

Standards und Werkzeuge Überblick

Josef Schneeberger

Studienstiftung Greifswald 2008

„Kulturerbe digital“

Übersicht

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

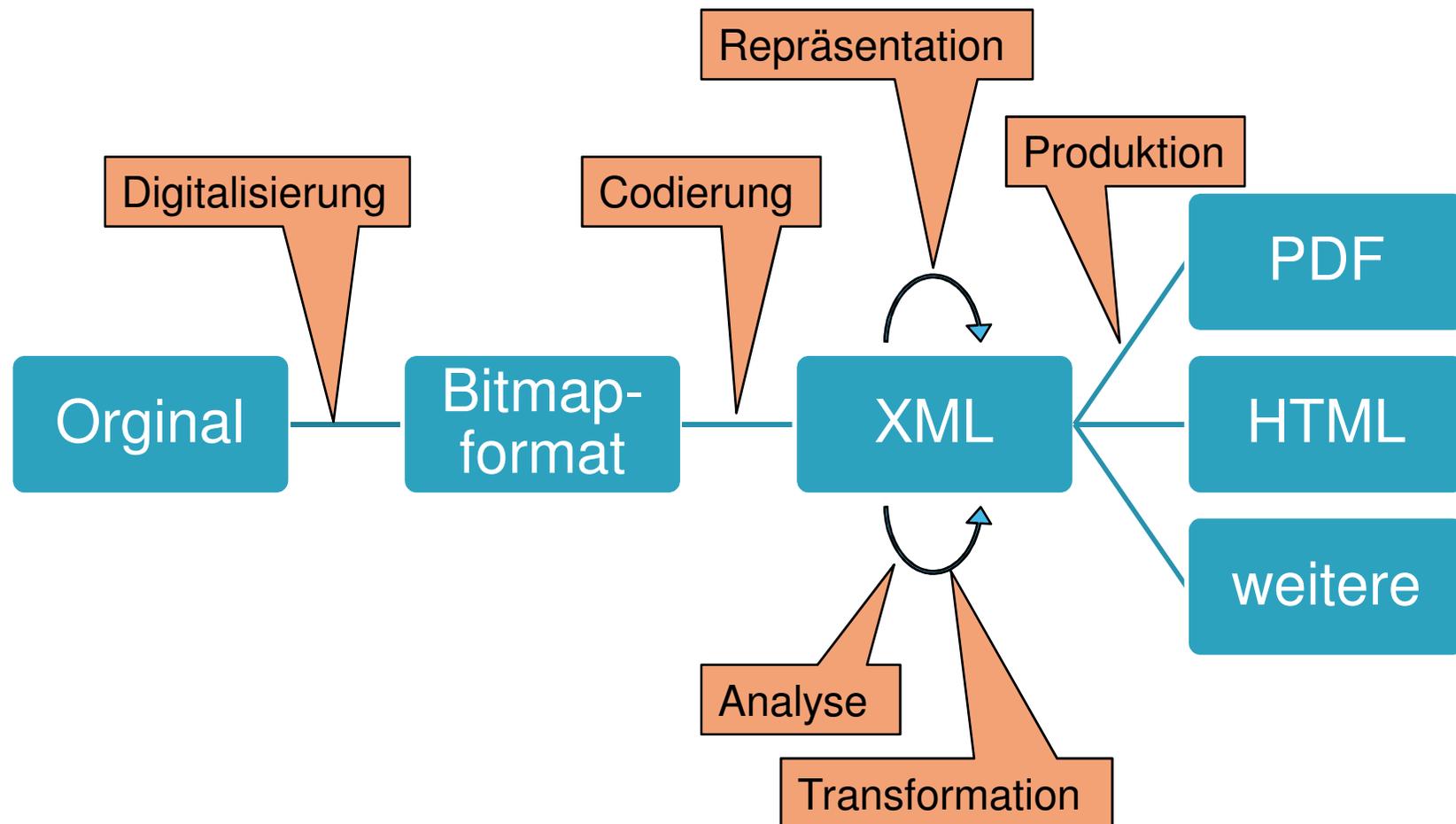
- ∅ Hinweise zur Erfassung des „deutschen Ptolemäus“



Studienstiftung
des deutschen Volkes

Arbeitsplan

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





Studienstiftung
des deutschen Volkes

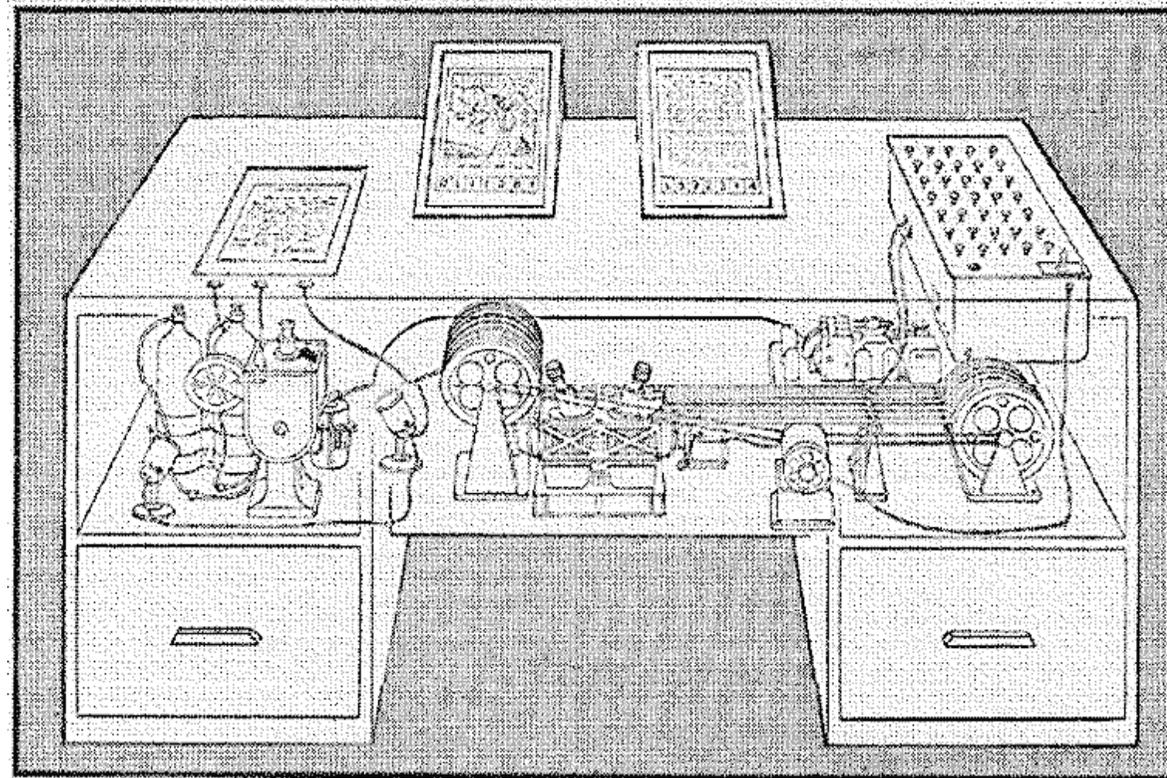
Digitale Dokumente

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?

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 supermicro-film 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

Kurze Geschichte: Hypertext

- n **1965** *Ted Nelson* prägt den Begriff *Hypertext* und entwirft das System *Xanadu* bezeichnet er als ("a magic place of literary memory").
- n **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".
- n **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.
- n **1982** Auslieferung des Dokumentationssystems *Document Examiner* (für sog. "Lisp-Maschinen") durch die Firma *Symbolics* als erstes Hypertext-System für Endanwender.
- n **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*).
- n **1988** Durch das System *HyperCard* von Apple erreicht das Hypermedia-Konzept auch die Benutzer von Personal Computern

Komponenten von Dokumenten

∅ Text (Absätze, Aufzählungen, Überschriften)

