STU FIIT
SLOVAK UNIVERSITY OF TECHNOLOGY IN BRATISLAVA
FACULTY OF INFORMATICS AND INFORMATION TECHNOLOGIES

VGG
Vision & Graphics Group
Institute of Computer Engineering and Applied Informatics

Základné metódy tvorby multimediálneho obsahu

Reprezentácia textu a textové dokumenty

Ing. Peter Kapec, PhD.

ZS 2020-21

Obsah

- Kódovanie textu
- Jednoduchý a štruktúrovaný text
- Značkovacie jazyky
- HTML
- Jazyky opisujúce stránky dokumentov
- Vedecké dokumenty

Kódovanie textu

ASCII

- American National Standards Institute (ANSI)
- 1968: ASCII - American Standard Code for Information Interchange
- 7-bit = 128 znakov
  - Románske / Latinské znaky, malé / veľké, číslice, kontrolné znaky, interpunkčné a iné znaky
- 8-bit = ďalšie znaky 128-255
  - Ne-románske znaky
  - ISO 8859-1 - jazyky západnej Európy
- Pre ne-európske jazyky - irelevantné
  - Hebrejský, čínsky (GB, Big-5), indické (ISCII)….

0	<NUL>	32	<SPC>	64	@	96	`	128	Ä	160	†	192	¿	224	‡
1	<SOH>	33	!	65	A	97	a	129	Å	161	°	193	¡	225	·
2	<STX>	34	"	66	B	98	b	130	Ç	162	¢	194	¬	226	‚
3	<ETX>	35	#	67	C	99	c	131	É	163	£	195	√	227	„
4	<EOT>	36	$	68	D	100	d	132	Ñ	164	§	196	ƒ	228	‰
5	<ENQ>	37	%	69	E	101	e	133	Ö	165	•	197	≈	229	Â
6	<ACK>	38	&	70	F	102	f	134	Ü	166	¶	198	Δ	230	Ê
7	<BEL>	39	'	71	G	103	g	135	á	167	ß	199	«	231	Á
8	<BS>	40	(	72	H	104	h	136	à	168	®	200	»	232	Ë
9	<TAB>	41	)	73	I	105	i	137	â	169	©	201	…	233	È
10	<LF>	42	*	74	J	106	j	138	ä	170	™	202		234	Í
11	<VT>	43	+	75	K	107	k	139	ã	171	´	203	À	235	Î
12	<FF>	44	,	76	L	108	l	140	å	172	¨	204	Ã	236	Ï
13	<CR>	45	-	77	M	109	m	141	ç	173	≠	205	Õ	237	Ì
14	<SO>	46	.	78	N	110	n	142	é	174	Æ	206	Œ	238	Ó
15	<SI>	47	/	79	O	111	o	143	è	175	Ø	207	œ	239	Ô
16	<DLE>	48	0	80	P	112	p	144	ê	176	∞	208	–	240	
17	<DC1>	49	1	81	Q	113	q	145	ë	177	±	209	—	241	Ò
18	<DC2>	50	2	82	R	114	r	146	í	178	≤	210	“	242	Ú
19	<DC3>	51	3	83	S	115	s	147	ì	179	≥	211	”	243	Û
20	<DC4>	52	4	84	T	116	t	148	î	180	¥	212	‘	244	Ù
21	<NAK>	53	5	85	U	117	u	149	ï	181	µ	213	’	245	ı
22	<SYN	54	6	86	V	118	v	150	ñ	182	∂	214	÷	246	ˆ
23	<ETB>	55	7	87	W	119	w	151	ó	183	∑	215	◊	247	˜
24	<CAN>	56	8	88	X	120	x	152	ò	184	∏	216	ÿ	248	¯
25	<EM>	57	9	89	Y	121	y	153	ô	185	π	217	Ÿ	249	˘
26	<SUB>	58	:	90	Z	122	z	154	ö	186	∫	218	/	250	˙
27	<ESC>	59	;	91	[	123	{	155	õ	187	ª	219	€	251	°
28	<FS>	60	<	92	\	124	|	156	ú	188	º	220	‹	252	¸
29	<GS>	61	=	93	]	125	}	157	ù	189	Ω	221	›	253	˝
30	<RS>	62	>	94	^	126	~	158	û	190	æ	222	fi	254	˛
31	<US>	63	?	95	_	127	<DEL>	159	ü	191	ø	223	fl	255	ˇ

Unicode

- 1993 - Unicode Consortium a ISO
- Štandard pre reprezentáciu všetkých jazykov používaných vo svete
  - ~94,000 znakov
- Súčasná práca
  - Podpora historických jazykov (egyptské hieroglyfy, Indo-Európske jazyky)
  - Notácia pre hudbu

