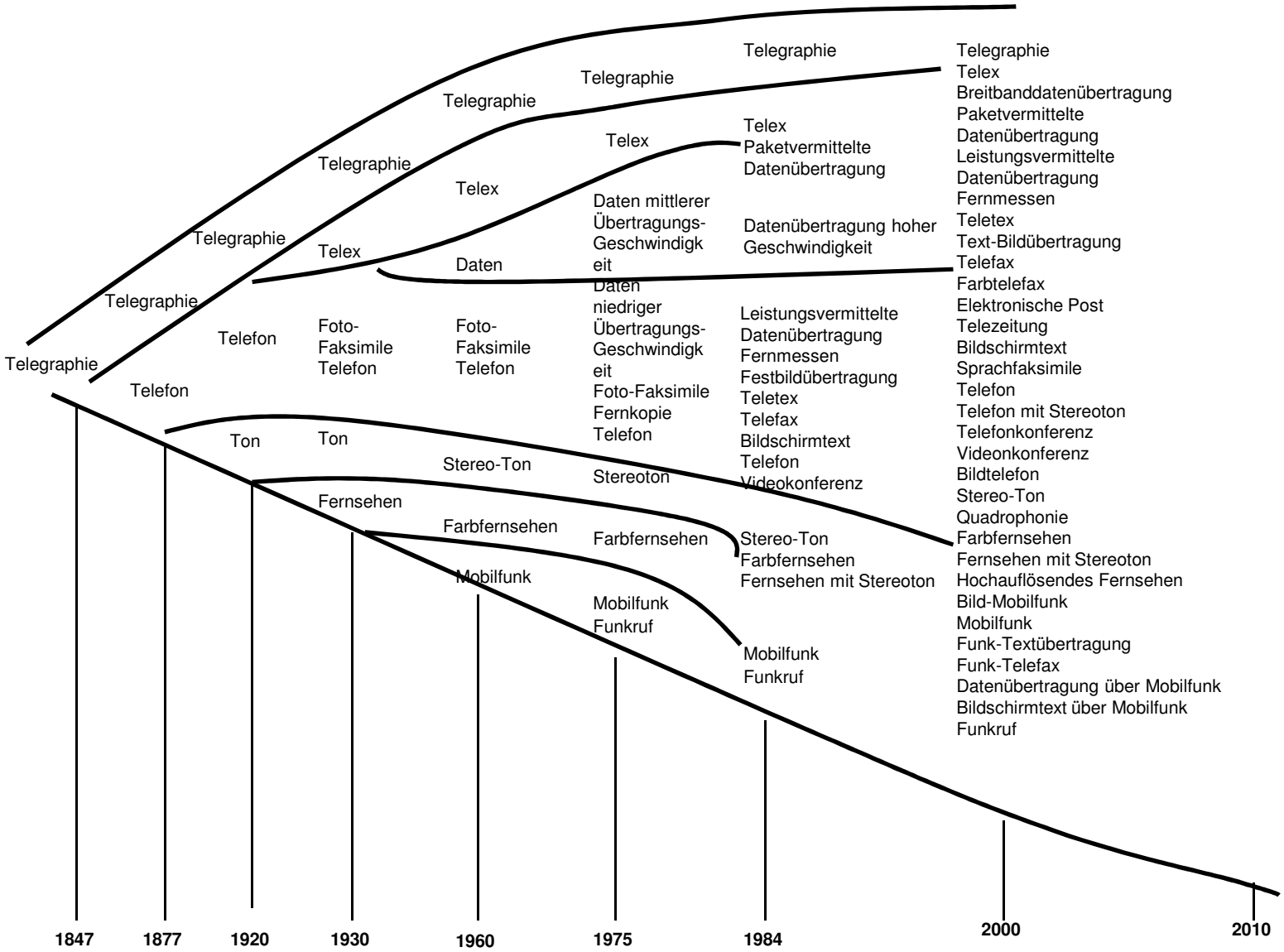
# Entlinearisierung und Technologie





Studienstiftung des deutschen Volkes

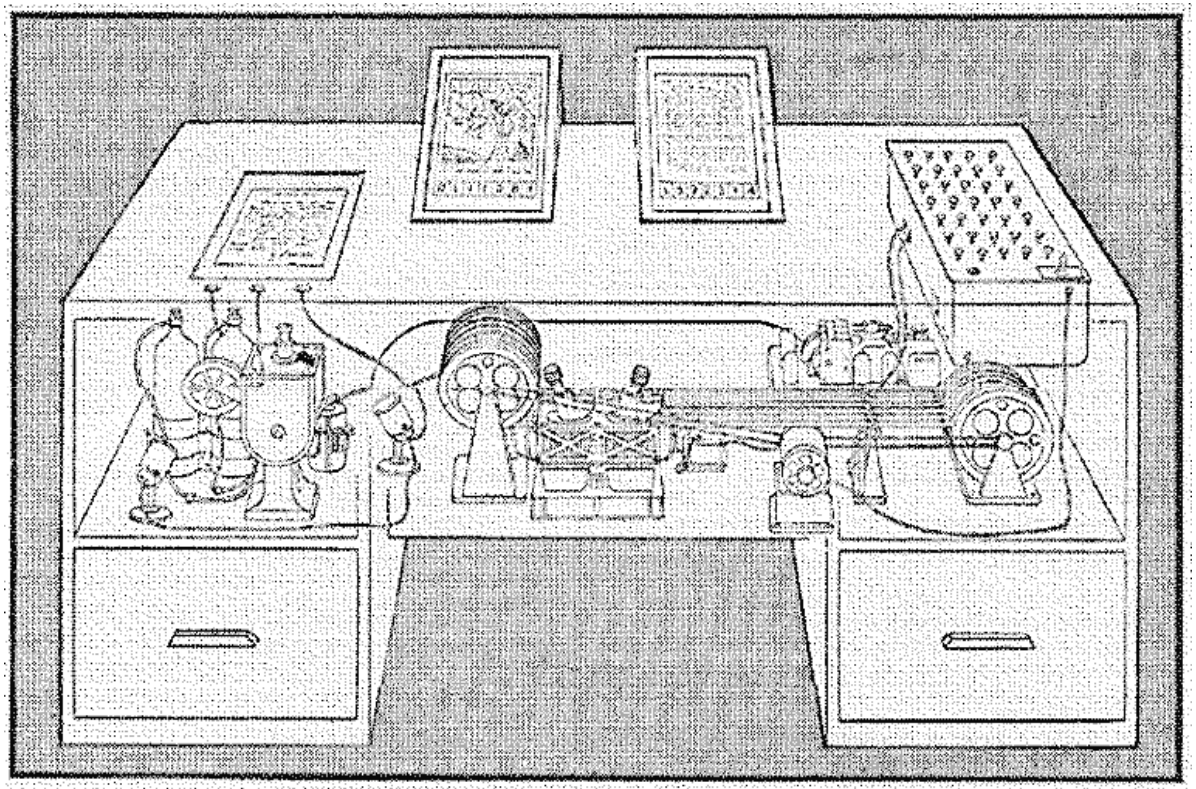
# Explosion digitaler Medien



Studienstiftung des deutschen Volkes

# Bush und Memex

(V. Bush: "As we may think", in Atlantic Monthly, 1945).



**MEMEX** in the form of a desk would instantly bring files and material on any subject to the operator's fingertips. Slanting translucent viewing screens magnify supermicrofilm filed by code numbers. At left is a mechanism which automatically photographs longhand notes, pictures and letters, then files them in the desk for future reference.

**AS WE MAY THINK** CONTINUED

Studienstiftung des deutschen Volkes

# Kurze Geschichte: Hypertext

- **1965** *Ted Nelson* prägt den Begriff *Hypertext* und entwirft das System *Xanadu* bezeichnet er als ("a magic place of literary memory").
- ´**1962-76** *Augment*-Projekt von *Doug Englebart* am Stanford Research Institute (SRI) in Kalifornien: Entwicklung des System *NLS* (oN Line System) als assoziativer Dokumentspeicher mit ca. 100.000 einzelnen "Einträgen".
- **1967** Eine Gruppe um *Andries van Dam* an der Brown University in Providence (Ostküste) entwickelt das sog. *Hypertext Editing System.* Das System benötigte 120 KByte Hauptspeicher und lief auf IBM/360-Großrechnern. Es diente u.a. zur Dokumentation der Apollo-Missionen.
- **1982** Auslieferung des Dokumentationssystems *Document Examiner* (für sog. "Lisp-Maschinen") durch die Firma *Symbolics* als erstes Hypertext-System für Endanwender.
- **1985** Am *XEROX Palo Alto Research Center* (PARC) wird das System *NoteCards* entwickelt, das speziell die Sammlung und assoziative Strukturierung von Ideen in offenen Diskussionen unterstützt (*idea processing*).
- **1988** Durch das System *HyperCard* von Apple erreicht das Hypermedia-Konzept auch die Benutzer von Personal Computern

Studienstiftung des deutschen Volkes

# Komponenten von Dokumenten

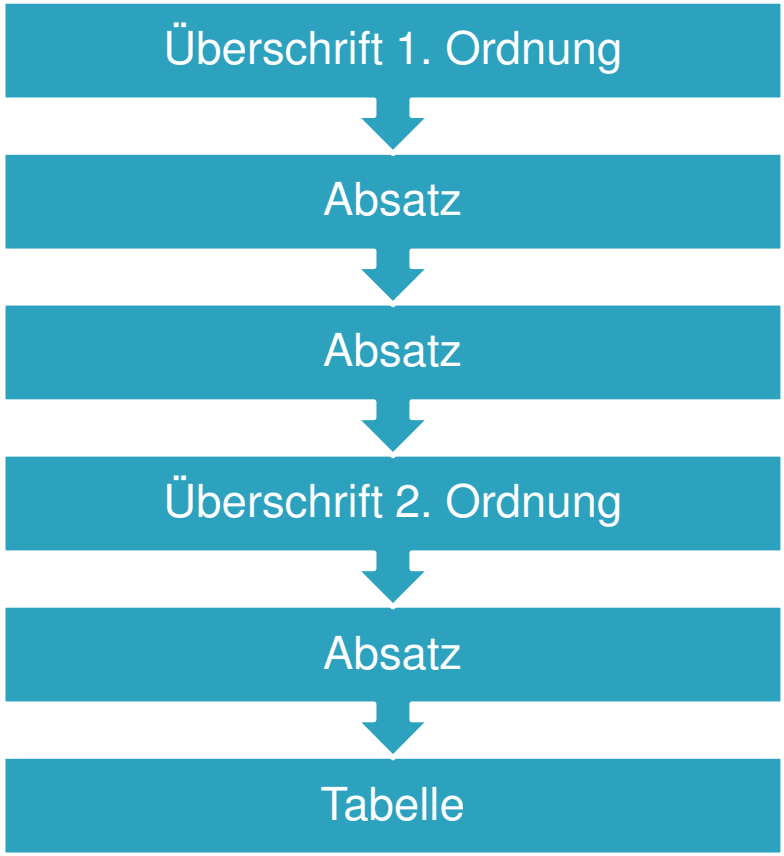
- Text (Absätze, Aufzählungen, Überschriften)
- Bilder
- Zeichnungen / Grafiken
- Tabellen
- Verweise
- Formeln

... das war‘s gibt es sonst noch etwas?