∅ Bilder

∅ Zeichnungen / Grafiken

∅ Tabellen

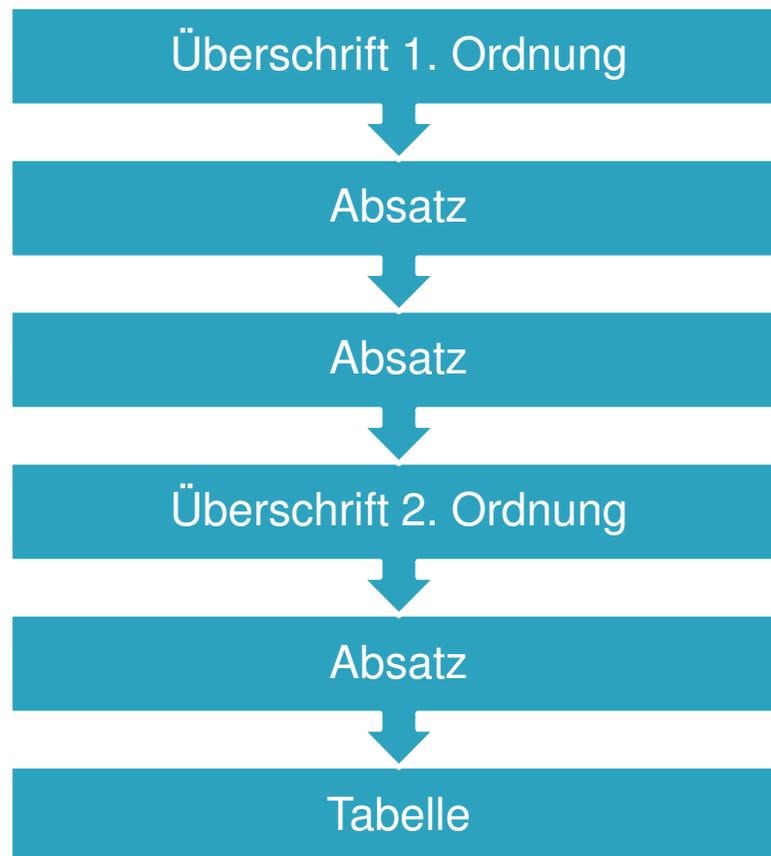
∅ Verweise

∅ Formeln

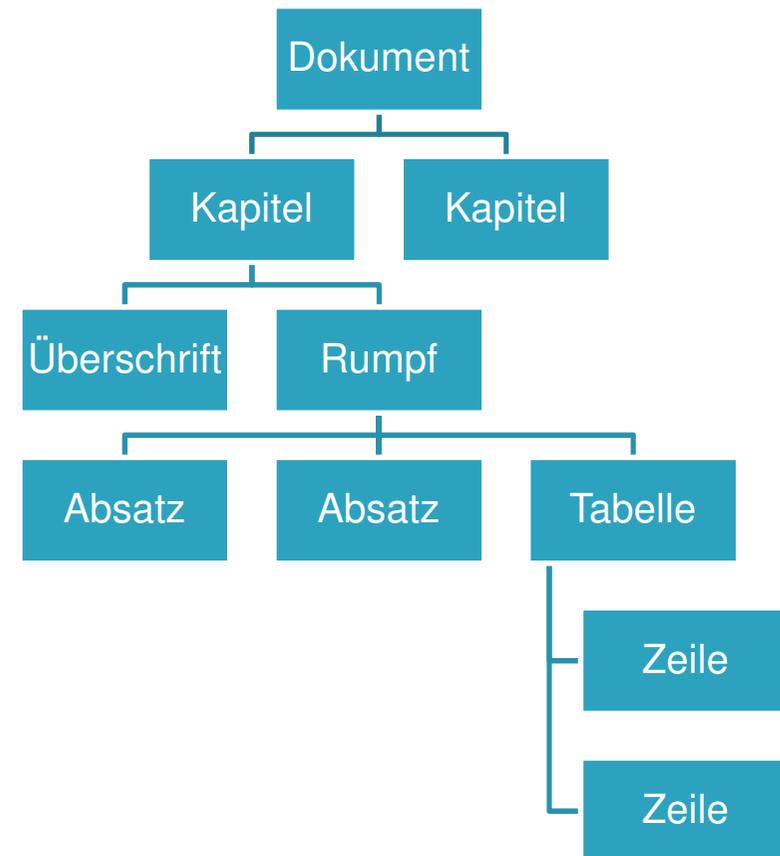
... das war's gibt es sonst noch etwas?

Zwei Repräsentationsansätze

∅ Eine lineare Sequenz von Absätzen (Word, Framemaker)

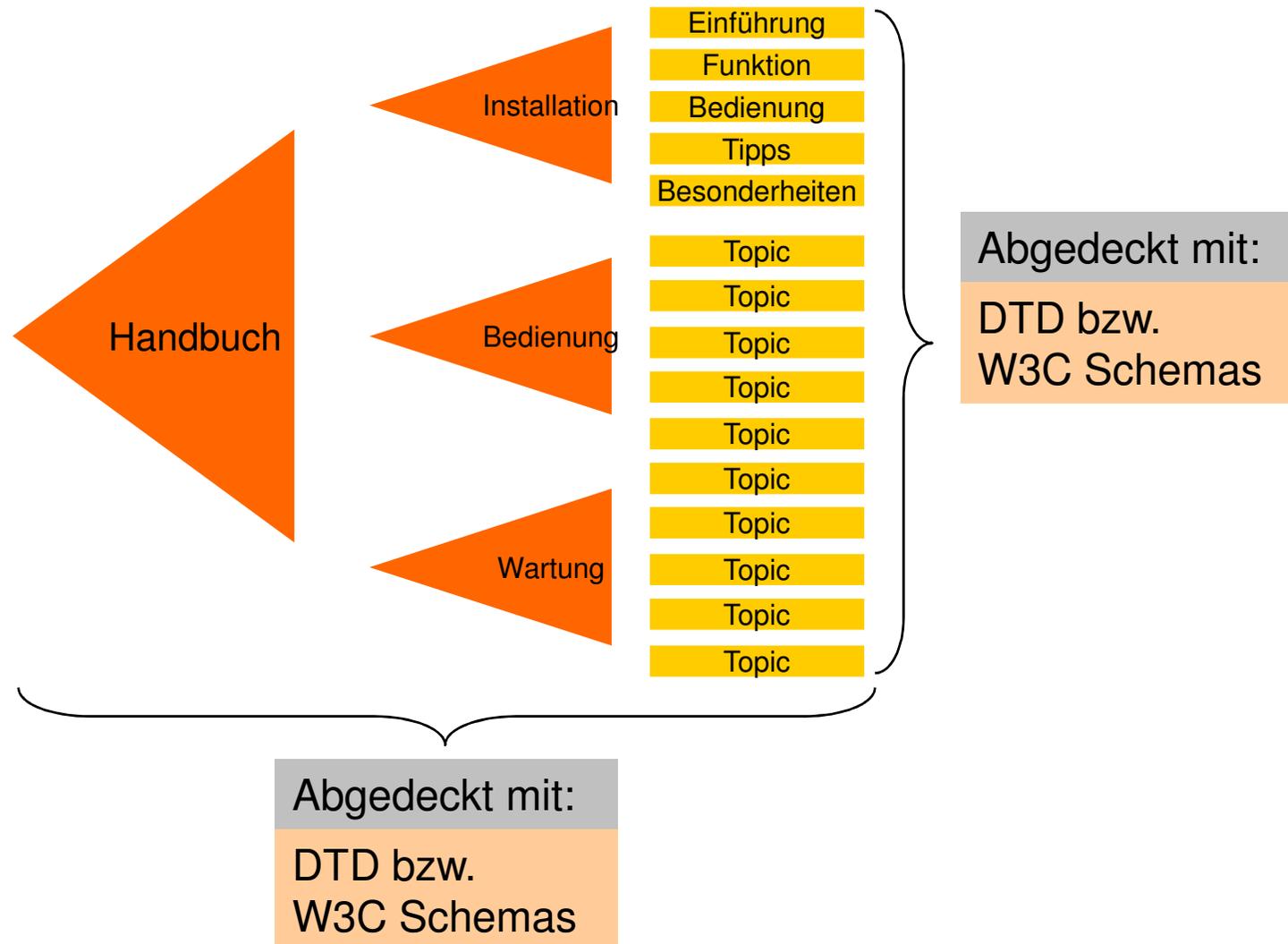


∅ Ein Baum

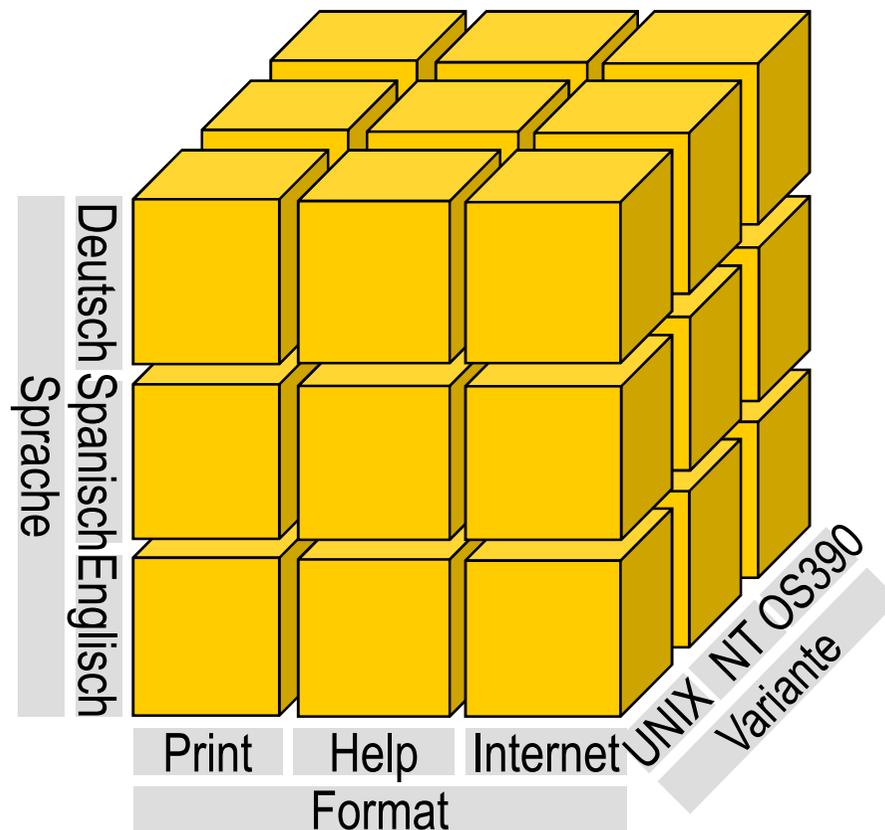


Dimensionen eines digitalen Dokuments

State of the Art: XML und DTDs



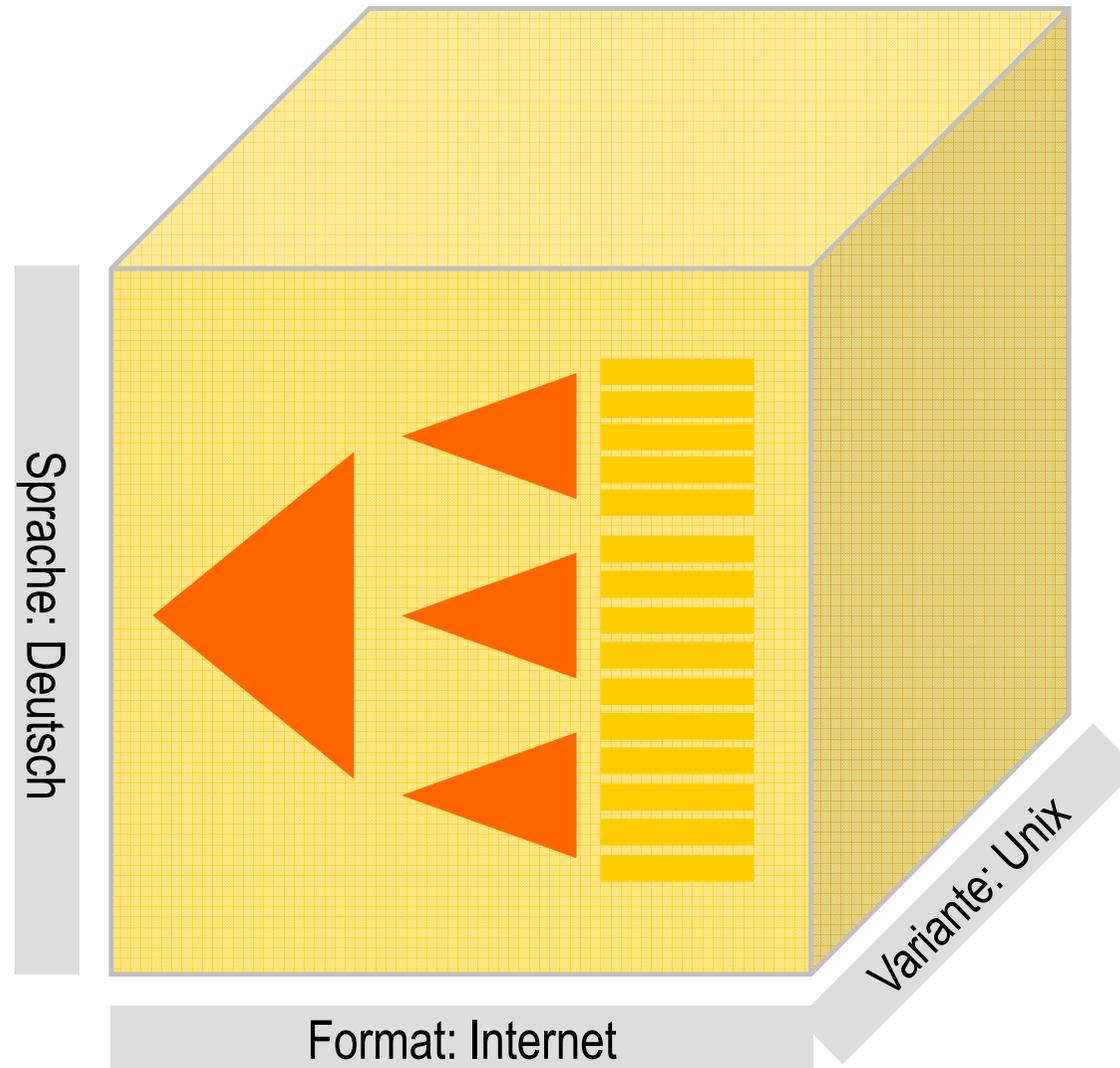
Informationsdimensionen



Nicht Abgedeckt
mit XML:

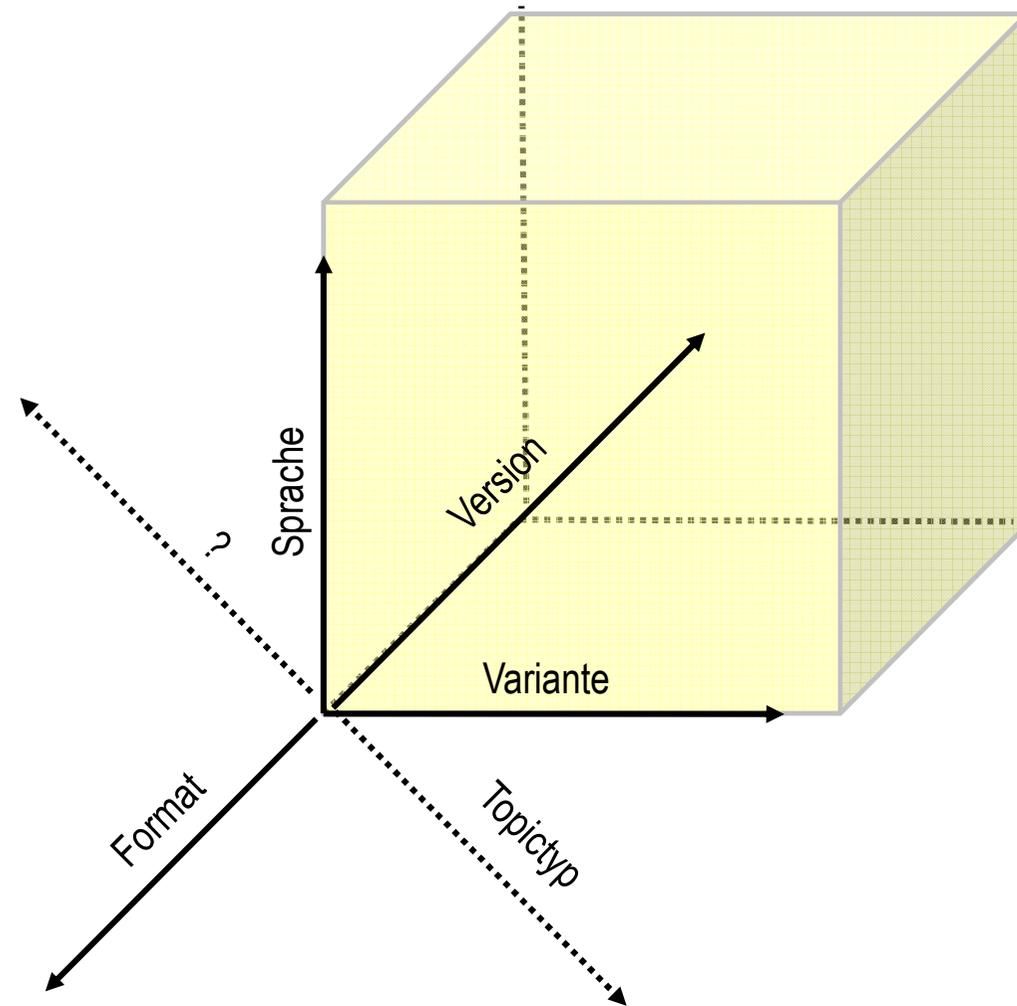
Dimensionen
und Aspekten

Eine Strukturbaum existiert in mehreren Dimensionen!

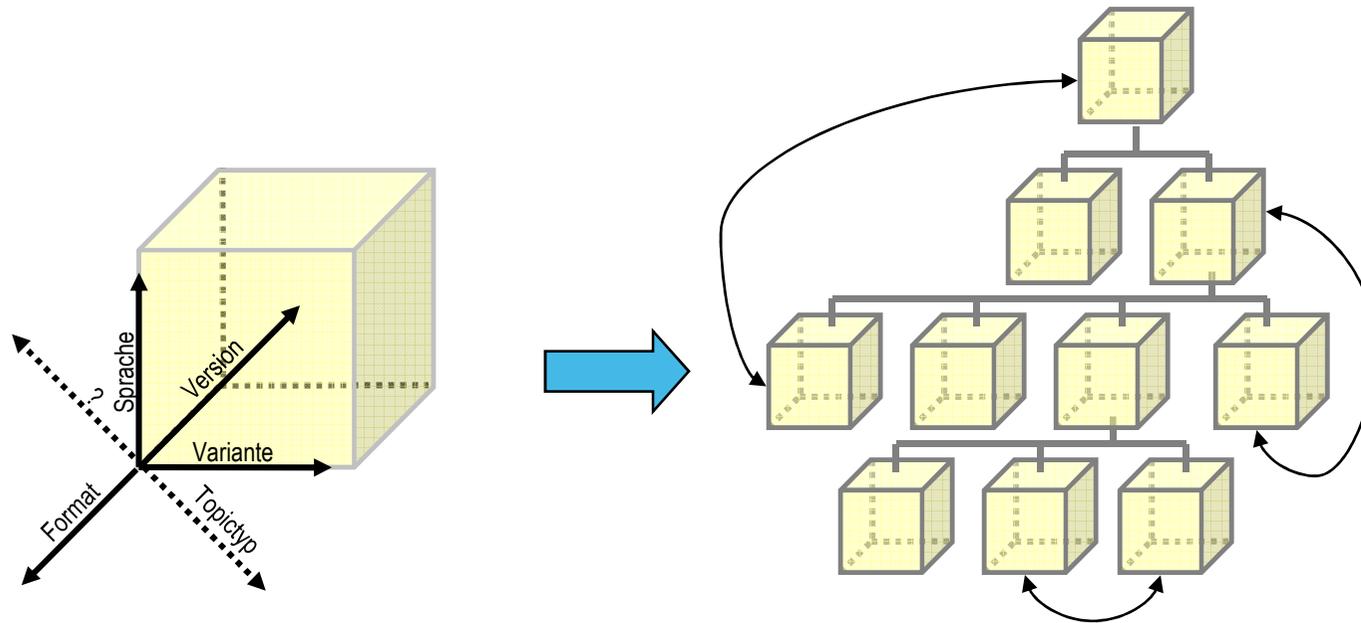


Was sind somit die eigentlichen Probleme?

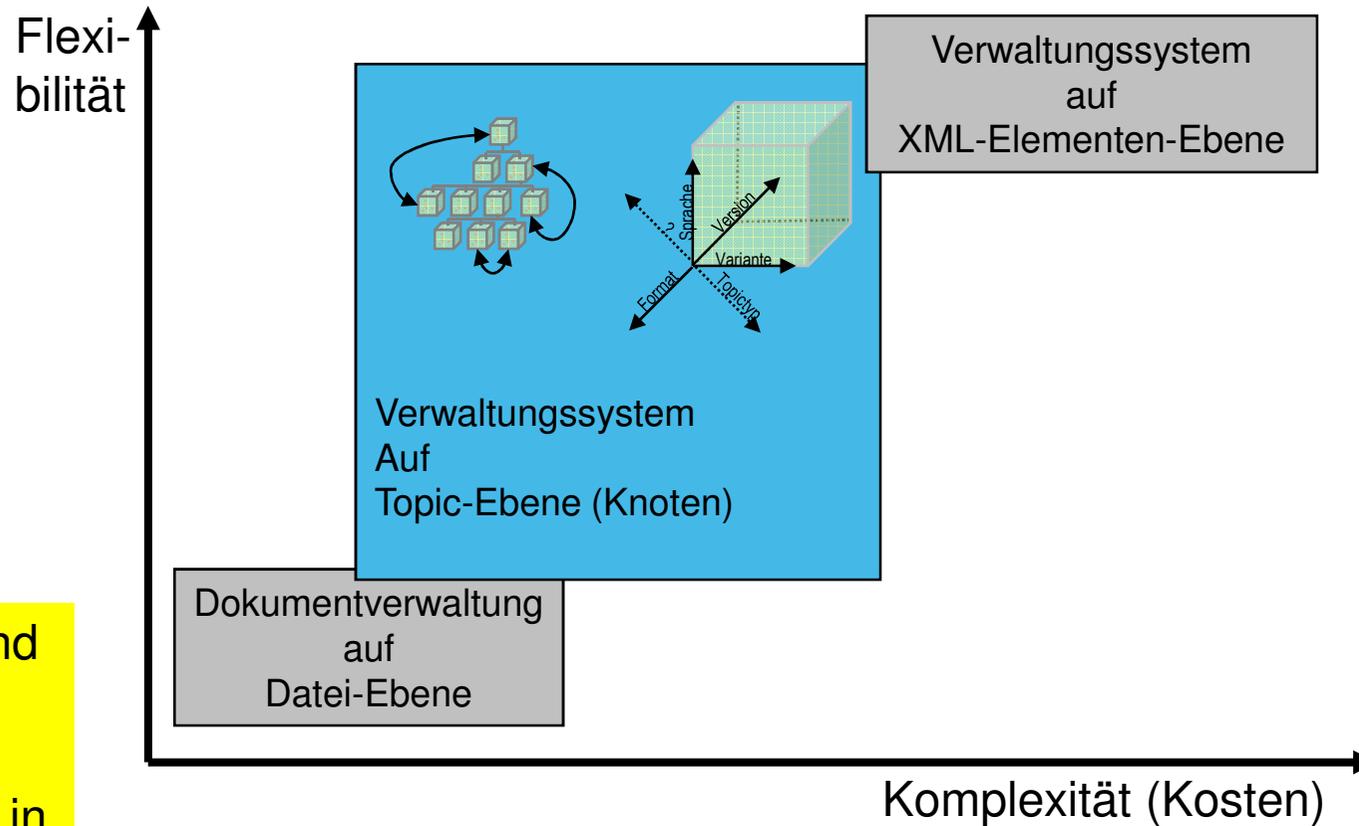
Wie behält man
die Übersicht
über ein solches
Dokument?