Unicode

- Univerzálnosť
  - *Round-trip compatibility:*
  - dokument v existujúcej znakovej sade → unicode dokument
  - unicode dokument → namapovaný späť na pôvodnú znakovú sadu
  - Napr.: ak znak **é** je reprezentovaný jedným kódom, tak aj unicode znak musí byť

Unicode Table

Unicode Table

	00	01	02	03	04	05	06	07	08	09	0A	0B	0C	0D	0E	0F	10	11	12	13	14	15	16	17	18	19	1A	1B	1C	1D	1E	1F	
0000																																	Symbols
0020		!	"	#	$	%	&	'	(	)	*	+	,	-	.	/	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?	Number
0040	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	[	\	]	^	_	Alphabet
0060	`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z	{	|	}	~		
0080	€		‚	ƒ	„	…	†	‡	ˆ	‰	Š	‹	Œ		Ž			‘	’	“	”	•	–	—	˜	™	š	›	œ		ž	Ÿ	
00A0		¡	¢	£	¤	¥	¦	§	¨	©	ª	«	¬		®	¯	°	±	²	³	´	µ	¶	·	¸	¹	º	»	¼	½	¾	¿	
00C0	À	Á	Â	Ã	Ä	Å	Æ	Ç	È	É	Ê	Ë	Ì	Í	Î	Ï	Ð	Ñ	Ò	Ó	Ô	Õ	Ö	×	Ø	Ù	Ú	Û	Ü	Ý	Þ	ß	Latin
00E0	à	á	â	ã	ä	å	æ	ç	è	é	ê	ë	ì	í	î	ï	ð	ñ	ò	ó	ô	õ	ö	÷	ø	ù	ú	û	ü	ý	þ	ÿ	
0100	Ā	ā	Ă	ă	Ą	ą	Ć	ć	Ĉ	ĉ	Ċ	ċ	Č	č	Ď	ď	Đ	đ	Ē	ē	Ĕ	ĕ	Ė	ė	Ę	ę	Ě	ě	Ĝ	ĝ	Ğ	ğ	
0120	Ġ	ġ	Ģ	ģ	Ĥ	ĥ	Ħ	ħ	Ĩ	ĩ	Ī	ī	Ĭ	ĭ	Į	į	İ	ı	Ĳ	ĳ	Ĵ	ĵ	Ķ	ķ	ĸ	Ĺ	ĺ	Ļ	ļ	Ľ	ľ	Ŀ	
0140	ŀ	Ł	ł	Ń	ń	Ņ	ņ	Ň	ň	ŉ	Ŋ	ŋ	Ō	ō	Ŏ	ŏ	Ő	ő	Œ	œ	Ŕ	ŕ	Ŗ	ŗ	Ř	ř	Ś	ś	Ŝ	ŝ	Ş	ş	
0160	Š	š	Ţ	ţ	Ť	ť	Ŧ	ŧ	Ũ	ũ	Ū	ū	Ŭ	ŭ	Ů	ů	Ű	ű	Ų	ų	Ŵ	ŵ	Ŷ	ŷ	Ÿ	Ź	ź	Ż	ż	Ž	ž	ſ	
0180	ƀ	Ɓ	Ƃ	ƃ	Ƅ	ƅ	Ɔ	Ƈ	ƈ	Ɖ	Ɗ	Ƌ	ƌ	ƍ	Ǝ	Ə	Ɛ	Ƒ	ƒ	Ɠ	Ɣ	ƕ	Ɩ	Ɨ	Ƙ	ƙ	ƚ	ƛ	Ɯ	Ɲ	ƞ	Ɵ	
01A0	Ơ	ơ	Ƣ	ƣ	Ƥ	ƥ	Ʀ	Ƨ	ƨ	Σ	ƪ	ƫ	Ƭ	ƭ	Ʈ	Ư	ư	Ʊ	Ʋ	Ƴ	ƴ	Ƶ	ƶ	Ʒ	Ƹ	ƹ	ƺ	ƻ	Ƽ	ƽ	ƾ	ƿ	
01C0	ǀ	ǁ	ǂ	ǃ	DŽ	Dž	dž	LJ	Lj	lj	NJ	Nj	nj	Ǎ	ǎ	Ǐ	ǐ	Ǒ	ǒ	Ǔ	ǔ	Ǖ	ǖ	Ǘ	ǘ	Ǚ	ǚ	Ǜ	ǜ	ǝ	Ǟ	ǟ	
01E0	Ǡ	ǡ	Ǣ	ǣ	Ǥ	ǥ	Ǧ	ǧ	Ǩ	ǩ	Ǫ	ǫ	Ǭ	ǭ	Ǯ	ǯ	ǰ	DZ	Dz	dz	Ǵ	ǵ	Ƕ	Ƿ	Ǹ	ǹ	Ǻ	ǻ	Ǽ	ǽ	Ǿ	ǿ	
0200	Ȁ	ȁ	Ȃ	ȃ	Ȅ	ȅ	Ȇ	ȇ	Ȉ	ȉ	Ȋ	ȋ	Ȍ	ȍ	Ȏ	ȏ	Ȑ	ȑ	Ȓ	ȓ	Ȕ	ȕ	Ȗ	ȗ	Ș	ș	Ț	ț	Ȝ	ȝ	Ȟ	ȟ	
0220	Ƞ	ȡ	Ȣ	ȣ	Ȥ	ȥ	Ȧ	ȧ	Ȩ	ȩ	Ȫ	ȫ	Ȭ	ȭ	Ȯ	ȯ	Ȱ	ȱ	Ȳ	ȳ	ȴ	ȵ	ȶ	ȷ	ȸ	ȹ	Ⱥ	Ȼ	ȼ	Ƚ	Ⱦ	ȿ	
0240	ɀ	Ɂ	ɂ	Ƀ	Ʉ	Ʌ	Ɇ	ɇ	Ɉ	ɉ	Ɋ	ɋ	Ɍ	ɍ	Ɏ	ɏ	ɐ	ɑ	ɒ	ɓ	ɔ	ɕ	ɖ	ɗ	ɘ	ə	ɚ	ɛ	ɜ	ɝ	ɞ	ɟ	