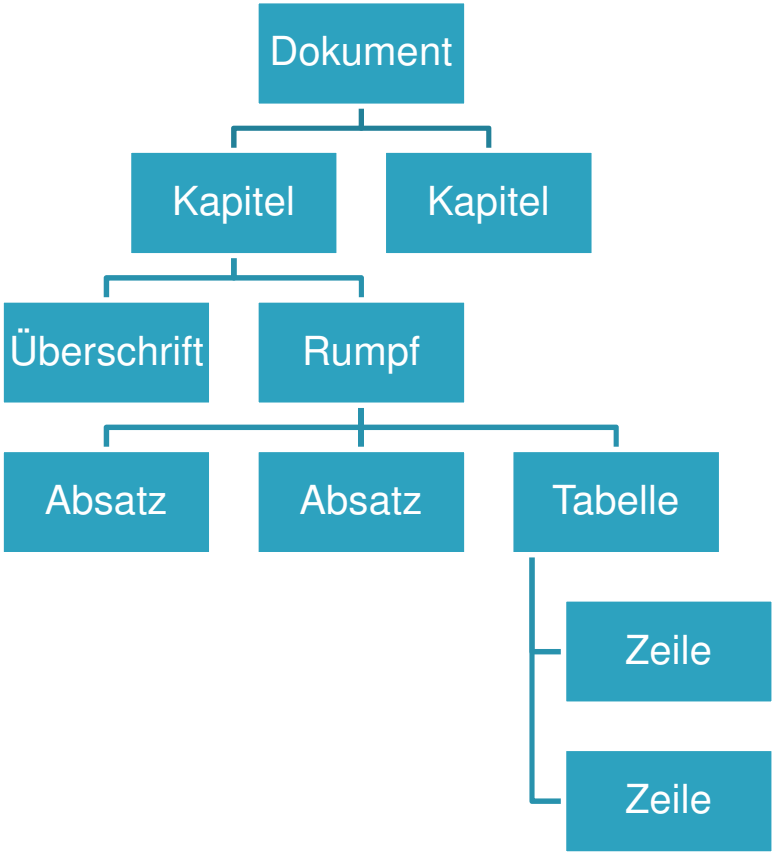
Studienstiftung des deutschen Volkes

# Zwei Repräsentationsansätze

- Eine lineare Sequenz von Absätzen (Word, Framemaker)