Dimensionen und Hierarchien



Wozu Knoten bzw. Topics?



Warum und wie speichert man XML in einer Datenbank?

Funktionsblöcke im CMS

Workflow	Metadaten für Status-Informationen, Aufgabenverwaltung als zentrale Steuerung, Event für kundenspezifische Workflow-Ereignisse
Publishing	Produktion von dynamischen Websites für Redaktionsanwendungen 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?

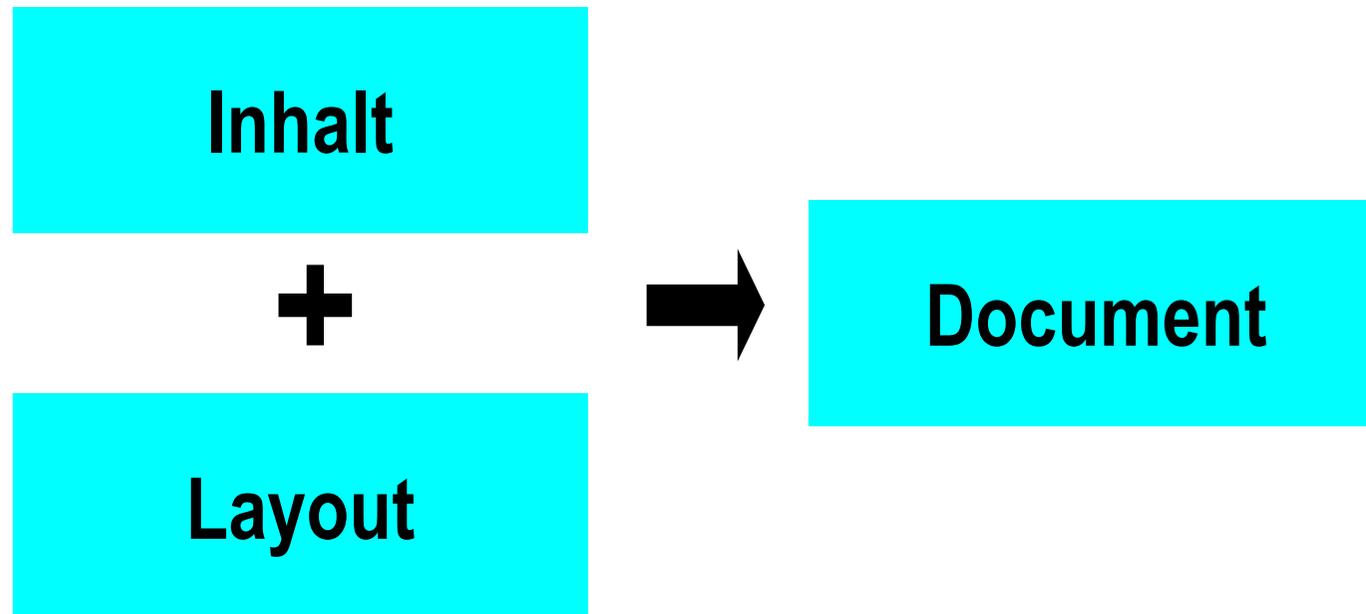
XML - Was ist das?

- ∅ 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

Philosophie von XML / SGML



XML steht für strukturierte Daten

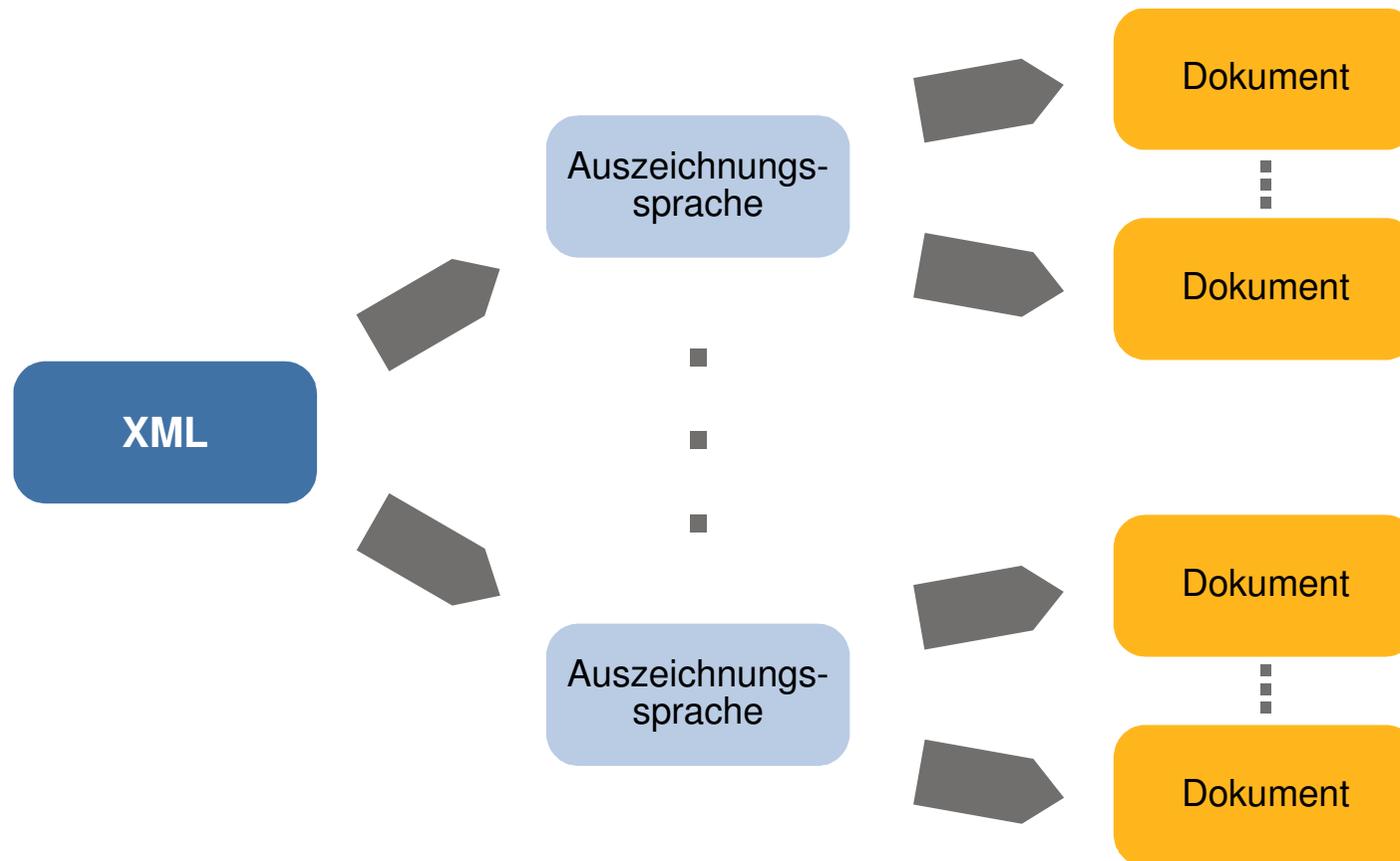
- ∅ Strukturierte Daten können beliebige Objekte enthalten
- ∅ XML ist keine Programmiersprache (aber eine formale Sprache)
- ∅ XML dient zur leichten Verarbeitung beliebiger Daten auf dem Computer
- ∅ XML ist
 - ∅ Erweiterbar
 - ∅ Plattformenabhängig
 - ∅ International (Unicode)

Aus [Mi02]