Unicode

- Dve časti (ISO 10646-1 a ISO 10646-2)
- ISO 10646-1 = Basic Multilingual Plane
  - 49,000 znakov
  - Západné jazyky
  - Stredo-východné jazyky: Latin, Greek, Cyrillic, Hebrew, Arabic
  - Chinese, Japanese, and Korean Hangul ideographs
  - Bengali, Thai, Ethiopic ...
  - Brailove písmo, matematické symboly, rôzne tvary
  - ...

Unicode

- 5 zón
  - alphabetic scripts
    - general scripts, symbols, CJK (Chinese-Japanese-Korean) phonetics, symbols
  - ideographic scripts
  - other characters
  - surrogates
  - reserved codes

Kódovanie Unicode znakov

- 32-bit – zatiaľ sa používa prvých 21-bit
- 32 *planes* po 65,536 znakov:
  - *Basic Multilingual Plane* (BMP) – všetky znaky v používaných jazykoch
  - *Supplementary Multilingual Plane* – historické jazyky, matematické znaky, hudobné symboly
  - *Supplementary Ideographic Plane* – historické čínske znaky
  - *Supplementary Special-Purpose Plane* – identifikovanie jazykov, špeciálne protokoly

UTF-8/16/32

- UTF: UCS Transformation Format
  - Akronym pre: UCS is Unicode Character Set
- UTF-32
  - 4byte, UTF-32BE (big-endian), UTF-32LE (little-endian)
  - Špeciálny znak: *byte-order mark (BOM)*
- UTF-16
  - 2byte, UTF-16BE (big-endian), UTF-16LE (little-endian)
    - *Surrogate* znak na adresovanie znakov mimo *BMP*
- UTF-8
  - 1byte, variabilná dĺžka (1 pre ASCII – 4 mimo BMP)

Mnoho ďalších...

- **US-ASCII** (základná angličtina)
- **UTF:** UTF-8/UTF-16/UTF-32 (Unicode, worldwide)
- **ISO štandardy:** ISO-8859-1 (Western Europe), ISO-8859-2 (Central Europe), ISO-8859-3 (Southern Europe), ISO-8859-4 (Northern Europe), ISO-8859-5 (Cyrillic), ISO-8859-6-i (Arabic), ISO-8859-7 (Greek), ISO-8859-8 (Hebrew, visual), ISO-8859-8-i (Hebrew, logical), ISO-8859-9 (Turkish), ISO-8859-10 (Latin 6), ISO-8859-11, (Latin/Thai), ISO-8859-13 (Latin 7, Balic Rim), ISO-8859-14 (Latin 8, Celtic), ISO-8859-15 (Latin 9), ISO-8859-16 (Latin 10), ISO-2022-jp (Japanese, e-mail), ISO-ir-111 (Cyrillic KOI-8)
- **Windows:** Windows-1250 (Central Europe), Windows-1251 (Cyrillic), Windows-1252 (Western Europe), Windows-1253 (Greek), Windows-1254 (Turkish), Windows-1255 (Hebrew), Windows-1256 (Arabic), Windows-1257 (Baltic Rim)
- **Východné jazyky:** EUC-JP (Japanese, Unix), Shift_JIS (Japanese, Win/Mac), EUC-kr (Korean), gb2312 (Chinese, simplified), gb18030 (Chinese, simplified), big5 (Chinese, traditional), Big5-HKSCS (Chinese, Hong Kong), tis-620 (Thai)
- **Iné:** koi8-r (Russian), koi8-u (Ukrainian), Macintosh (MacRoman)

Jednoduchý a štruktúrovaný text

Jednoduchý textový dokument

- Zvyčajne ASCII, dnes už UTF-8
- Rozdielne platformy – iný „nový riadok“
  - Windows: CRLF
  - Unix: LF
  - Mac OS: CR

- Odsadenie
  - Medzera vs tabulátor (šírka tabulátora?)
- Súborový systém – môže rozlišovať malé a veľké písmená (a mnoho iných odlišností)

Štruktúrovaný textový dokument

Nadpis kapitoly 1

Pod-nadpis 1.1

Tu je samotný text pod-kapitoly 1.1, ktorý je tiež odsadený od kraja stránky...