- Ein Baum

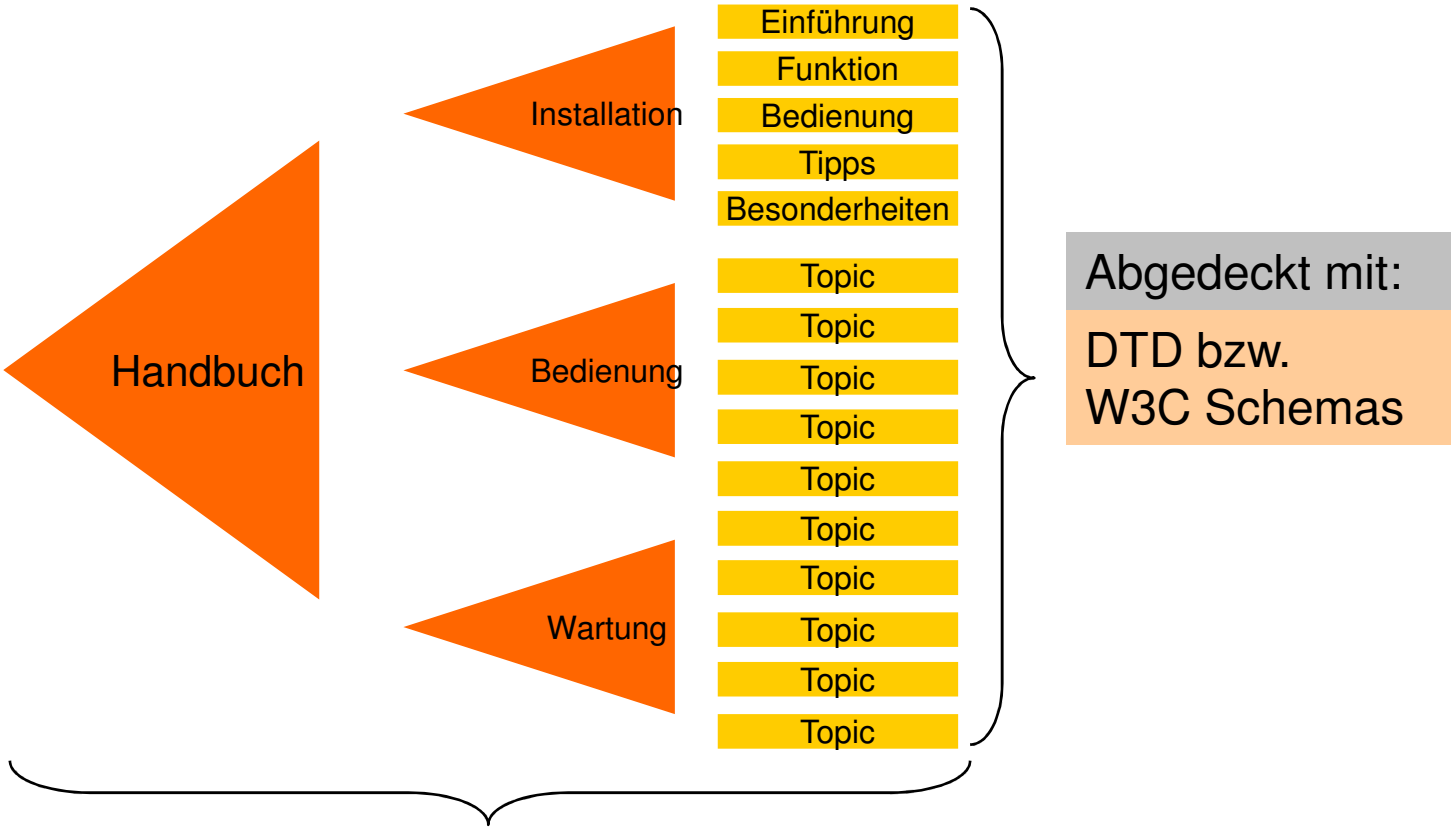


Studienstiftung des deutschen Volkes

# Dimensionen eines digitalen Dokuments

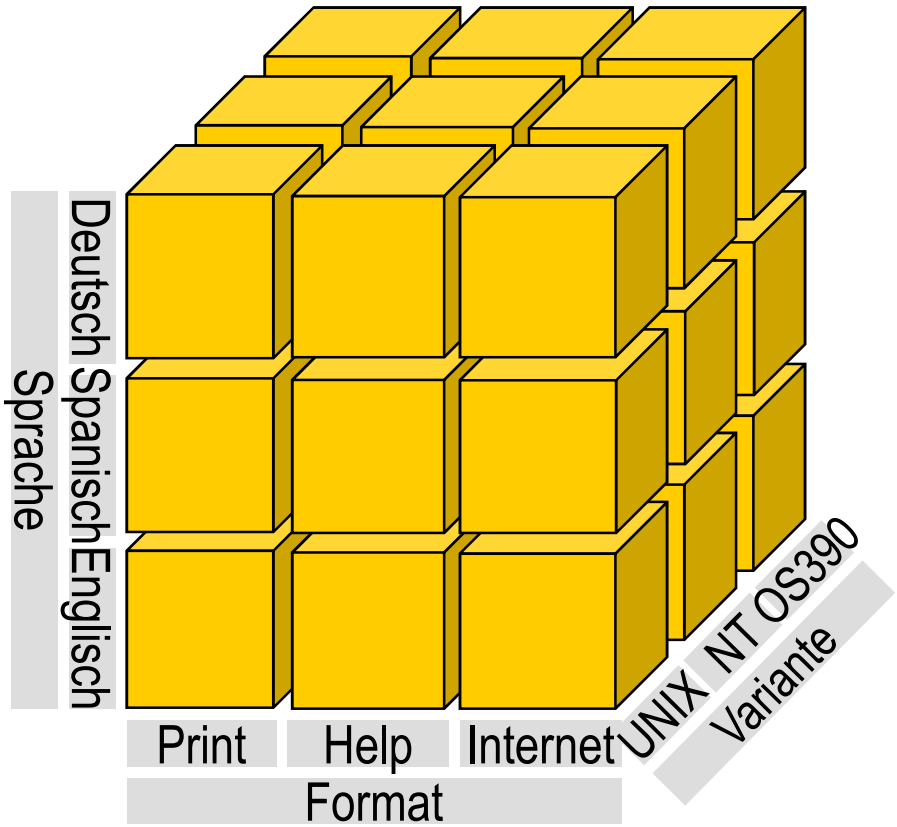
Studienstiftung des deutschen Volkes

# State of the Art: XML und DTDs



Studienstiftung des deutschen Volkes

# Informationsdimensionen

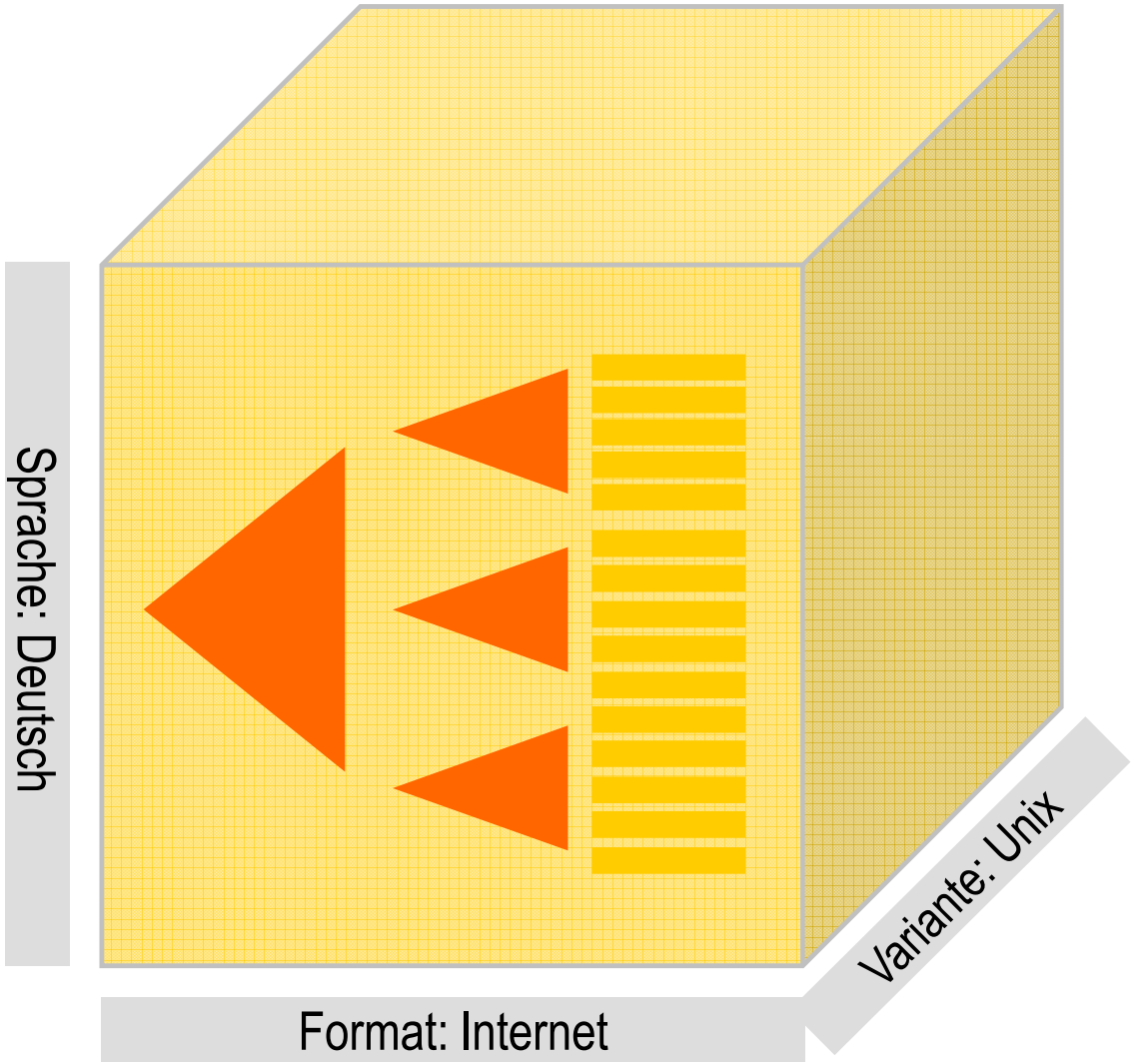



Nicht Abgedeckt mit XML:

Dimensionen und Aspekten

Studienstiftung des deutschen Volkes

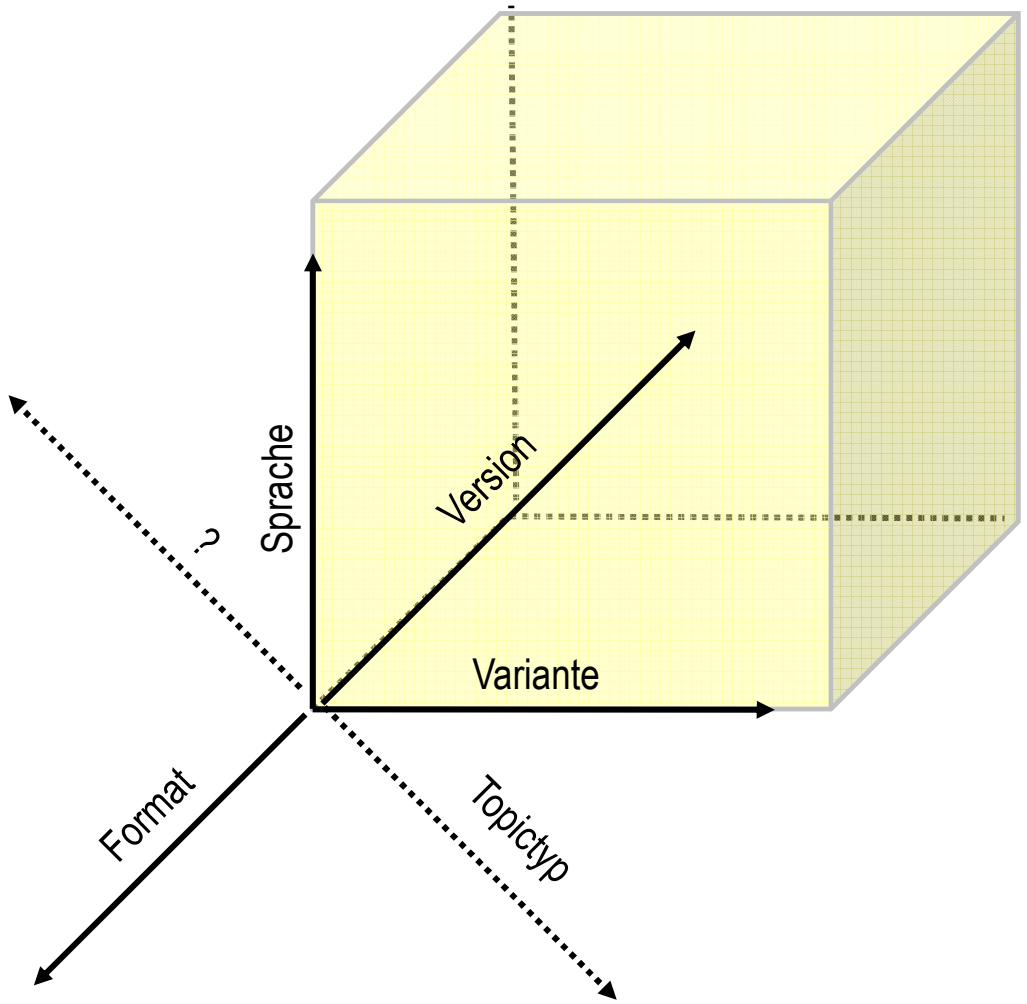
# Eine Strukturbaum existiert in mehreren Dimensionen!



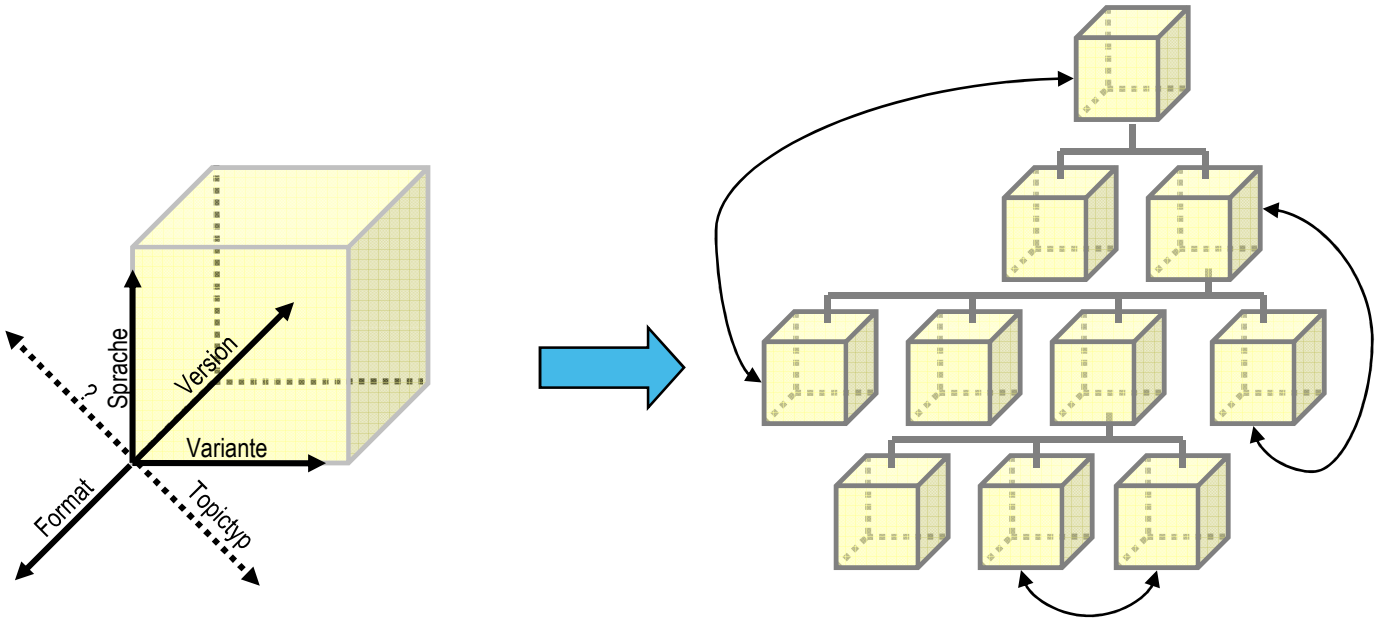
Studienstiftung des deutschen Volkes

# Was sind somit die eigentlichen Probleme?

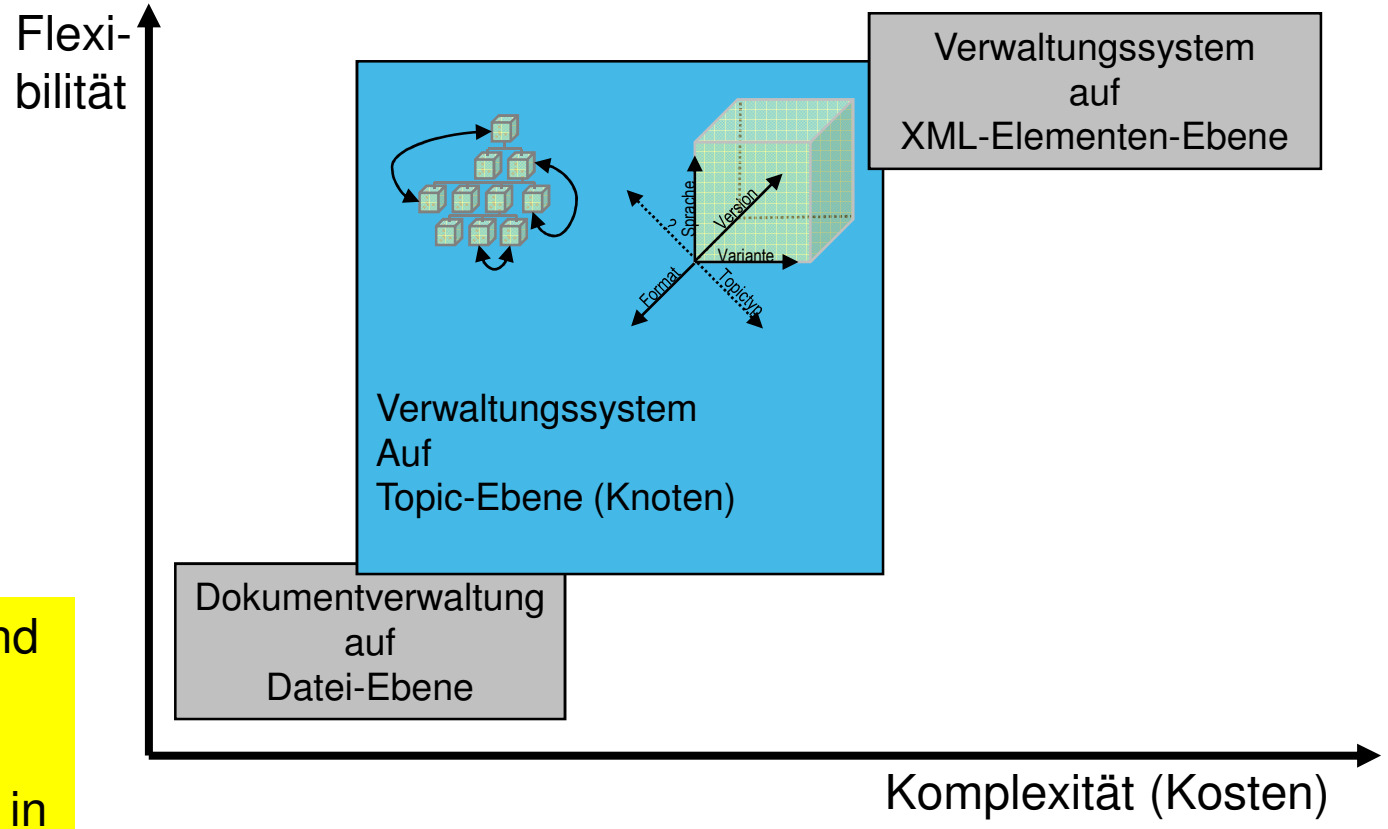
Wie behält man die Übersicht über ein solches Dokument?