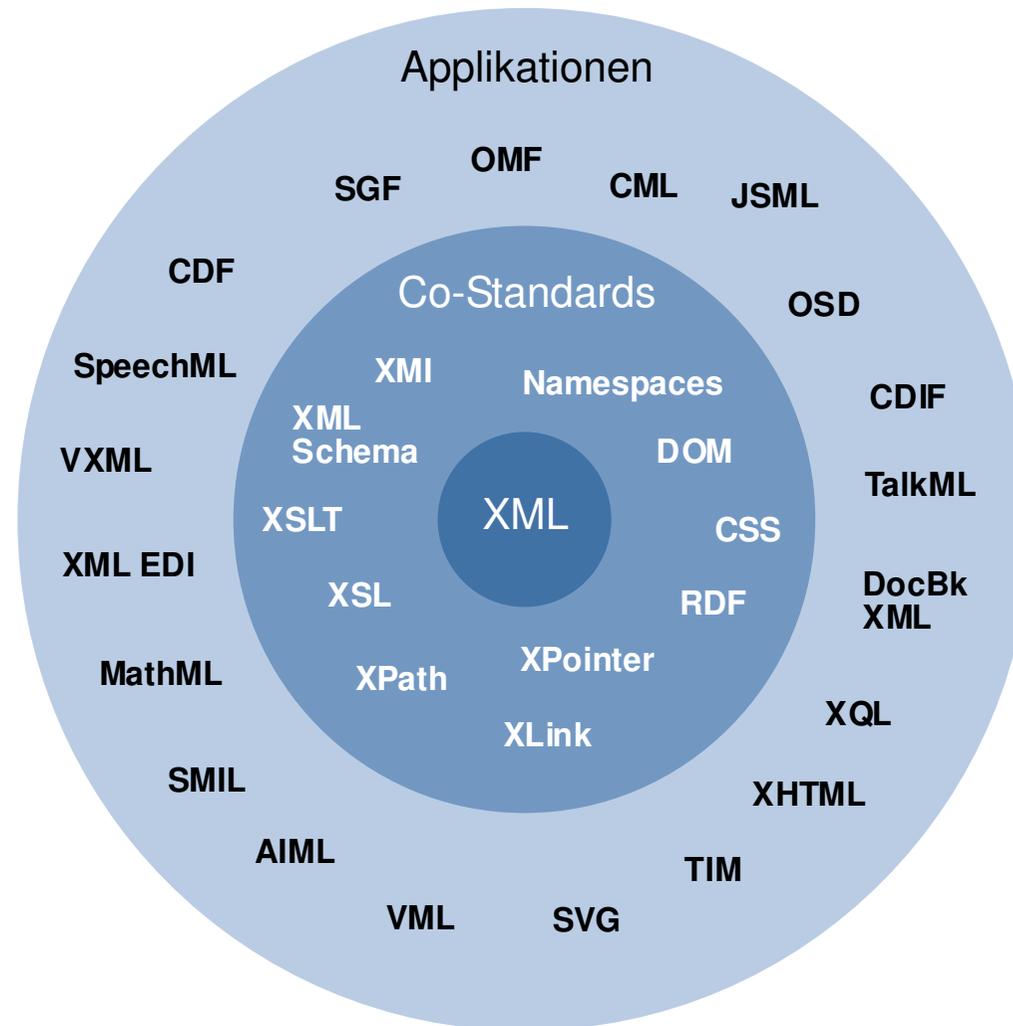
Es gibt (sehr) viele Werkzeuge für XML

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

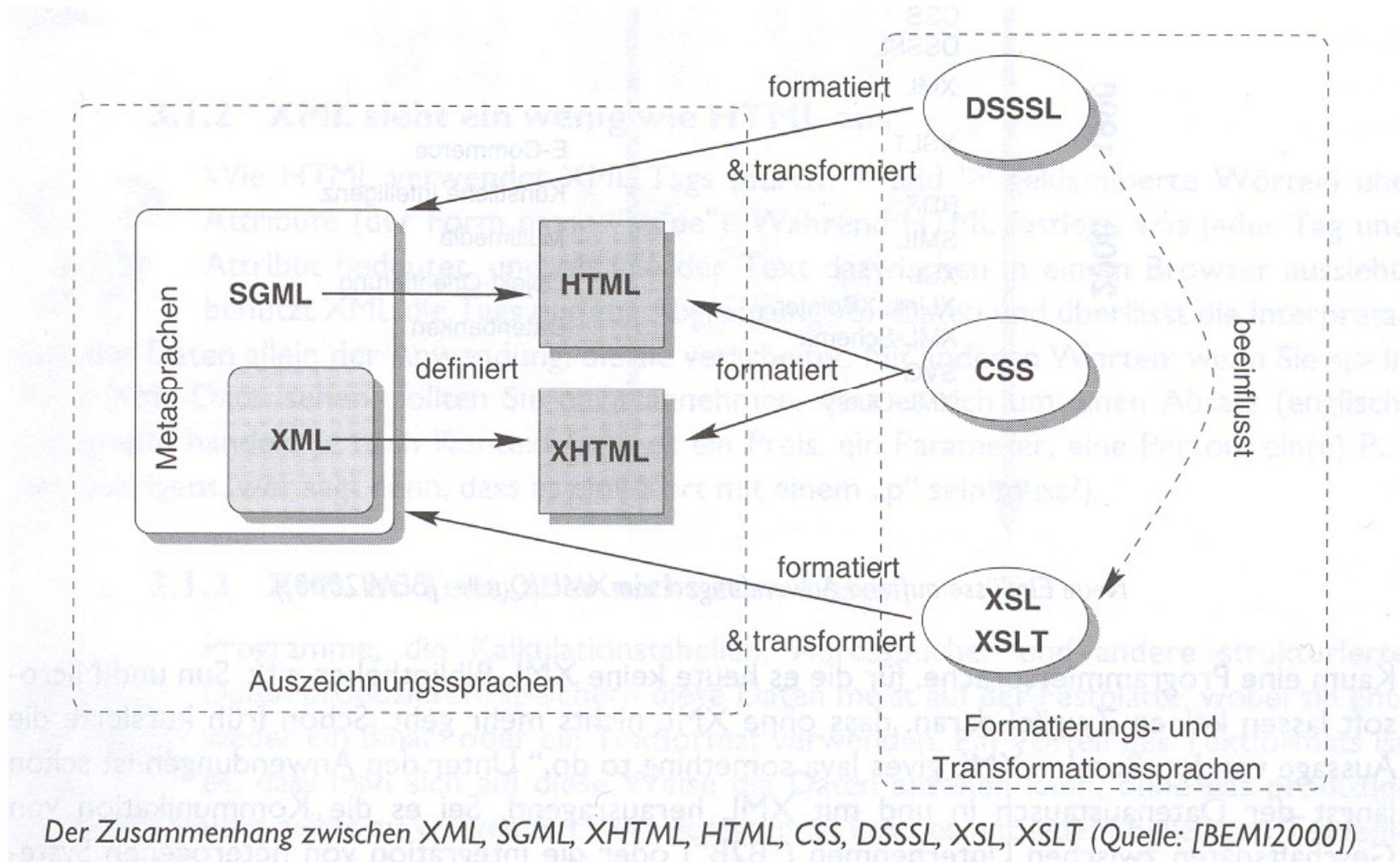
Auszeichnungssprachen (3)



Ein Blick in das XML-Universum



XML SGML XHTML HTML CSS DSSSL XSL XSLT



Aus [BeMi00]

Diskussion

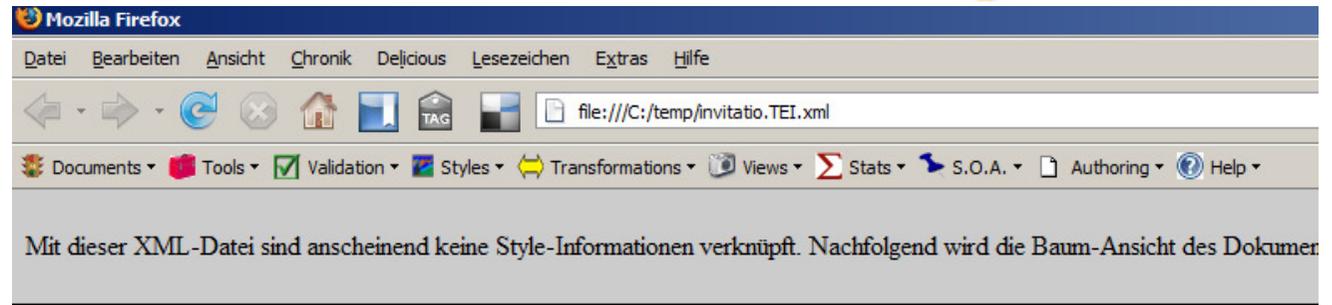
- ∅ Die Auszeichnung von Dokumenten ermöglicht die flexible Verwendungen von Dokumenten:
 - ∅ Import / Export mit Datenbanken
 - ∅ Wissensrepräsentation
 - ∅ Automatische Extraktion und Verknüpfung
- ∅ Probleme mit XML :
 - ∅ Wenn das Schema/DTD geändert wird, dann müssen alle existierenden Dokumente 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

Browser – Darstellung von XML

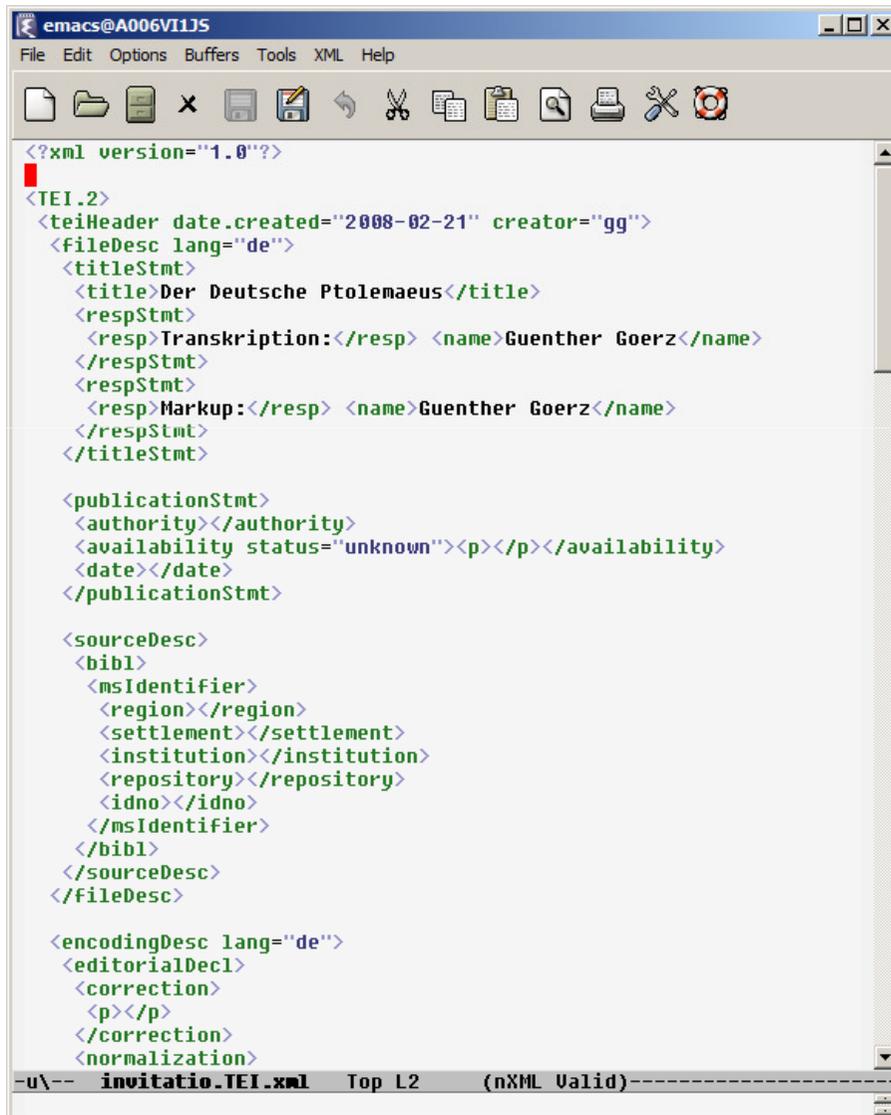


```

- <TEI.2>
  - <teiHeader date.created="2008-02-21" creator="gg">
    - <fileDesc lang="de">
      - <titleStm>
        <title>Der Deutsche Ptolemaeus</title>
      - <respStm>
        <resp>Transkription:</resp>
        <name>Guenther Goerz</name>
      </respStm>
      + <respStm></respStm>
    </titleStm>
    + <publicationStm></publicationStm>
    + <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
  </p>
  </div1>
  </body>
  </text>
  </group>
  </text>
  </fileDesc>
  </teiHeader>
  </TEI.2>