Nadpis kapitoly 2

Pod-nadpis 2.1

Ďalší text pokračuje tu...

Pod-nadpis 2.2

A ďalší text pokračuje tu...

Nadpis kapitoly 3

Táto kapitola nemá pod-kapitoly.

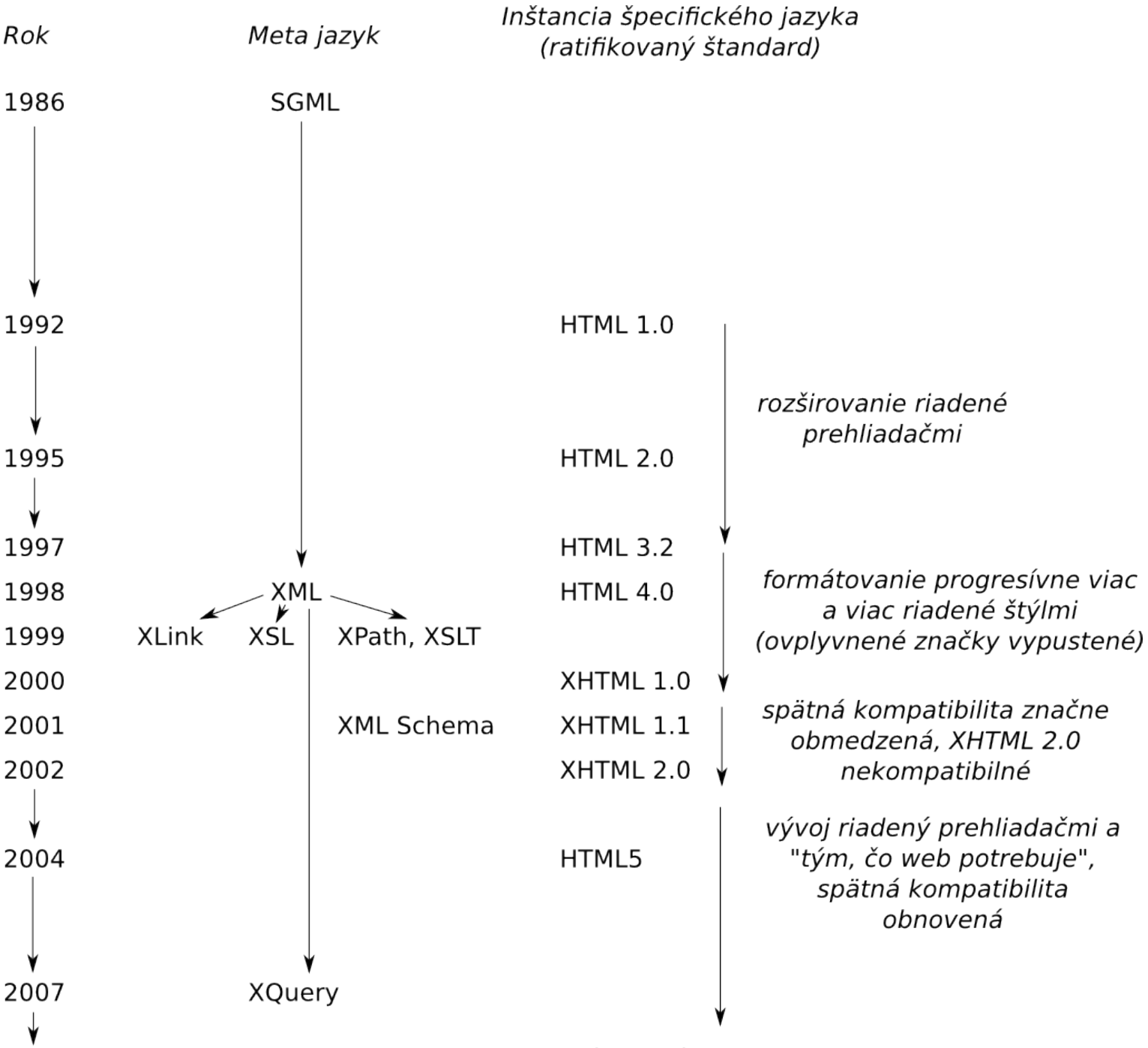
Jednoduché značkovacie jazyky

Markdown:

```
# First-level heading

#### Fourth-level heading

This is a paragraph.

It has two sentences.


This is another paragraph.

It also has  two sentences.


*emphasis* or _emphasis_  (e.g., italics)
```

```
* An item in a bulleted (unordered) list
    * A subitem, indented with 4 spaces
* Another item in a bulleted list


1. An item in an enumerated (ordered) list
2. Another item in an enumerated list


[link to google](http://www.google.com)
```