# Dimensionen und Hierarchien



# Wozu Knoten bzw. Topics?




Warum und wie speichert man XML in einer Datenbank?

Studienstiftung des deutschen Volkes

# Funktionsblöcke im CMS

| | |
|---|---|
| Workflow | **Metadaten für Status-Informationen, Aufgabenverwaltung als zentrale Steuerung, Event für kundenspezifische Workflow-Ereignisse** |
| Publishing | **Produktion von dynamischen Websites für Redakionsanwendungen und Intranet über SchemaText WebServer, Anbindung von anderen Portalen** |
| Producing | **Produktion von statischen Exportformaten für Print (Word, FrameMaker, Interleaf), Online Help (WinHelp, HTMLHelp, JavaHelp) und beliebige andere Formate, Datenbanken** |
| Composing | **Grafische Oberfläche für Strukturbäume und Netzwerke, XML-Ressourcen als Bausteine verwenden, Sprachverwaltung, Variantensteuerung, Linkmanagement** |
| Authoring | **Beliebige XML-Editoren, Word als XML-Editor, XML-Fragmente bearbeiten, Linkmanagement, Ressourcen-Management mit Editor-Integration, Beliebige Ressourcen-Editoren** |
| DMS / XML | **Import, Export von BLOBs; Versionen, Metadaten, Rechte, Verwendungsnachweise, XML-Speicherung, XML-Delta, Delta-Visualisierung, Zugang über LAN und Web, XML und Volltextsuche** |
| Administration | **Basis-Einstellungen, Benutzerverwaltung (Rollen, Rechte), Entwicklungsumgebung für Scheme-Scripte und Applicationserver API, Parsing, Import, Export, Datenmodellierung, Metadaten** |

Studienstiftung
des deutschen Volkes

# Warum XML?

# XML - Was ist das?

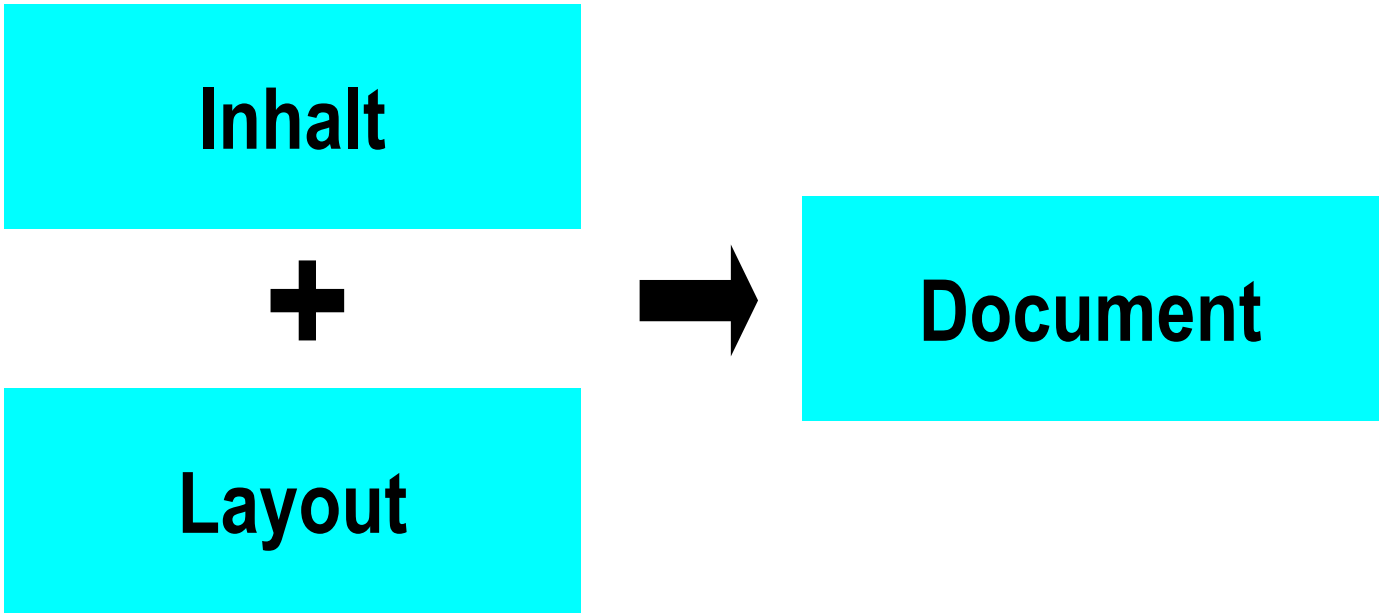
- XML = eXtensible Markup Language
- Format zur Strukturierung von Dokumenten und Daten
- Gründe für XML
  - Anwendungsspezifische Auszeichnungsmöglichkeiten
  - Trennung von Inhalt und Layout
  - Erweiterbarkeit

# Historie

- 1969: GML von Goldfarb, Mosher, Lorie
- 1989: SGML wird ISO Standard 8879
- 1989: HTML von Berners-Lee
- 1994: HTML 2.0: erste echte SGML-Anwendung (DTD)
- 1996: XML 1.0
- 1998: XML 1.0 wird W3C Standard
- 1999: XSLT, XPath, XHTML

Studienstiftung des deutschen Volkes

# Philosophie von XML / SGML



Studienstiftung des deutschen Volkes

# XML steht für strukturierte Daten

- Strukturierte Daten können beliebige Objekte enthalten
- XML ist keine Programmiersprache (aber eine formale Sprache)
- XML dient zur leichten Verarbeitung beliebiger Daten auf dem Computer
- XML ist
  - Erweiterbar
  - Plattformenabhängig
  - International (Unicode)

Aus [Mi02]

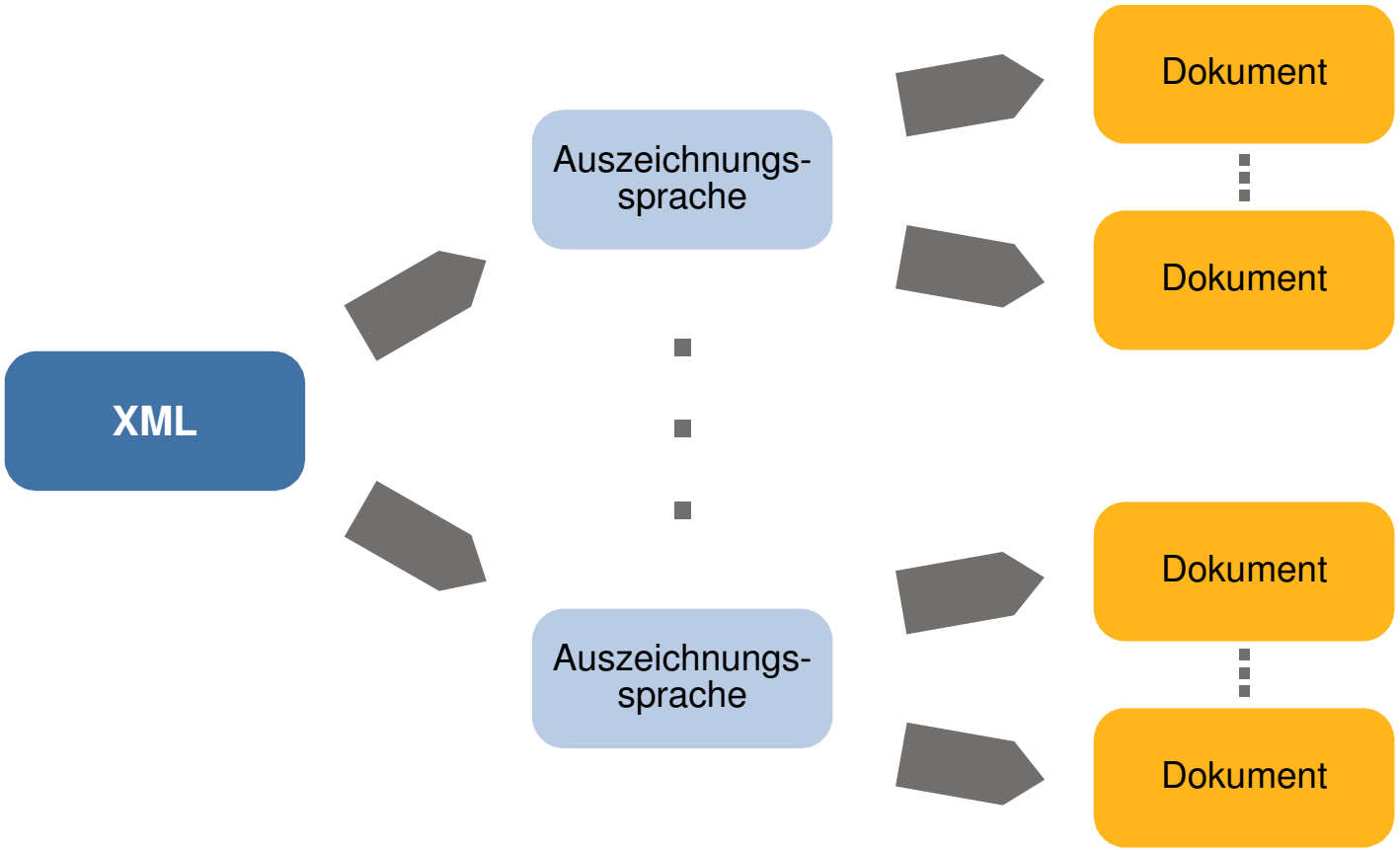
Studienstiftung des deutschen Volkes

# Es gibt (sehr) viele Werkzeuge für XML

- In (fast) jede Programmiersprache integriert
- Alle Browser können XML darstellen
- Datenbanken zur Speicherung von XML
- Protokolle zur Übertragung von XML
- Darstellung mit Schablonen (Templates)