```

Editoren für XML (Texteditor, Emacs, etc.)



```
emacs@A006VI1J5
File Edit Options Buffers Tools XML Help
[Icons]
<?xml version="1.0"?>
<TEI.2>
  <teiHeader date.created="2008-02-21" creator="gg">
    <fileDesc lang="de">
      <titleStmnt>
        <title>Der Deutsche Ptolemaeus</title>
        <respStmnt>
          <resp>Transkription:</resp> <name>Guenther Goerz</name>
        </respStmnt>
        <respStmnt>
          <resp>Markup:</resp> <name>Guenther Goerz</name>
        </respStmnt>
      </titleStmnt>

      <publicationStmnt>
        <authority></authority>
        <availability status="unknown"><p></p></availability>
        <date></date>
      </publicationStmnt>

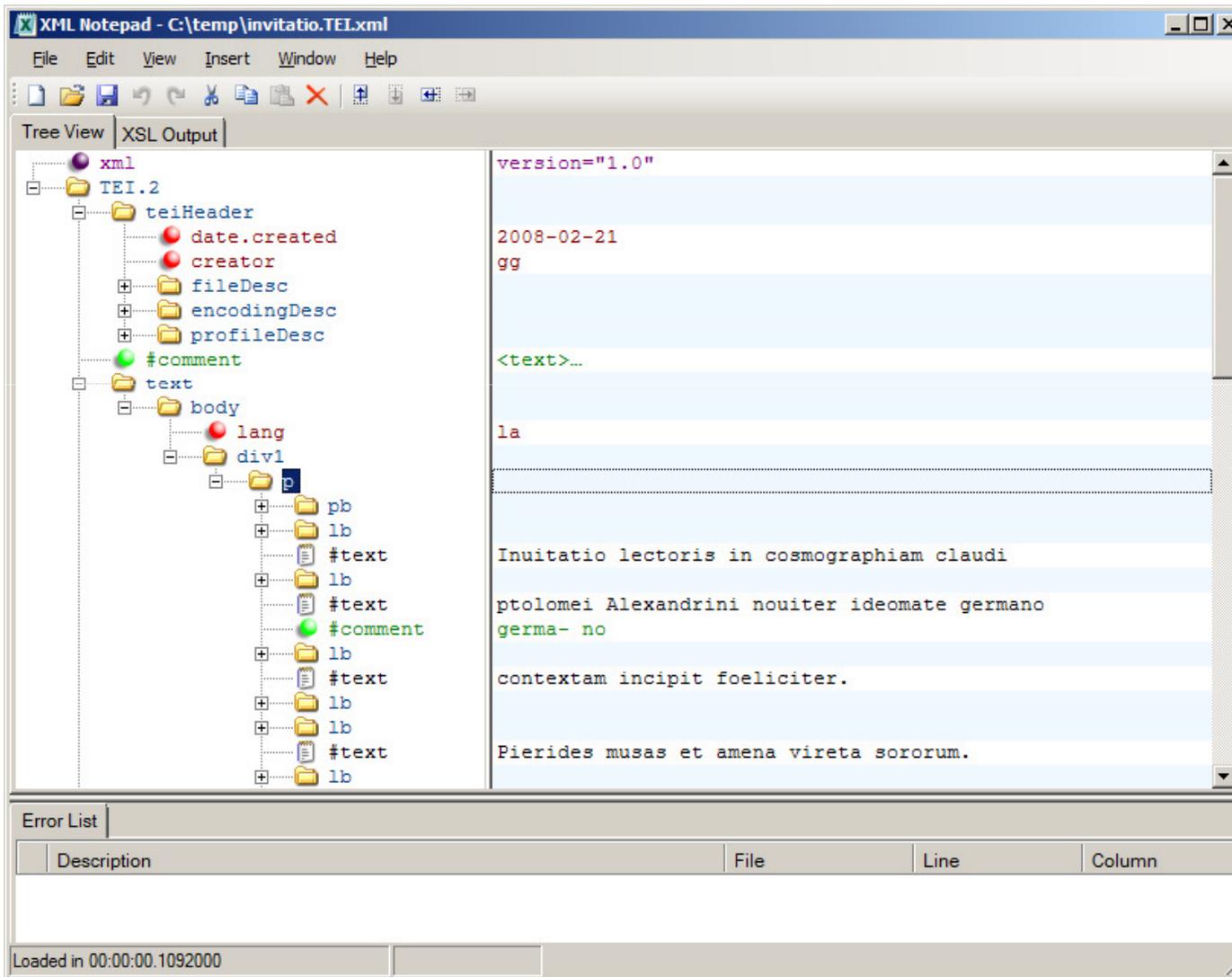
      <sourceDesc>
        <bibl>
          <msIdentifier>
            <region></region>
            <settlement></settlement>
            <institution></institution>
            <repository></repository>
            <idno></idno>
          </msIdentifier>
        </bibl>
      </sourceDesc>
    </fileDesc>

    <encodingDesc lang="de">
      <editorialDecl>
        <correction>
          <p></p>
        </correction>
        <normalization>

```

-u\-- **invitatio.TEI.xml** Top L2 (nXML Valid)

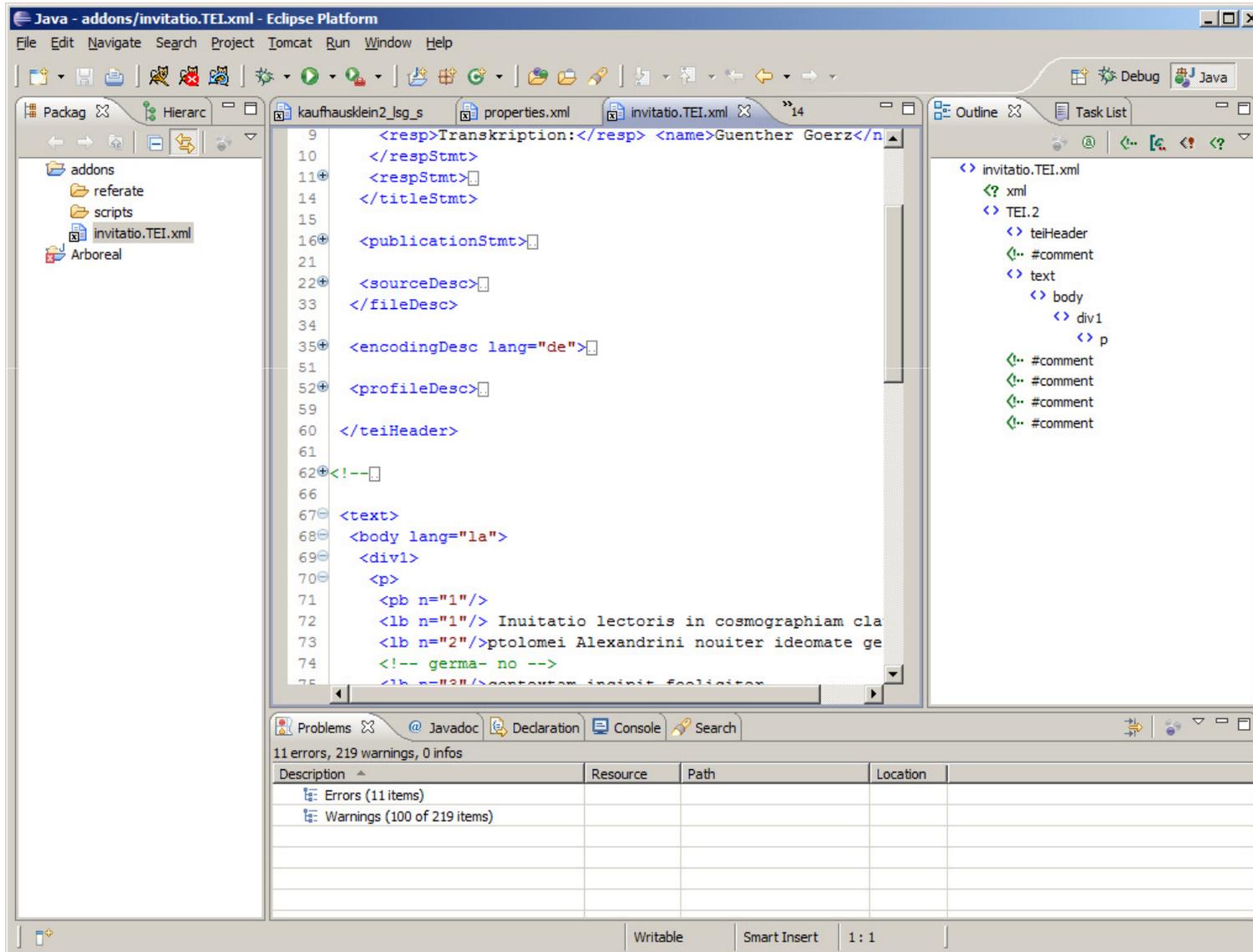
Editoren für XML (MS XML Notepad)



The screenshot shows the MS XML Notepad interface with the following components:

- Tree View:** A hierarchical tree structure of the XML document. The root is `xml`, followed by `TEI.2`. Under `TEI.2` are `teiHeader` (containing `date.created`, `creator`, `fileDesc`, `encodingDesc`, and `profileDesc`), `#comment`, and `text`. The `text` element contains `body`, which includes `lang` (set to `la`) and `div1`. `div1` contains a `pb` element, followed by several `lb` (line break) elements and `#text` elements containing Latin text.
- XSL Output:** The right pane displays the XML content in a light blue background. It shows the `version="1.0"` attribute, the `date.created` and `creator` values, the `<text>...` tag, the `la` language attribute, and the Latin text: `Inuitatio lectoris in cosmographiam claudi ptolomei Alexandrini nouiter ideomate germano germa- no contextam incipit foeliciter. Pierides musas et amena vireta sororum.`
- Error List:** A table at the bottom with columns for Description, File, Line, and Column. It is currently empty.
- Status Bar:** Shows "Loaded in 00:00:00.1092000".