Značkovacie jazyky


Rok
Meta jazyk
Inštancia špecifického jazyka
(ratifikovaný štandard)
1986
SGML
1992
HTML 1.0
rozširovanie riadené
prehliadačmi
1995
HTML 2.0
1997
HTML 3.2
1998
XML
HTML 4.0
formátovanie progresívne viac
a viac riadené štýlmi
(ovplyvnené značky vypustené)
1999
XLink
XSL
XPath, XSLT
2000
XHTML 1.0
2001
XML Schema
XHTML 1.1
spätná kompatibilita značne
obmedzená, XHTML 2.0
nekompatibilné
2002
XHTML 2.0
2004
HTML5
vývoj riadený prehliadačmi a
"tým, čo web potrebuje",
spätná kompatibilita
obnovená
2007
XQuery
a ďalej
budúce revízie

Standard Generalized Markup Language (SGML)

- 1986 – ISO štandard
- Portabilný spôsob reprezentácie dokumentov v PC
- Priamo opisuje štruktúru dokumentu namiesto „dočasného“ opisu ako napr. formátovanie (závislé od štruktúry)
- Základ pre mnoho značkovacích jazykov
  - Meta-jazyk pre opis značkovacích jazykov

SGML - Charakteristika

- Deskriptívne značky
  - Dokument obsahuje objekty rôznych tried
    - Elementy – generické typy elementov
    - Kapitoly, nadpisy, referencie, grafické objekty...
    - Značky (tag) v tvare <...> ohraničujú elementy začiatočný (start) koncový (end) tag:

      <QUOTATION>Full speed ahead</QUOTATION>
  - Elementy – samo-opisné

SGML - Charakteristika

- Hierarchická štruktúra
  - Elementy v elementoch – vytvárajú strom

```
<CHAPTER>
    <TITLE>Getting Started</TITLE>
    <SECTION>
        <TITLE>Overview of SGML</TITLE>
        <P> SGML is ...</P>
    </SECTION>
</CHAPTER>
```

- Flexibilita
  - Používateľom / aplikačne definované elementy

SGML - Charakteristika

- Formálna špecifikácia
  - Document Type Declaration (DTD)
  - Definícia typov elementov a atribútov
  - Možno validovať SGML dokument, či spĺňa DTD
- Čitateľná reprezentácia aj pre ľudí
  - Textový dokument
  - Zrozumiteľné značky: <EMPH>very</EMPH>

SGML – štruktúra vs. formátovanie

<SK><DS><IN L="+I0" R="-10">The quotation...

<SK><IL L="+10">The following paragraph...

<DOC>

.sk;.ds;.in +i0 -i0;

The quotation...

.sk;.il +i0;

The following paragraph..