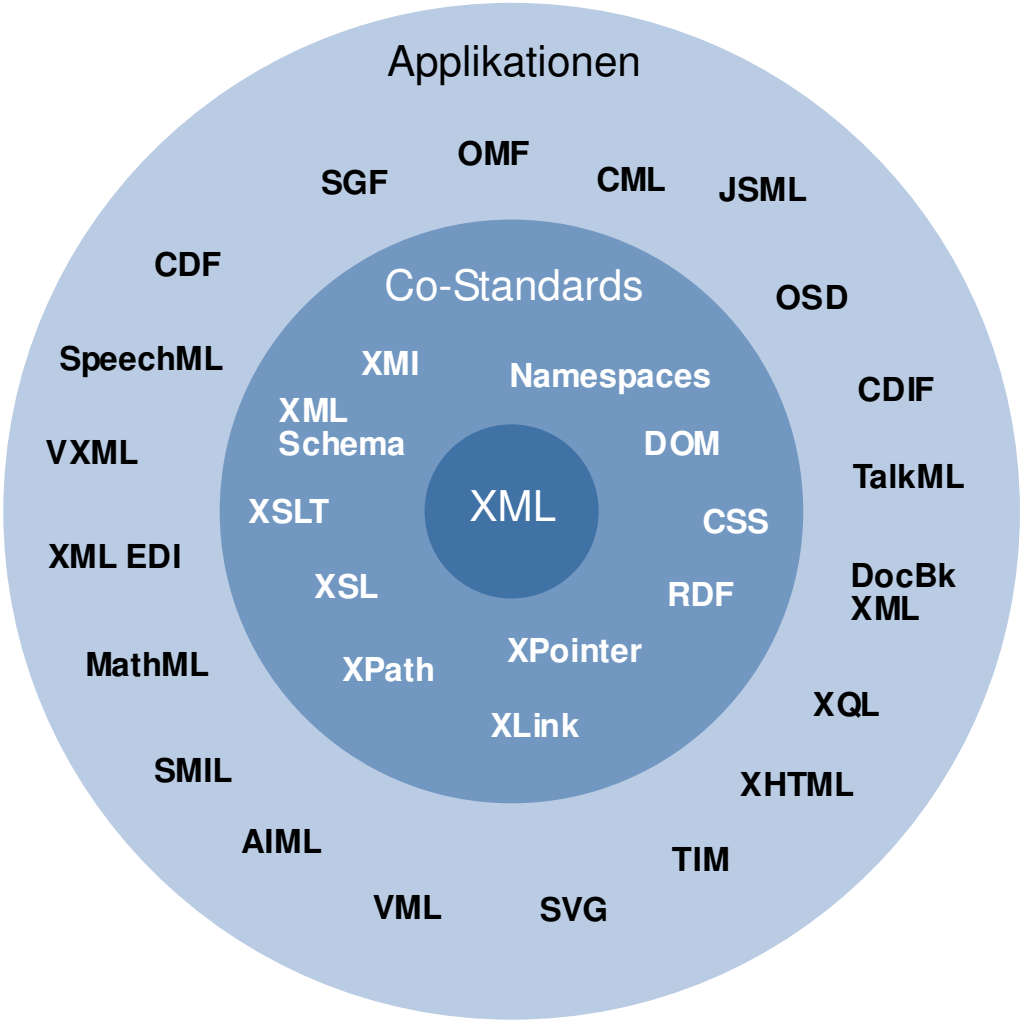
Studienstiftung des deutschen Volkes

# Auszeichnungssprachen (3)

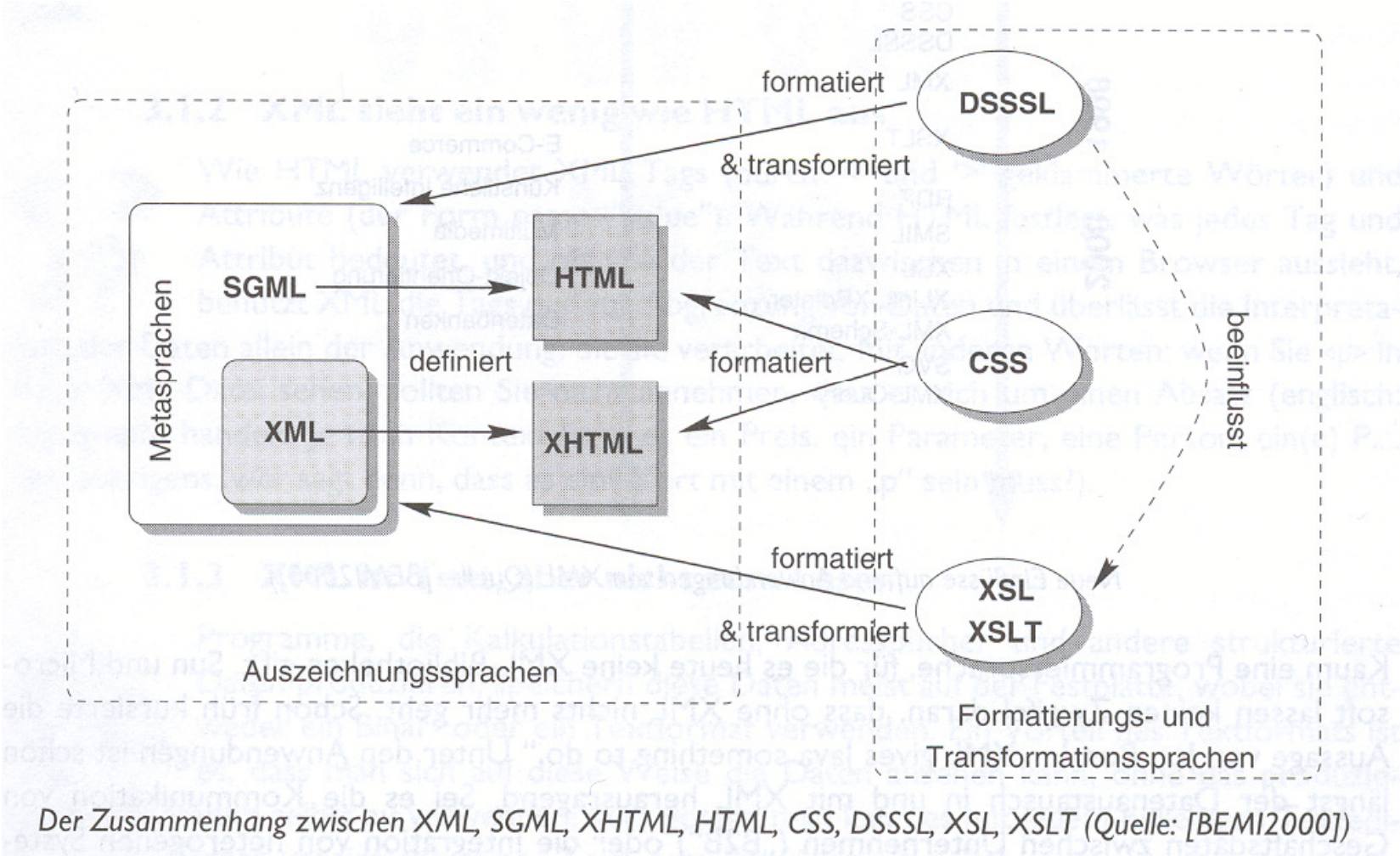


Studienstiftung des deutschen Volkes

# Ein Blick in das XML-Universum



# XML SGML XHTML HTML CSS DSSSL XSL XSLT




*Der Zusammenhang zwischen XML, SGML, XHTML, HTML, CSS, DSSSL, XSL, XSLT (Quelle: [BEMI2000])*

Aus [BeMi00]

Studienstiftung des deutschen Volkes

# Diskussion

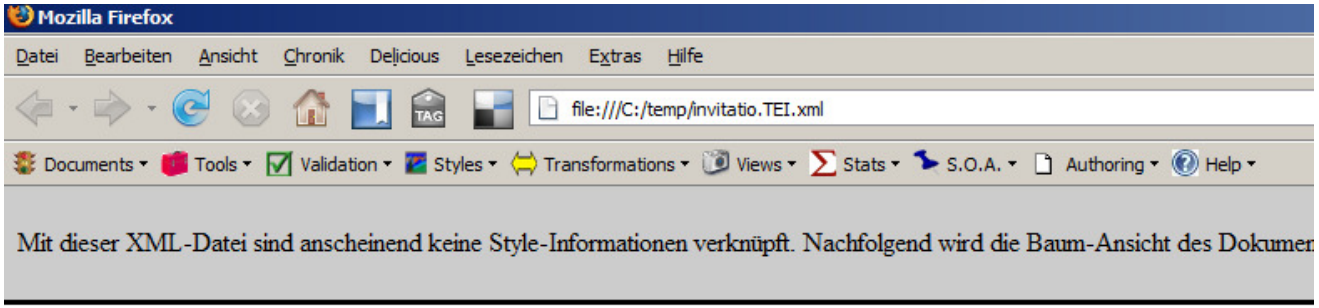
- Die Auszeichnung von Dokumenten ermöglicht die flexible Verwendungen von Dokumenten:
  - Import / Export mit Datenbanken
  - Wissensrepräsentation
  - Automatische Extraktion und Verknüpfung
- Probleme mit XML :
  - Wenn das Schema/DTD geändert wird, dann müssen alle exisitierenden Dokumente angepasst werden.
  - Schema/DTD nötig um ein Dokument zu verarbeiten.
- Ein Dokumentstandart bedeutet nicht unbedingt, dass Dokumente standardisiert werden (Beispiel Microsoft).