Editoren für XML (Eclipse – mit Schema Unterstützung)



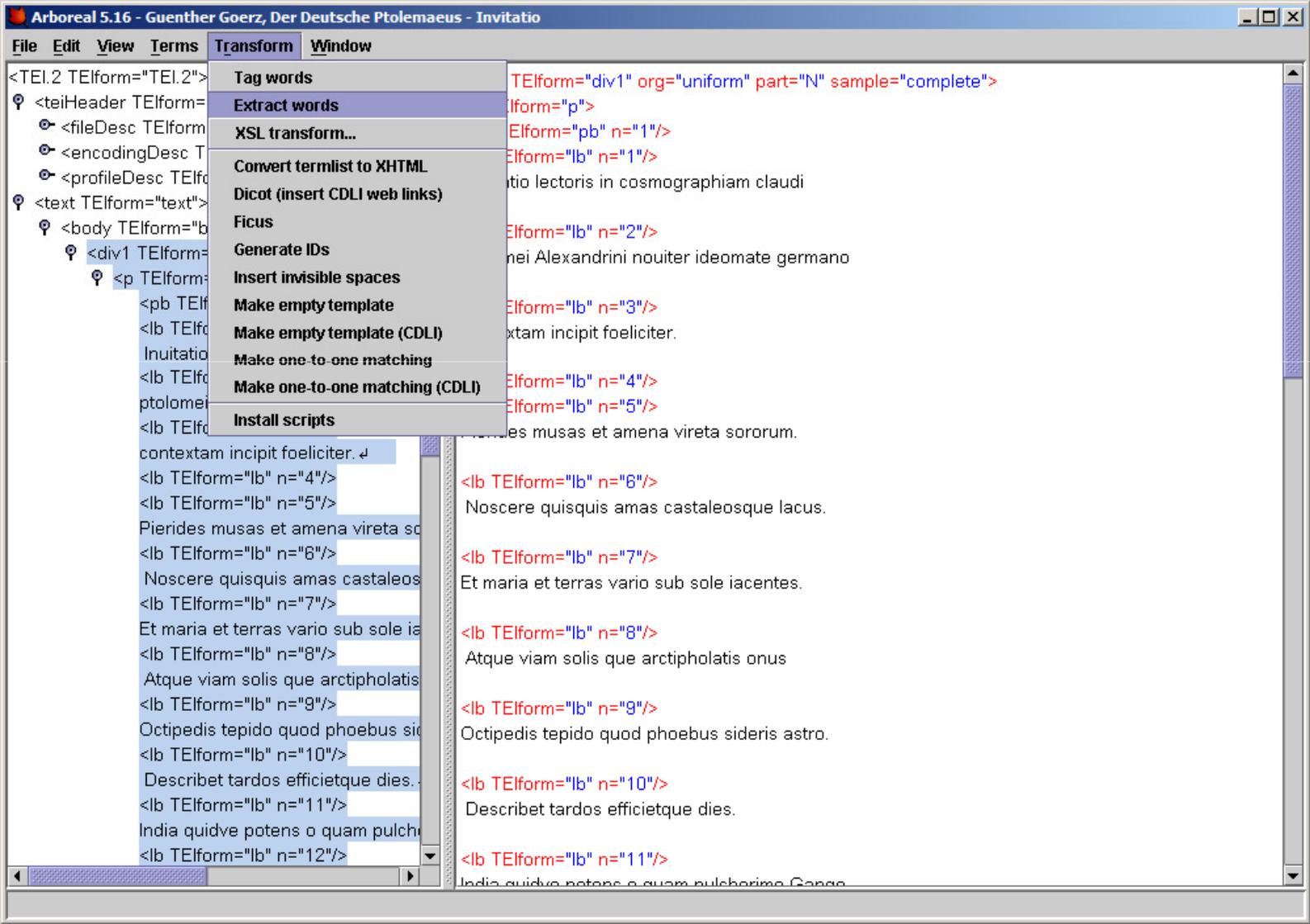
The screenshot shows the Eclipse IDE interface with the following components:

- Left Panel (Package Explorer):** Shows a project structure with folders 'referate', 'scripts', and 'Arboreal', and a file 'invitatio.TEI.xml'.
- Central Editor:** Displays the XML code for 'invitatio.TEI.xml'. The code includes elements like `<resp>`, `<respStmt>`, `<titleStmt>`, `<publicationStmt>`, `<sourceDesc>`, `<fileDesc>`, `<encodingDesc lang="de">`, `<profileDesc>`, `</teiHeader>`, `<!-->`, `<text>`, `<body lang="la">`, `<div1>`, `<p>`, `<pb n="1"/>`, `<lb n="1"/>`, `<lb n="2"/>`, and `<!-- germa- no -->`.
- Right Panel (Outline):** Shows a tree view of the XML document structure, including 'xml', 'TEI.2', 'teiHeader', '#comment', 'text', 'body', 'div1', and 'p'.
- Bottom Panel (Problems):** Displays a table with 11 errors and 219 warnings. The table has columns for Description, Resource, Path, and Location.

Description	Resource	Path	Location
Errors (11 items)			
Warnings (100 of 219 items)			

XML Editieren / Verarbeiten (Oxygen)

Verarbeitung / Transformation (Arboreal)



The screenshot shows the Arboreal 5.16 application window. The title bar reads "Arboreal 5.16 - Guenther Goerz, Der Deutsche Ptolemaeus - Invitatio". The menu bar includes "File", "Edit", "View", "Terms", "Transform", and "Window". The "Transform" menu is open, displaying the following options:

- Tag words
- Extract words
- XSL transform...
- Convert termlist to XHTML
- Dicot (insert CDLI web links)
- Ficus
- Generate IDs
- Insert invisible spaces
- Make empty template
- Make empty template (CDLI)
- Make one-to-one matching
- Make one-to-one matching (CDLI)
- Install scripts

The main window displays XML code on the left and the resulting HTML output on the right. The XML code includes tags like `<teiHeader TEIform=`, `<fileDesc TEIform=`, `<encodingDesc T`, `<profileDesc TEIfo`, `<text TEIform="text">`, `<body TEIform="b`, `<div1 TEIform=`, `<p TEIform=`, `<pb TEI`, `<lb TEIfo`, `Inuitatio`, `<lb TEIfo`, `ptolomei`, `<lb TEIfo`, `contextam incipit foeliciter.`, `<lb TEIform="lb" n="4"/>`, `<lb TEIform="lb" n="5"/>`, `Pierides musas et amena vireta sc`, `<lb TEIform="lb" n="6"/>`, `Noscere quisquis amas castaleos`, `<lb TEIform="lb" n="7"/>`, `Et maria et terras vario sub sole ia`, `<lb TEIform="lb" n="8"/>`, `Atque viam solis que arctipholatis`, `<lb TEIform="lb" n="9"/>`, `Octipedis tepido quod phoebus sid`, `<lb TEIform="lb" n="10"/>`, `Describet tardos efficietque dies.`, `<lb TEIform="lb" n="11"/>`, `India quidve potens o quam pulch`, `<lb TEIform="lb" n="12"/>`.

The HTML output on the right shows the rendered text with corresponding `<lb TEIform="lb" n="1"/>` through `<lb TEIform="lb" n="11"/>` tags, along with the text: `TEIform="div1" org="uniform" part="N" sample="complete">`, `form="p">`, `form="pb" n="1"/>`, `form="lb" n="1"/>`, `atio lectoris in cosmographiam claudi`, `form="lb" n="2"/>`, `nei Alexandrini nouiter ideomate germano`, `form="lb" n="3"/>`, `xtam incipit foeliciter.`, `form="lb" n="4"/>`, `form="lb" n="5"/>`, `es musas et amena vireta sororum.`, `Noscere quisquis amas castaleosque lacus.`, `Et maria et terras vario sub sole iacentes.`, `Atque viam solis que arctipholatis onus`, `Octipedis tepido quod phoebus sideris astro.`, `Describet tardos efficietque dies.`, `India quidve potens o quam pulcherime Gange`.

Kooperative Entwicklung von Inhalten

∅ Verwaltung von Versionen

∅ Wiki – werden wir evtl. einrichten

Arbeit mit dem Text

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>

Basiselemente

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

Optionale Elemente

Tag	Wofür
<code><rs ...> <name></code>	Namen, Orte
<code><date ...> <time ...></code>	Zeitpunkte

Diskussion

Unicode

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] <small>000</small>	[DLE] <small>001</small>	[SP] <small>002</small>	0 <small>003</small>	@ <small>004</small>	P <small>005</small>	` <small>006</small>	p <small>007</small>
1	[SOH] <small>008</small>	[DC1] <small>009</small>	! <small>010</small>	1 <small>011</small>	A <small>012</small>	Q <small>013</small>	a <small>014</small>	q <small>015</small>
2	[STX] <small>016</small>	[DC2] <small>017</small>	" <small>018</small>	2 <small>019</small>	B <small>020</small>	R <small>021</small>	b <small>022</small>	r <small>023</small>
3	[ETX] <small>024</small>	[DC3] <small>025</small>	# <small>026</small>	3 <small>027</small>	C <small>028</small>	S <small>029</small>	c <small>030</small>	s <small>031</small>
4	[EDT] <small>032</small>	[DC4] <small>033</small>	\$ <small>034</small>	4 <small>035</small>	D <small>036</small>	T <small>037</small>	d <small>038</small>	t <small>039</small>
5	[ENQ] <small>040</small>	[NAK] <small>041</small>	% <small>042</small>	5 <small>043</small>	E <small>044</small>	U <small>045</small>	e <small>046</small>	u <small>047</small>
6	[ACK] <small>048</small>	[SYN] <small>049</small>	& <small>050</small>	6 <small>051</small>	F <small>052</small>	V <small>053</small>	f <small>054</small>	v <small>055</small>
7	[BEL] <small>056</small>	[ETB] <small>057</small>	' <small>058</small>	7 <small>059</small>	G <small>060</small>	W <small>061</small>	g <small>062</small>	w <small>063</small>
8	[BS] <small>064</small>	[CAN] <small>065</small>	(<small>066</small>	8 <small>067</small>	H <small>068</small>	X <small>069</small>	h <small>070</small>	x <small>071</small>
9	[HT] <small>072</small>	[EM] <small>073</small>) <small>074</small>	9 <small>075</small>	I <small>076</small>	Y <small>077</small>	i <small>078</small>	y <small>079</small>
A	[LF] <small>080</small>	[SUB] <small>081</small>	* <small>082</small>	: <small>083</small>	J <small>084</small>	Z <small>085</small>	j <small>086</small>	z <small>087</small>
B	[VT] <small>088</small>	[ESC] <small>089</small>	+ <small>090</small>	; <small>091</small>	K <small>092</small>	[<small>093</small>	k <small>094</small>	{ <small>095</small>
C	[FF] <small>096</small>	[FS] <small>097</small>	, <small>098</small>	< <small>099</small>	L <small>100</small>	\ <small>101</small>	l <small>102</small>	<small>103</small>
D	[CR] <small>104</small>	[GS] <small>105</small>	- <small>106</small>	= <small>107</small>	M <small>108</small>] <small>109</small>	m <small>110</small>	} <small>111</small>
E	[SO] <small>112</small>	[RS] <small>113</small>	. <small>114</small>	> <small>115</small>	N <small>116</small>	^ <small>117</small>	n <small>118</small>	~ <small>119</small>
F	[SI] <small>120</small>	[US] <small>121</small>	/ <small>122</small>	? <small>123</small>	O <small>124</small>	_ <small>125</small>	o <small>126</small>	[DEL] <small>127</small>