<DOC>

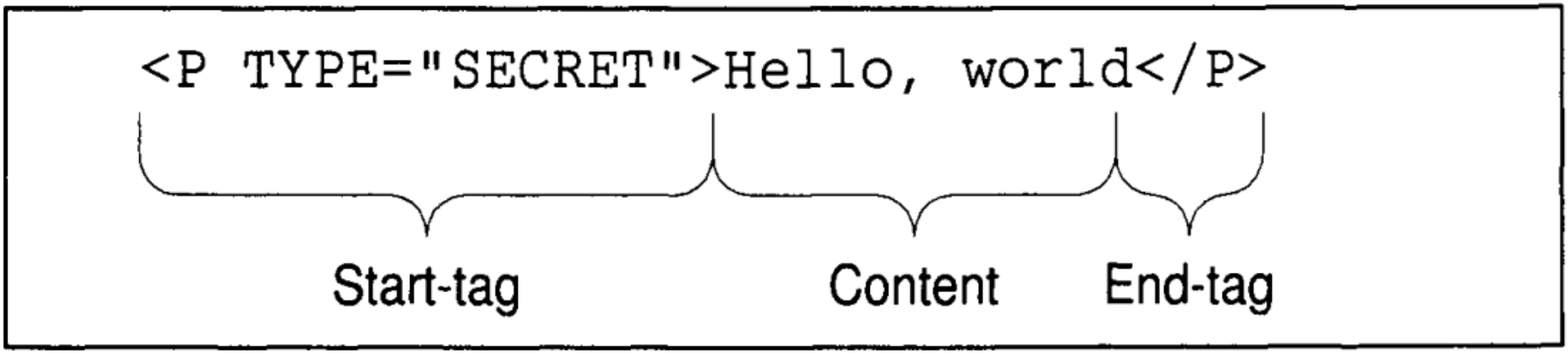
<QUO>The quotation...</QUO>

<P>The following paragrap...</P>

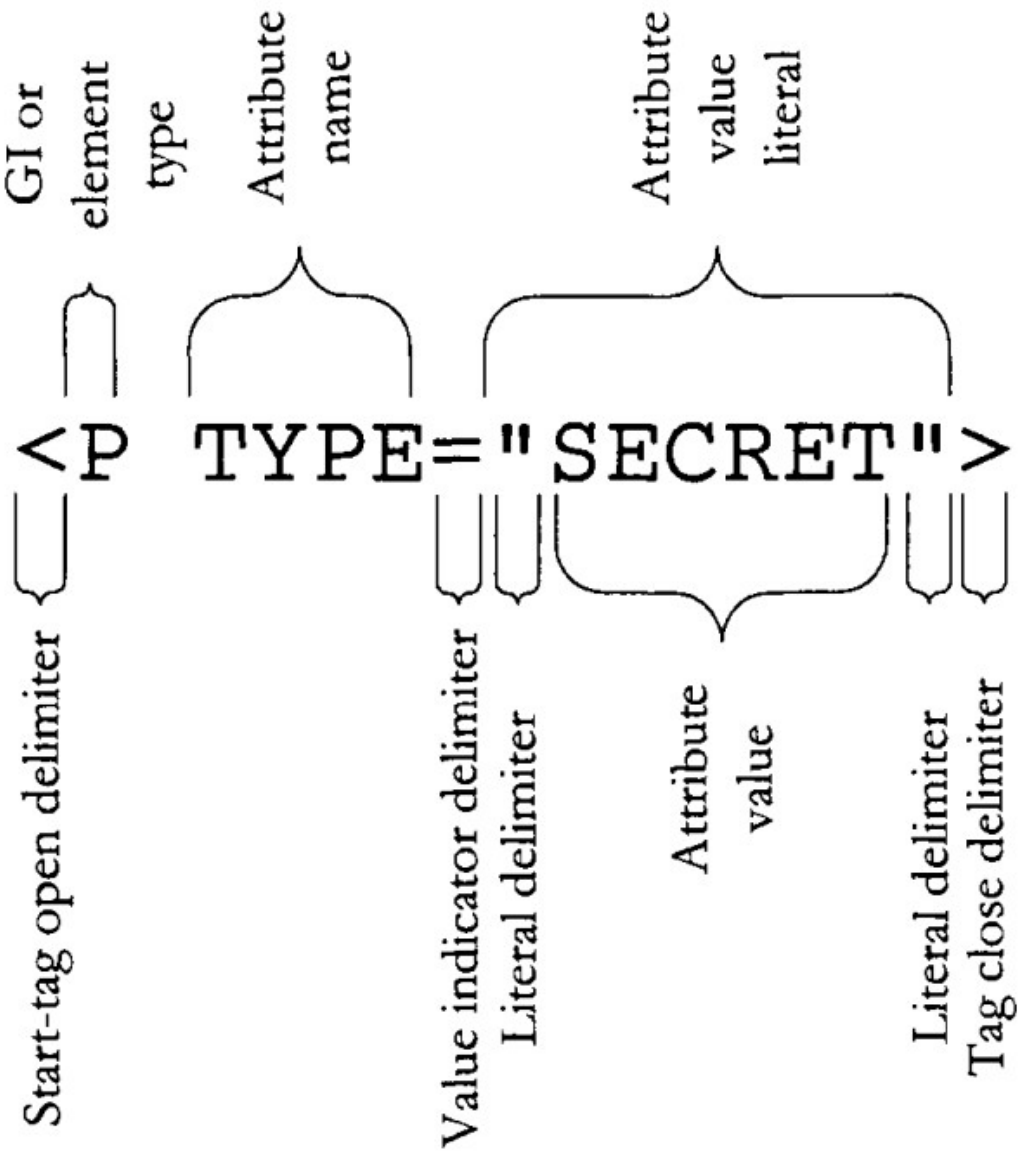
SGML – štruktúra súboru

- SGML deklarácie
  - Znaková sada, definovanie delimiterov, …
- DTD
- Inštancia dokumentu = obsah dokumentu
  - Elementy
  - Atribúty
  - Entity
  - Dáta

SGML - elementy



SGML - elementy



- **Príklad**

```
<LIST TYPE="BULLETED">
    <ITEM>...</ITEM>
</LIST>
```

SGML – DTD

- Document Type Definition (DTD)
- Definovanie elementov – pravidlá určujúce:
  - ktoré iné elementy môže element obsahovať
  - v akom poradí
  - v ako počte
  - či môže, a ako, obsahovať textový obsah
- Operátory medzi elementmi:

  "," sekvencia, | or, & and

  ? voliteľný, + jeden a viac, * nula a viac

  none presne jeden element

SGML – DTD príklad

```
<?xml version="1.0"?>
<!DOCTYPE note [
   <!ELEMENT note (to,from,heading,body)>
   <!ELEMENT to (#PCDATA)>
   <!ELEMENT from (#PCDATA)>
   <!ELEMENT heading (#PCDATA)>
   <!ELEMENT body (#PCDATA)>
]>
<note>
   <to>Tove</to>
   <from>Jani</from>
   <heading>Reminder</heading>
   <body>Don't forget me this weekend</body>
</note>
```

SGML – príklad 2

January 27, 1993

Dear Jean Luc,

How are you doing?
Isn't it *about time* you visited?

See you soon,

Genise

SGML - príklad 2

```
<letter filecode="97022701">
    <date>January 27, 1997</date>
    <greeting>&salute; Jean Luc,</greeting>
    <body>
        <p>How are you doing?</p>
        <p>Isn't it <emph>about time</emph> you visited?</p>
    </body>
    <closing>See you soon,</closing>
    <sig>Genise</sig>
</letter>
```

SGML – príklad 2 - DTD

```
<!DOCTYPE letter [
<!-- declare root element type for document -->
<!ELEMENT letter - (date, greeting, body, closing, sig)>
<!ATTLIST letter
                filecode NUMBER #REQUIRED
                secret   (yes Ino) "no">
<!ELEMENT - - body (p)*>
<!ELEMENT (date, greeting, closing, sig) - - (#PCDATA)>
<!ELEMENT p - - (#PCDATA I emph)*>
<!ELEMENT emph  - -  (#PCDATA)>
<I-- Provide a handy, replaceable greeting -->
<!ENTITY salute "Dear">
]>
```

Extensible Markup Language (XML)

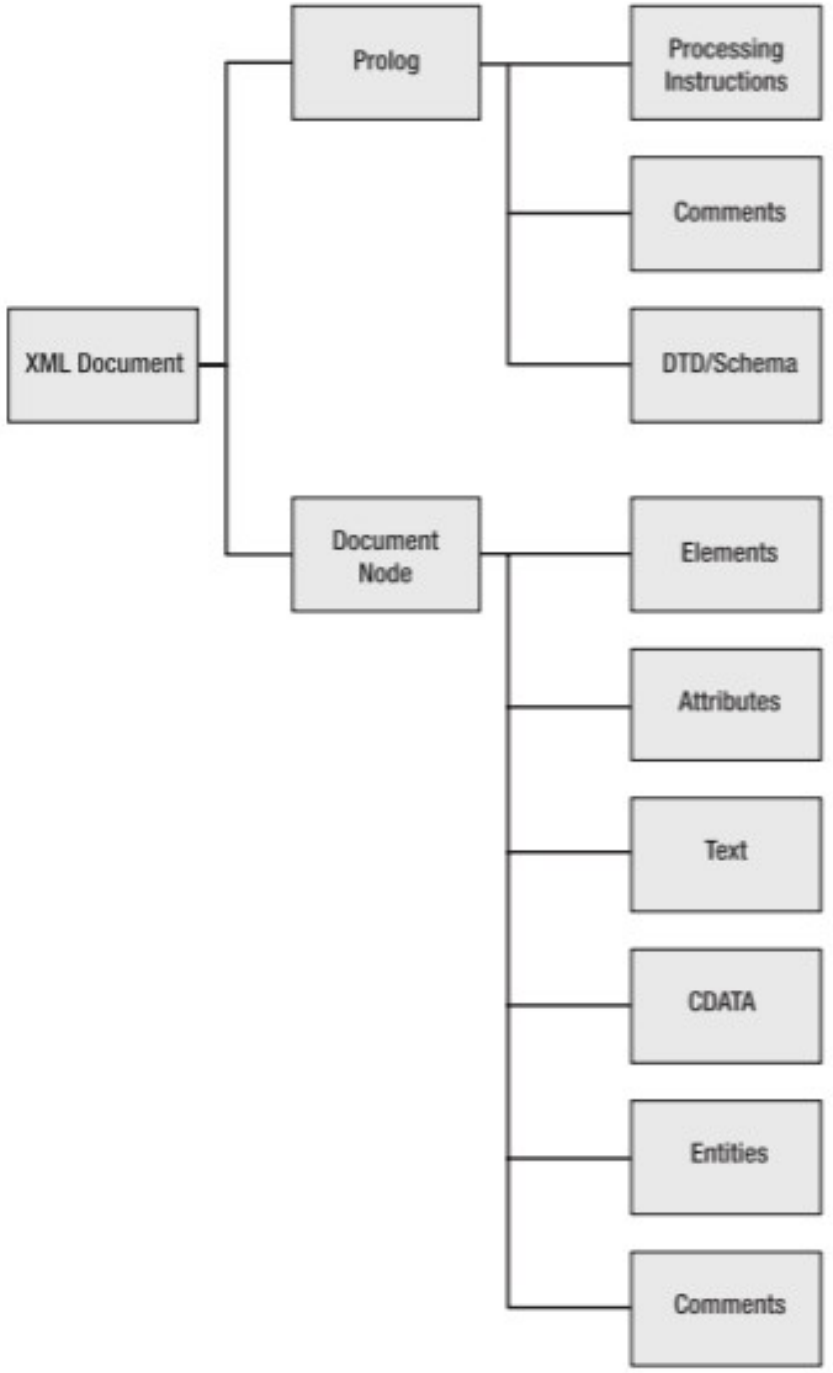
- Jazyk na opis pravidiel pre vytvorenie iných jazykov
- Správne definované dokumenty:
  - Obsahuje jeden a viac elementov
  - Obsahuje jediný dokument element, ktorý obsahuje ďalšie elementy
  - Každý element je správne ukončený
  - Elementy sú *case-sensitive*
  - Atribúty sú v “” a niesu prázdne