Studienstiftung
des deutschen Volkes

# Werkzeuge

Studienstiftung des deutschen Volkes

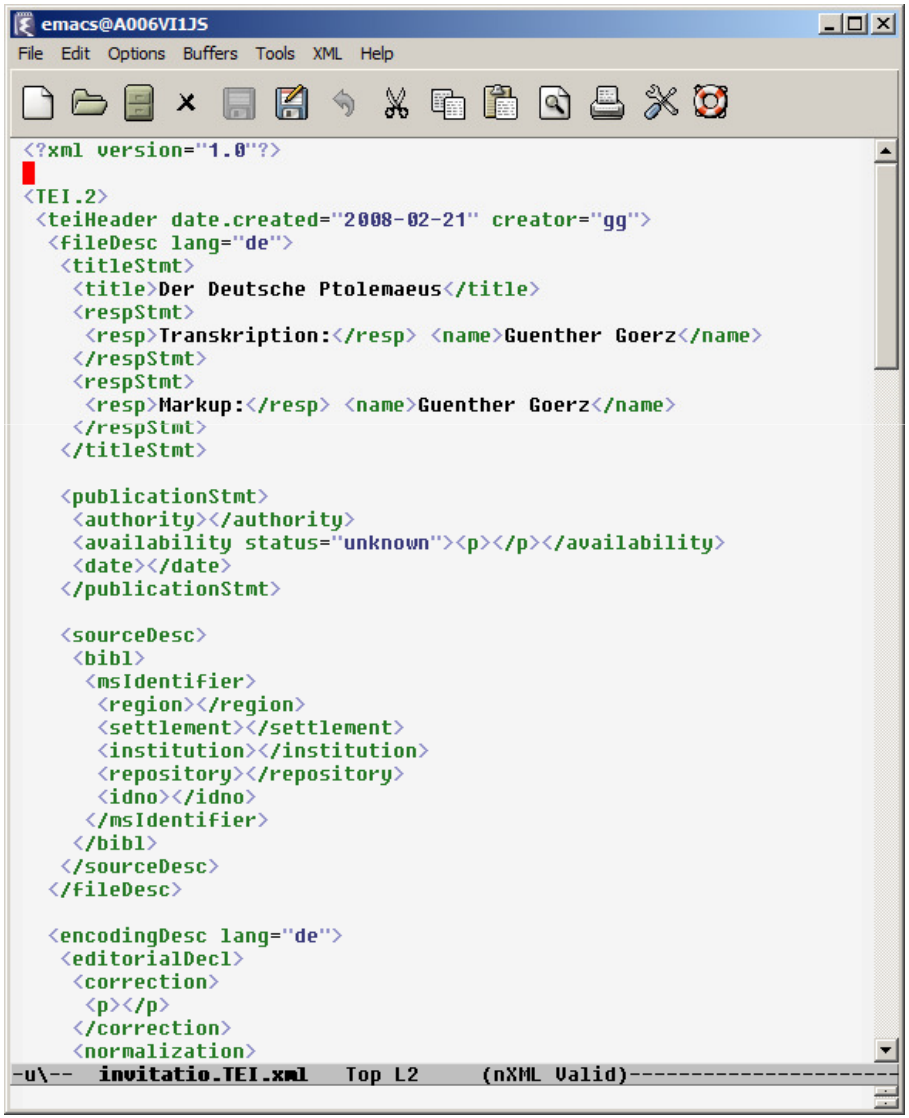
# Browser – Darstellung von XML




```
− <TEI.2>
  − <teiHeader date.created="2008-02-21" creator="gg">
    − <fileDesc lang="de">
      − <titleStmt>
          <title>Der Deutsche Ptolemaeus</title>
        − <respStmt>
            <resp>Transkription:</resp>
            <name>Guenther Goerz</name>
          </respStmt>
        + <respStmt></respStmt>
        </titleStmt>
      + <publicationStmt></publicationStmt>
      + <sourceDesc></sourceDesc>
      </fileDesc>
    + <encodingDesc lang="de"></encodingDesc>
    + <profileDesc></profileDesc>
    </teiHeader>
    <!--
     <text>
     <group>
    -->
  − <text>
    − <body lang="la">
      − <div1>
        − <p>
            <pb n="1"/>
            <lb n="1"/>
            Inuitatio lectoris in cosmographiam claudi
            <lb n="2"/>
            ptolomei Alexandrini nouiter ideomate germano
```