	008	009	00A	00B	00C	00D	00E	00F
0	[XXX] <small>008</small>	[DC5] <small>009</small>	[NB SP] <small>00A</small>	◊ <small>00B</small>	À <small>00C</small>	Ā <small>00D</small>	à <small>00E</small>	ā <small>00F</small>
1	[XXX] <small>010</small>	[PU1] <small>011</small>	¡ <small>012</small>	± <small>013</small>	Á <small>014</small>	Ñ <small>015</small>	á <small>016</small>	ñ <small>017</small>
2	[BPH] <small>018</small>	[PU2] <small>019</small>	¢ <small>020</small>	² <small>021</small>	Â <small>022</small>	Ò <small>023</small>	â <small>024</small>	ò <small>025</small>
3	[NBH] <small>026</small>	[STS] <small>027</small>	£ <small>028</small>	³ <small>029</small>	Ã <small>030</small>	Ó <small>031</small>	ã <small>032</small>	ó <small>033</small>
4	[XXX] <small>034</small>	[CCH] <small>035</small>	¤ <small>036</small>	´ <small>037</small>	Ä <small>038</small>	Ô <small>039</small>	ä <small>040</small>	ô <small>041</small>
5	[NEL] <small>042</small>	[MW] <small>043</small>	¥ <small>044</small>	µ <small>045</small>	Å <small>046</small>	Õ <small>047</small>	å <small>048</small>	õ <small>049</small>
6	[SSA] <small>050</small>	[SPA] <small>051</small>	¦ <small>052</small>	¶ <small>053</small>	Æ <small>054</small>	Ö <small>055</small>	æ <small>056</small>	ö <small>057</small>
7	[ESA] <small>058</small>	[EPA] <small>059</small>	§ <small>060</small>	• <small>061</small>	Ç <small>062</small>	× <small>063</small>	ç <small>064</small>	÷ <small>065</small>
8	[HTS] <small>066</small>	[SOS] <small>067</small>	** <small>068</small>	ˆ <small>069</small>	È <small>070</small>	Ø <small>071</small>	è <small>072</small>	ø <small>073</small>
9	[HTJ] <small>074</small>	[XXX] <small>075</small>	© <small>076</small>	¹ <small>077</small>	É <small>078</small>	Ù <small>079</small>	é <small>080</small>	ù <small>081</small>
A	[VTS] <small>082</small>	[SCI] <small>083</small>	ª <small>084</small>	º <small>085</small>	Ê <small>086</small>	Ú <small>087</small>	ê <small>088</small>	ú <small>089</small>
B	[PLD] <small>090</small>	[OSI] <small>091</small>	« <small>092</small>	» <small>093</small>	Ë <small>094</small>	Û <small>095</small>	ë <small>096</small>	û <small>097</small>
C	[PLU] <small>098</small>	[ST] <small>099</small>	¬ <small>100</small>	¼ <small>101</small>	Ì <small>102</small>	Ü <small>103</small>	ì <small>104</small>	ü <small>105</small>
D	[RI] <small>106</small>	[OSC] <small>107</small>	[SHV] <small>108</small>	½ <small>109</small>	Í <small>110</small>	Ý <small>111</small>	í <small>112</small>	ý <small>113</small>
E	[SS2] <small>114</small>	[PM] <small>115</small>	® <small>116</small>	¾ <small>117</small>	Î <small>118</small>	Ë <small>119</small>	î <small>120</small>	ë <small>121</small>
F	[SS3] <small>122</small>	[APC] <small>123</small>	– <small>124</small>	¿ <small>125</small>	Ï <small>126</small>	ß <small>127</small>	ï <small>128</small>	ÿ <small>129</small>

Weitere Zeichensätze

Unicode – ist ein
allgemeines System
zur Definition von
Zeichensätzen

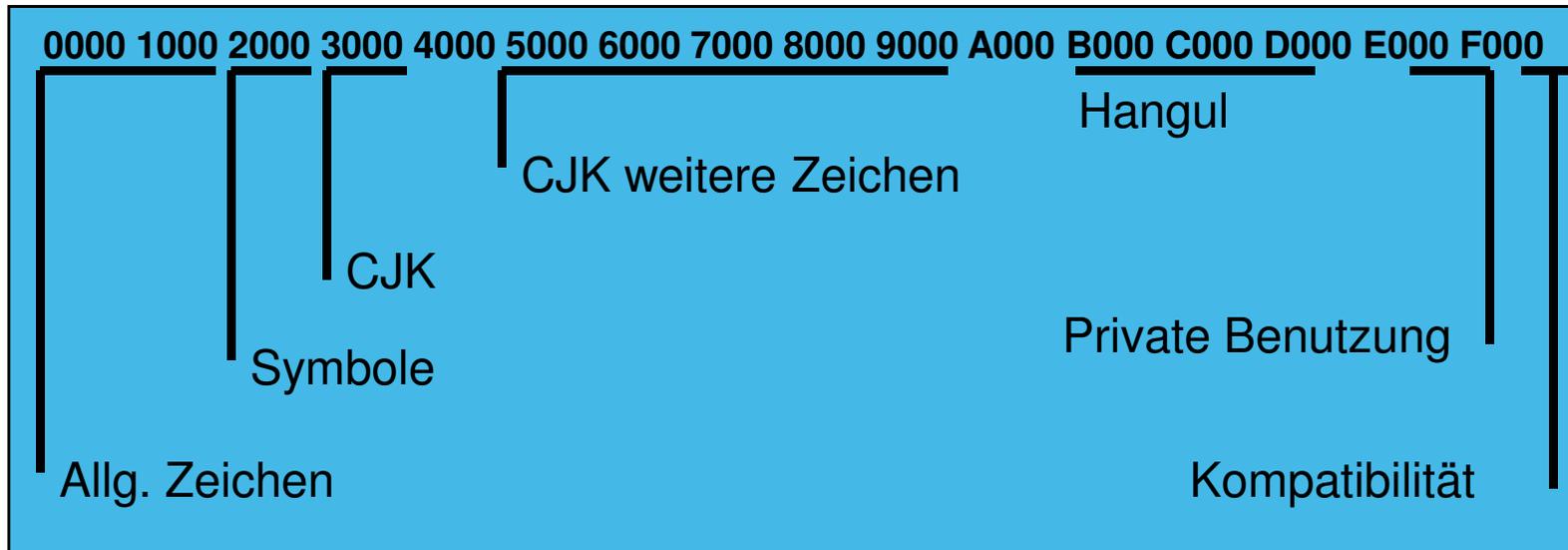
	270	271	272	273	274	275	276	277	278	279	27A	27B
0	☞	☛	☛	☛	☛	☛	☛	☛	①	②	➡	➡
1	✂	☞	☛	☛	☛	☛	☛	☛	③	④	➡	➡
2	✂	☞	☛	☛	☛	☛	☛	☛	⑤	⑥	➡	➡
3	✂	☞	☛	☛	☛	☛	☛	☛	⑦	⑧	➡	➡
4	✂	☞	☛	☛	☛	☛	☛	☛	⑨	⑩	➡	➡
5	✂	☞	☛	☛	☛	☛	☛	☛	⑪	⑫	➡	➡
6	☞	☛	☛	☛	☛	☛	☛	☛	⑬	⑭	➡	➡
7	☞	☛	☛	☛	☛	☛	☛	☛	⑮	⑯	➡	➡
8	✂	☞	☛	☛	☛	☛	☛	☛	⑰	⑱	➡	➡
9	☞	☛	☛	☛	☛	☛	☛	☛	⑲	⑳	➡	➡
A	☞	☛	☛	☛	☛	☛	☛	☛	㉑	㉒	➡	➡
B	☞	☛	☛	☛	☛	☛	☛	☛	㉓	㉔	➡	➡
C	☞	☛	☛	☛	☛	☛	☛	☛	㉕	㉖	➡	➡
D	☞	☛	☛	☛	☛	☛	☛	☛	㉗	㉘	➡	➡
E	☞	☛	☛	☛	☛	☛	☛	☛	㉙	㉚	➡	➡
F	☞	☛	☛	☛	☛	☛	☛	☛	㉛	㉜	➡	➡

	060	061	062	063	064	065	066	067
0				ذ	-	و	٠	١
1		ء	ر	ف	و	ا		أ
2		آ	ز	ق	و	٢		أ
3		أ	س	ك	و	٣		أ
4		و	ع	ل	و	٤		أ
5		ا	ص	م	و	٥		أ
6		ي	ن			٦		أ
7		ا	ط	ح		٧		أ
8		ب	ظ	و		٨		أ
9		ة	ع	ي		٩		أ
A		ن	غ	ي		١٠		أ
B		:	ن	و				أ
C		ح		و				أ
D		ح		و		*		أ
E		ح		و				أ
F		؟	د	و				أ

	304	305	306	307	308	309
0	ぐ	だ	ば	む	ゐ	
1	あ	け	ち	ば	め	ゑ
2	あ	げ	ち	ひ	も	を
3	い	こ	っ	び	ゃ	ん
4	い	ご	っ	び	ゃ	づ
5	う	さ	づ	ぶ	ゆ	
6	う	さ	て	ぶ	ゆ	
7	え	し	で	ぶ	よ	
8	え	じ	と	へ	よ	
9	お	す	ど	べ	ら	お
A	お	ず	な	ぺ	り	お
B	か	せ	に	ほ	る	ゝ
C	が	ぜ	ぬ	ぼ	れ	ゝ
D	き	そ	ね	ほ	ろ	ゝ
E	き	ぞ	の	ま	わ	ゝ
F	く	た	は	み	わ	

	037	038	039	03A	03B	03C	03D	03E	03F
0			ı	Π	ϖ	π	ϐ	ϑ	κ
1			A	P	α	ρ	ϑ	λ	ϑ
2			B		β	ς	Υ	Ϟ	ϑ
3			Γ	Σ	γ	σ	Υ	Ϟ	j
4	'	'	Δ	T	δ	τ	ÿ	ϑ	
5	'	^	E	Y	ε	υ	φ	ϑ	
6			A	Z	Φ	ζ	φ	ϑ	
7			·	H	X	η	χ	ϑ	
8			E	Θ	Ψ	θ	ψ		ϑ
9			H	I	Ω	ι	ω		ϑ
A			I	K	İ	κ	ı	S	X
B			Λ	ÿ	λ	ü	ς	τ	
C			O	M	ά	μ	ό	F	σ
D			N	έ	ν	ύ	φ	σ	
E	;	Y	E	ή	ξ	ώ	κ	τ	
F			Ω	O	i	o		4	†

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 ©)