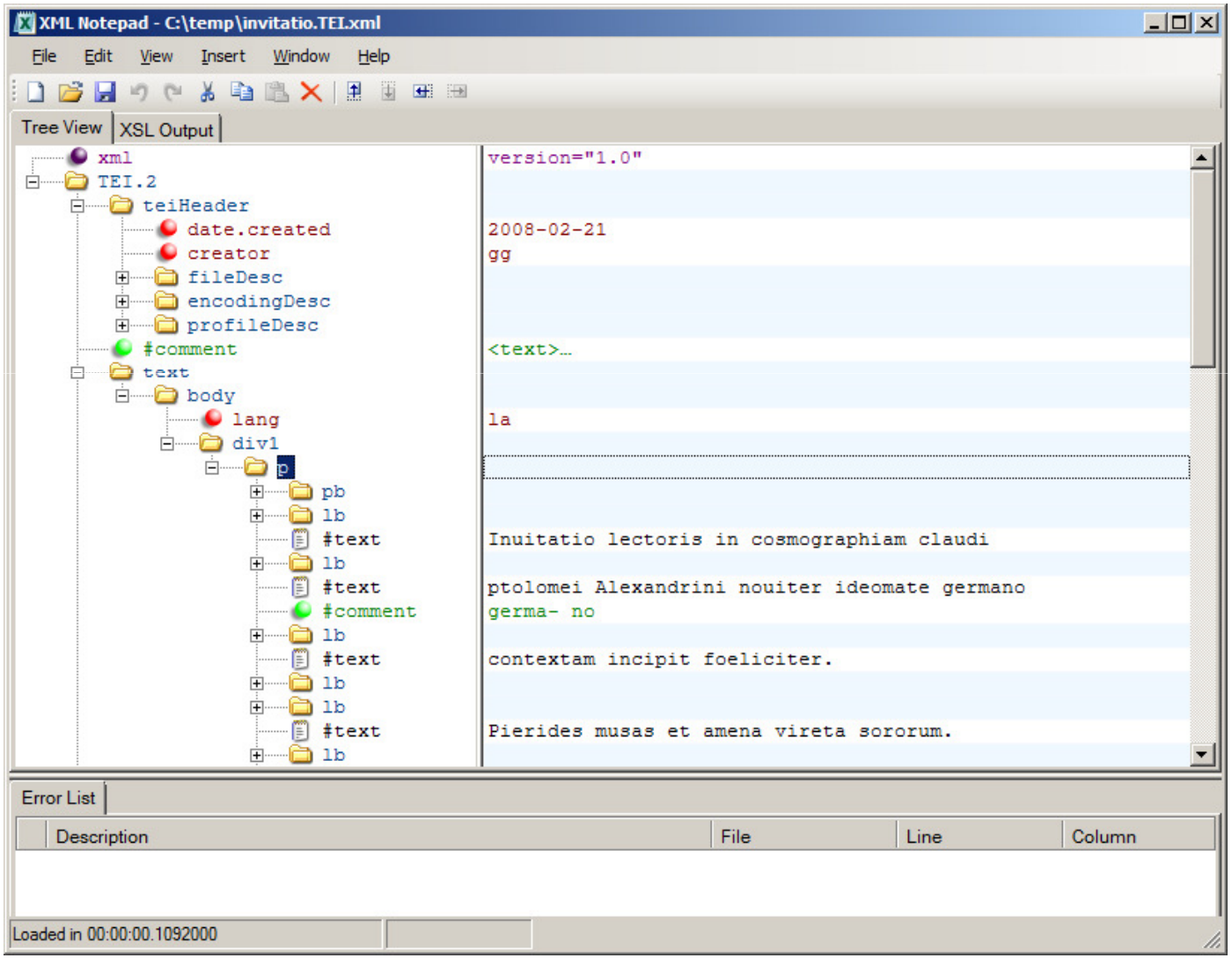
Studienstiftung des deutschen Volkes

# Editoren für XML (Texteditor, Emacs, etc.)



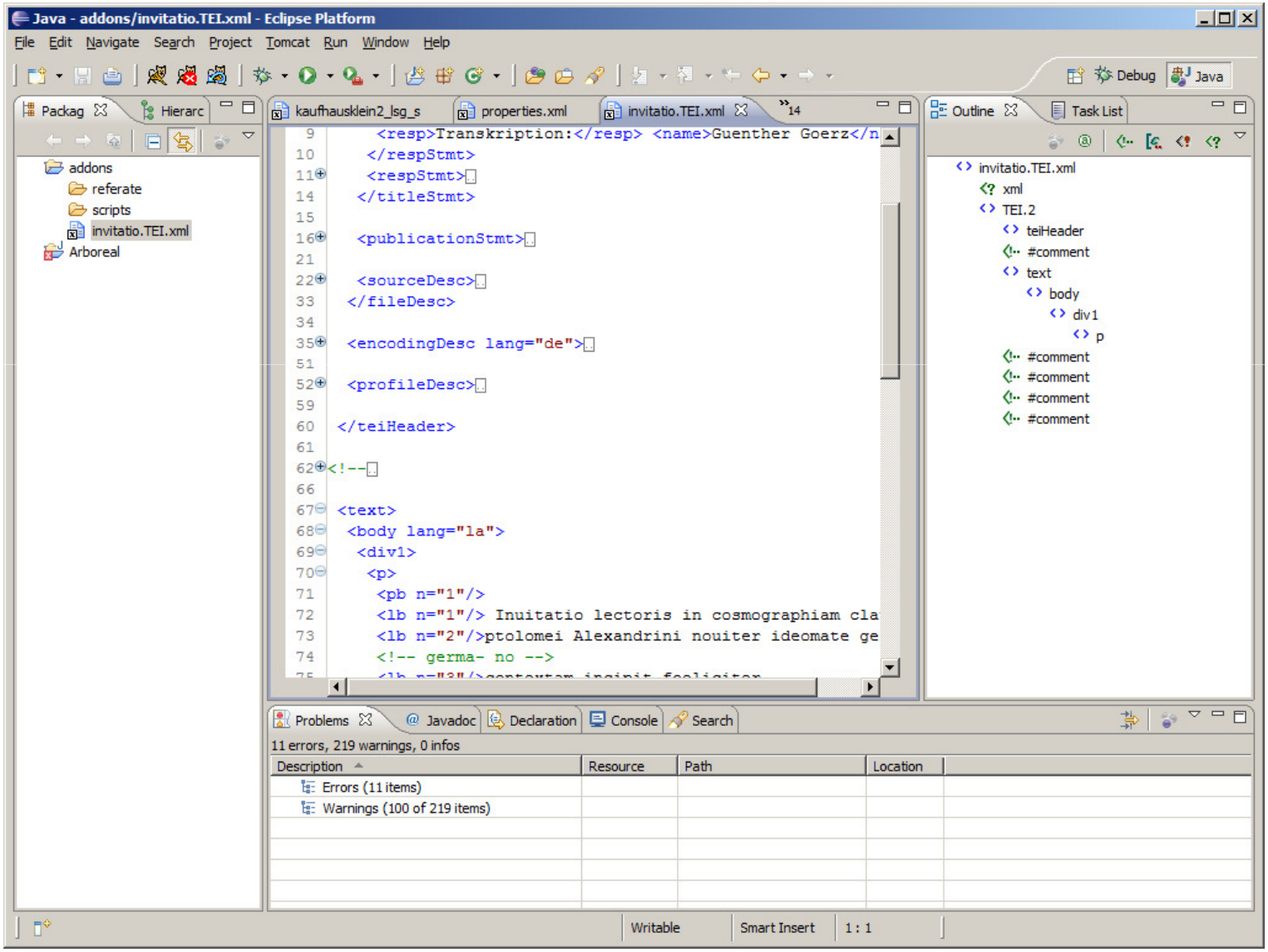
Studienstiftung des deutschen Volkes

# Editoren für XML (MS XML Notepad)



Studienstiftung des deutschen Volkes

# Editoren für XML (Eclipse – mit Schema Unterstützung)

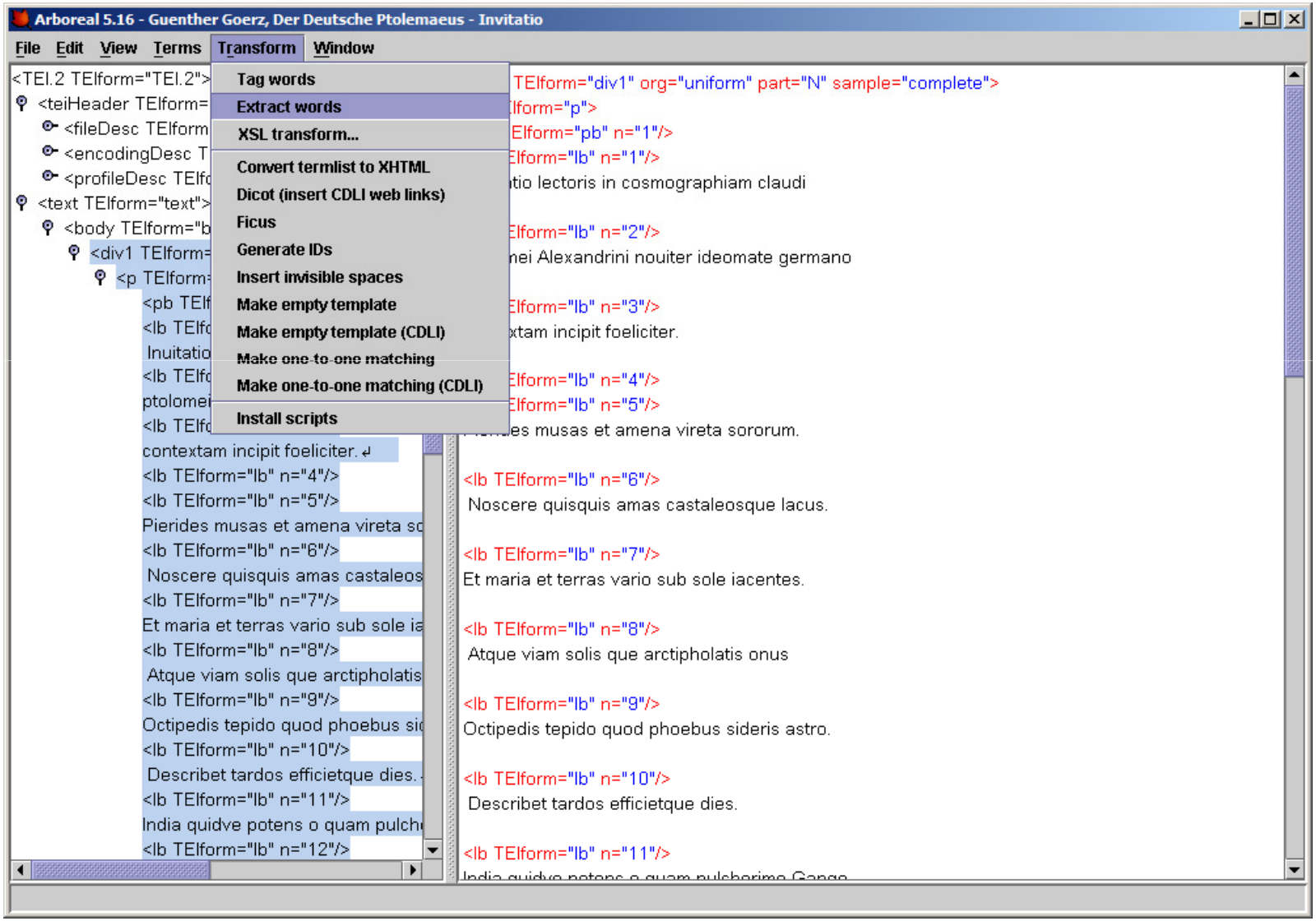


Studienstiftung des deutschen Volkes

# XML Editieren / Verarbeiten (Oxygen)

Studienstiftung des deutschen Volkes

# Verarbeitung / Transformation (Arboreal)



Studienstiftung des deutschen Volkes

# Kooperative Entwicklung von Inhalten

- Verwaltung von Versionen
- Wiki – werden wir evtl. einrichten

Studienstiftung des deutschen Volkes

# Arbeit mit dem Text

Studienstiftung des deutschen Volkes

# Erfassen der Dokumente

- Die Erfassung sollte möglichst weitgehend der Vorlage entsprechen.
  - Alle Zeichen (auch Druckfehler)
  - Zeilenumbrüche
  - Nummerierung der Zeilen und Blätter
- Auszeichnung
  - von Wörtern mit <w>

Studienstiftung des deutschen Volkes

# Basiselemente

| Tag | Wofür |
|---|---|
|  ... <br><br>@type<br>@n | Gliederungselement (von Überschrift zu Überschrift)<br>kapitel |
|  ...  | Absatz / Paragraph |
| <choice><orig>..<reg>.. | Silbentrennung / neue Zeile |
| <choice><abbr>..<expand>.. | Abkürzungen |
| <pb /> | Neue Seite |
| <lb /> | Neue Zeile |
| <hi type="initial">x</hi> | Initiale |
| <head>...</lb>...</head> | Kapitelüberschrift |
| <q lang="lat">...</q> | (lateinisches) Zitat |
| <milestone type="space"/> | Abstand in einer Zeile |

Studienstiftung des deutschen Volkes

# Optionale Elemente

| Tag | Wofür |
|---|---|
| <rs ...> <name> | Namen, Orte |
| <date ...> <time ...> | Zeitpunkte |

Studienstiftung des deutschen Volkes

# Diskussion

Studienstiftung
des deutschen Volkes

# Unicode

Studienstiftung des deutschen Volkes

# Unicode

- ISO-Standard
- 16-Bit-Code zur Darstellung von Schriftzeichen für Zwecke der Informationsverarbeitung für die wichtigsten Sprachen der Welt:
  - Z.Z. Unicode 2.1-Standard
  - enthält rund 40.000 alphabetischen Zeichen, ideographischen Zeichen und Symbole
  - Codierungen für weitere Zeichen werden laufend entwickelt

# Zeichensätze

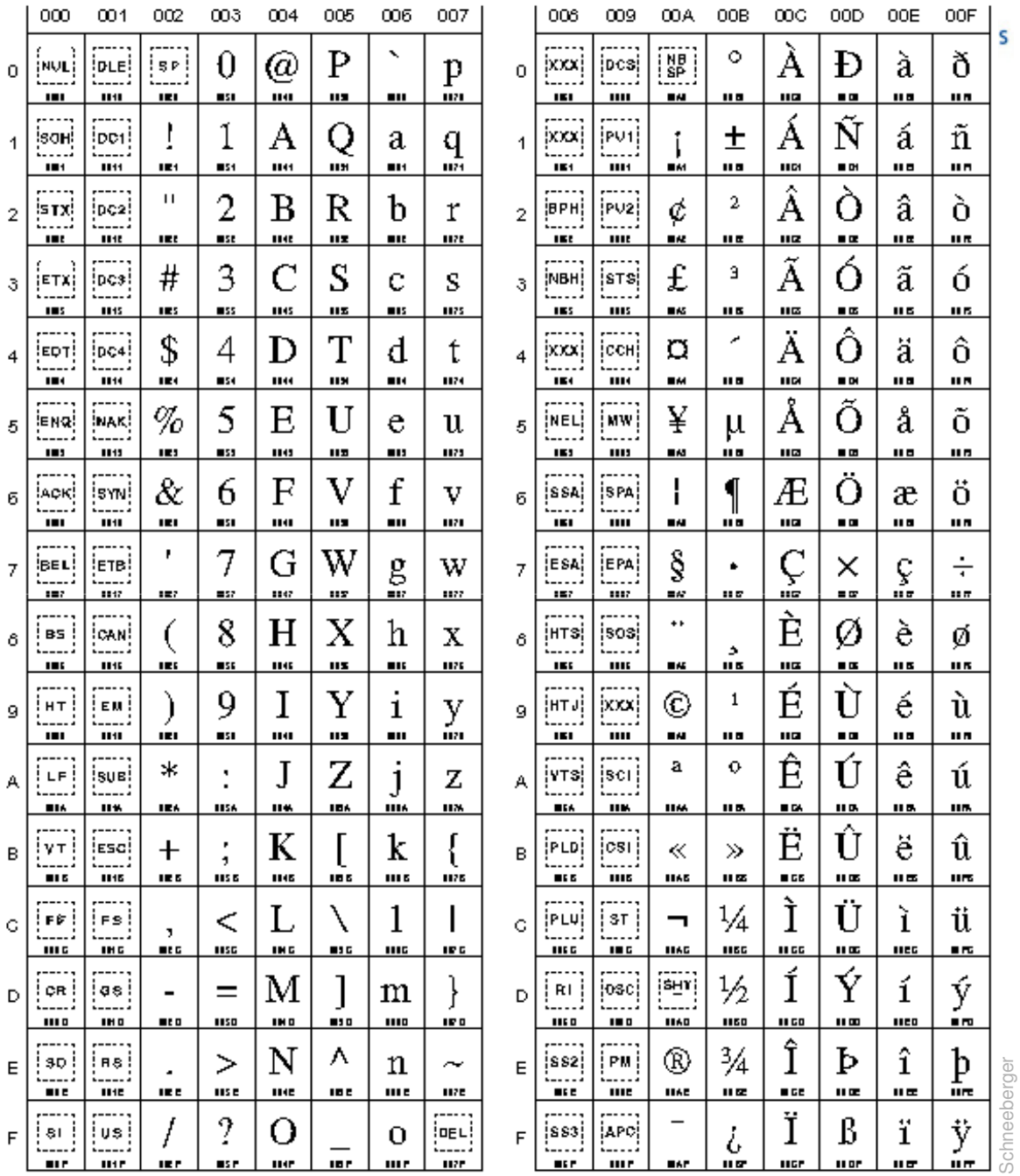
- American Standard Code for Information Interchange (ASCII)
- Die Tastatur Amerikanischer Schreibmaschinen

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | A | B | C | D | E | F |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | NUL | SOH | STX | ETX | EOT | ENQ | ACK | BEL | BS | HT | LF | VT | FF | CR | SO | SI |
| 1 | DLE | DC1 | DC2 | DC3 | DC4 | NAK | SYN | ETB | CAN | EM | SUB | ESC | FS | GS | RS | US |
| 2 | SPC | ! | " | # | $ | % | & | ' | ( | ) | * | + | , | - | . | / |
| 3 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | : | ; | < | = | > | ? |
| 4 | @ | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O |
| 5 | P | Q | R | S | T | U | V | W | X | Y | Z | [ | \ | ] | ^ | _ |
| 6 | ` | a | b | c | d | e | f | g | h | i | j | k | l | m | n | o |
| 7 | p | q | r | s | t | u | v | w | x | y | z | { | \| | } | ~ | DEL |

## Character Sets

ISO-Latin-1

... der NATO Zeichensatz

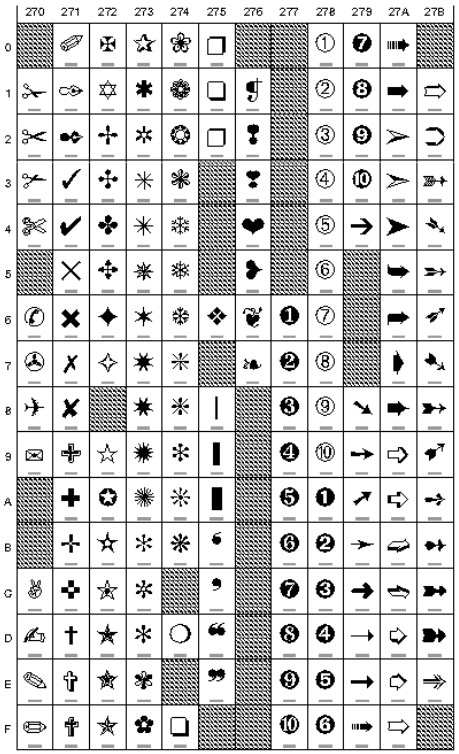


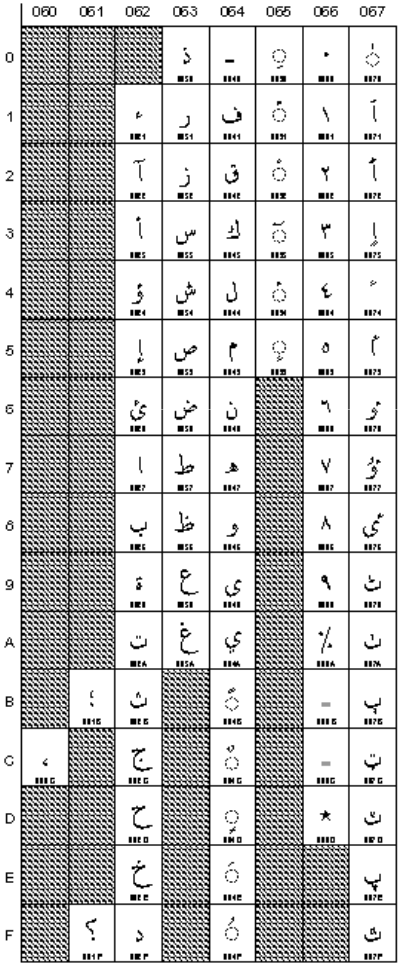
| | 000 | 001 | 002 | 003 | 004 | 005 | 006 | 007 |
|---|---|---|---|---|---|---|---|---|
| 0 | NUL | DLE | SP | 0 | @ | P | ` | p |
| 1 | SOH | DC1 | ! | 1 | A | Q | a | q |
| 2 | STX | DC2 | " | 2 | B | R | b | r |
| 3 | ETX | DC3 | # | 3 | C | S | c | s |
| 4 | EOT | DC4 | $ | 4 | D | T | d | t |
| 5 | ENQ | NAK | % | 5 | E | U | e | u |
| 6 | ACK | SYN | & | 6 | F | V | f | v |
| 7 | BEL | ETB | ' | 7 | G | W | g | w |
| 8 | BS | CAN | ( | 8 | H | X | h | x |
| 9 | HT | EM | ) | 9 | I | Y | i | y |
| A | LF | SUB | * | : | J | Z | j | z |
| B | VT | ESC | + | ; | K | [ | k | { |
| C | FF | FS | , | < | L | \ | l | \| |
| D | CR | GS | - | = | M | ] | m | } |
| E | SO | RS | . | > | N | ^ | n | ~ |
| F | SI | US | / | ? | O | _ | o | DEL |

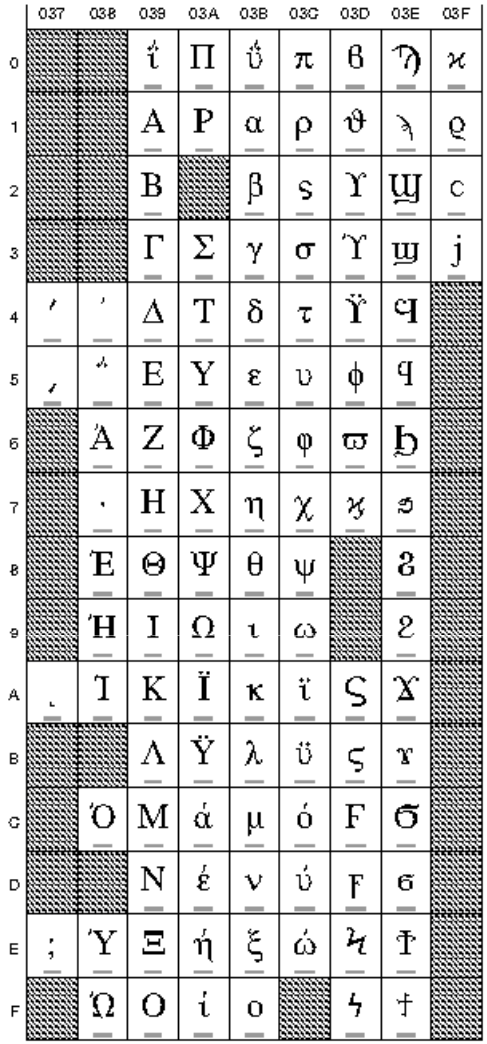
| | 006 | 009 | 00A | 00B | 00C | 00D | 00E | 00F |
|---|---|---|---|---|---|---|---|---|
| 0 | XXX | DCS | NB SP | ° | À | Ð | à | ð |
| 1 | XXX | PU1 | ¡ | ± | Á | Ñ | á | ñ |
| 2 | BPH | PU2 | ¢ | ² | Â | Ò | â | ò |
| 3 | NBH | STS | £ | ³ | Ã | Ó | ã | ó |
| 4 | XXX | CCH | ¤ | ´ | Ä | Ô | ä | ô |
| 5 | NEL | MW | ¥ | µ | Å | Õ | å | õ |
| 6 | SSA | SPA | ¦ | ¶ | Æ | Ö | æ | ö |
| 7 | ESA | EPA | § | · | Ç | × | ç | ÷ |
| 8 | HTS | SOS | ¨ | ¸ | È | Ø | è | ø |
| 9 | HTJ | XXX | © | ¹ | É | Ù | é | ù |
| A | VTS | SCI | ª | º | Ê | Ú | ê | ú |
| B | PLD | CSI | « | » | Ë | Û | ë | û |
| C | PLU | ST | ¬ | ¼ | Ì | Ü | ì | ü |
| D | RI | OSC | SHY | ½ | Í | Ý | í | ý |
| E | SS2 | PM | ® | ¾ | Î | Þ | î | þ |
| F | SS3 | APC | ¯ | ¿ | Ï | ß | ï | ÿ |

# Weitere Zeichensätze

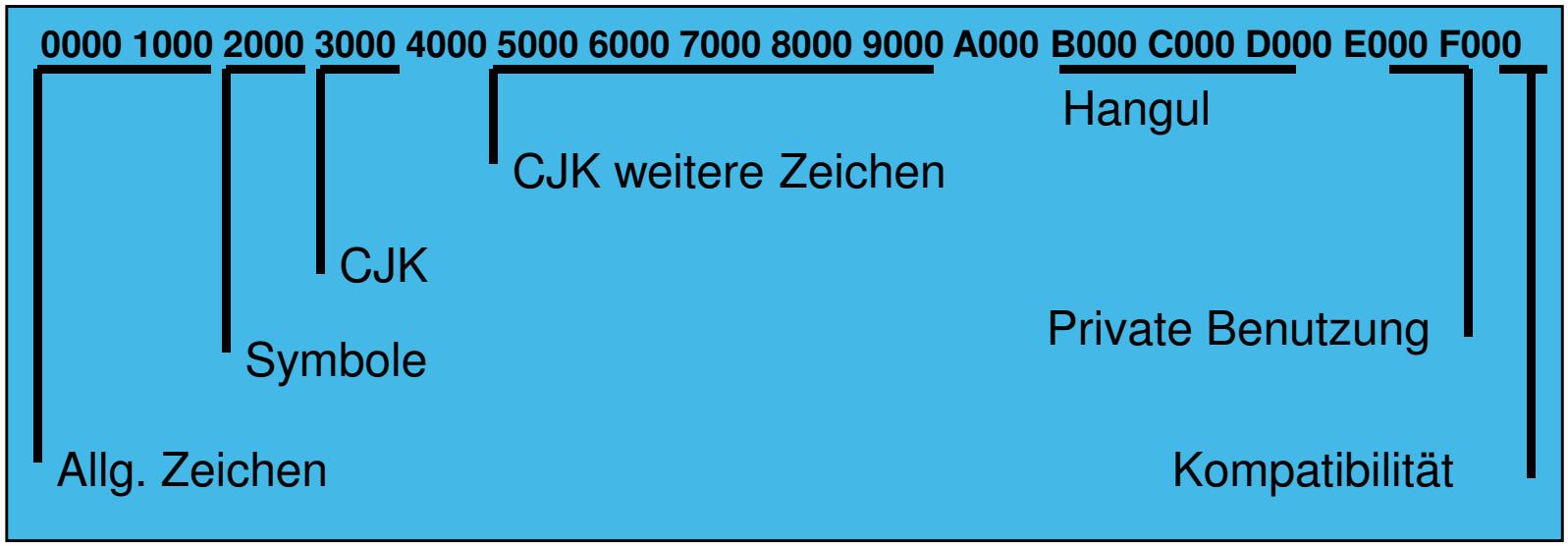
Unicode – ist ein allgemeines System zur Definition von Zeichensätzen







# The Unicode Spectrum




- Unicode == ISO 10646
- 38,886 16-bit characters (20,902 CJK)
- Every character ever available with a computer
- 1 million "surrogate" characters

# Unicode for Programmers

- 16-bit formats: UTF-16 and UCS-2 (wchar_t in C, char in Java)
- 8-bit format: UTF-8 (char in C)
- Go to www.unicode.org and buy the book!

# Unicode and XML

```
<?xml version="1.0" encoding="ISO-8859-1" ?>
```

- XML processors required to read UTF-8 and UTF-16!
- ... but ASCII, EBCDIC, JIS, KO18-R, Big5, etc. are all full of Unicode characters ...
- ... so they are legal XML too ...
- ... but you have to tell the processor!

# Unicode und Encodings

- Unicode in Programmen
  - UCS-2: two-byte characters
  - UCS-4: four-byte characters (future)
- Unicode in Dateien
  - UTF-8: ASCII is ASCII, rest are 1- to 4-bytes
  - UTF-16: two octets per character, initial
- ASCII with numeric character references works, too (© for ©)