Extensible Markup Language (XML)

```
<?xml version="1.0" encoding="UTF-8"?>

<!-- This XML document describes a DVD
library -->

<library>

  <DVD id="1">

    <title>Breakfast at Tiffany's</title>

    <format>Movie</format>

    <genre>Classic</genre>

  </DVD>

  <DVD id="2">

    <title>Contact</title>

    <format>Movie</format>

    <genre>Science fiction</genre>

  </DVD>

  <DVD id="3">

    <title>Little Britain</title>

    <format>TV Series</format>

    <genre>Comedy</genre>

  </DVD>

</library>
```

Extensible Markup Language (XML)

- Stromová štruktúra



Hypertext Markup Language (HTML)

- „volná“ syntax v HTML 3.2

```
<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 3.2//EN">
<title>An HTML 3.2 example</title>
<body bgcolor="#FFF6F0"
 text="#000000"
 link="#C00000">
<h1 align=center>Example header</h1>

 <p><A HREF=http://www.example.com/><img align=left border=0 alt="Example:"
width=102 height=52
src=http://www.example.com/images/author.jpg></A> <i>The Author</i>
</body>
```

XHTML

- XHTML 1.0 – reformulácia HTML 4.0 do XML
- XHTML 1.1 – modularizácia elementov

```
<?xml version="1.0" encoding="utf-8"?>
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.1//EN" ↩
 "http://www.w3.org/TR/xhtml11/DTD/xhtml11.dtd">
<html xmlns="http://www.w3.org/1999/xhtml" xml:lang="en">
  <head>
    <title>XHTML 1.1 sample document title</title>
    <meta http-equiv="Content-Type" content="application/xhtml+xml; charset=utf-8" />
  </head>
  <body>
    <p>
    XHTML 1.1 sample document body
    </p>
  </body>
</html>
```

- XTHML+MathML+SVG

(X)HTML5

- HTML5 možno zapisovať v:
  - HTML formáte
  - XML formáte: XHTML5
- Nielen štruktúra:
  - Podpora multimédií
  - API
  - Web aplikácie

```
<?xml version="1.0" encoding="utf-8"?>
<!DOCTYPE html>
<html xmlns="http://www.w3.org/1999/xhtml">
  <head>
    <title>An XHTML5 example</title>
    <meta charset="UTF-8" />
  </head>
  <body>
    <header>
      <h1>Document sample</h1>
    </header>
    <section>
      <article>
        <h2>Article1</h2>
          The first article of the document.
      </article>
      <article>
        <h2>Article2</h2>
          The second article of the document.
      </article>
    </section>
    <footer>
      Copyright © 2011 John Smith. All rights reserved.
    </footer>
  </body>
</html>
```

Rozdiely v (X)HTML(5)

- Elementy uzavreté
  - XHTML

    ```
    <p>This is the first paragraph.</p>
    <br />
    <p>This is the second one.</p>
    ```

  - HTML

    ```
    <p>This is the first paragraph. <p>This is the second one.
    <br>
    ```

- V správnom poradí
  - XHTML

    ```
    <p>Part of this <strong>bold text should be <em>italic as well</em></strong>.</p>
    ```

  - HTML

    ```
    <p>Part of this <strong>bold text should be <em>italic as well</strong></em>.</p>
    <br> unterminated elements are incorrect in XHTML <hr>
    ```

Rozdiely v (X)HTML(5)

- Atribúty v “”
  - XHTML správne

    ```
    <input type="checkbox" name="checkbox" id="checkbox"
    value="True" checked="checked" />
    ```

  - XHTML nesprávne

    ```
    <input type=checkbox name=checkbox id=checkbox value=True
    checked />
    ```

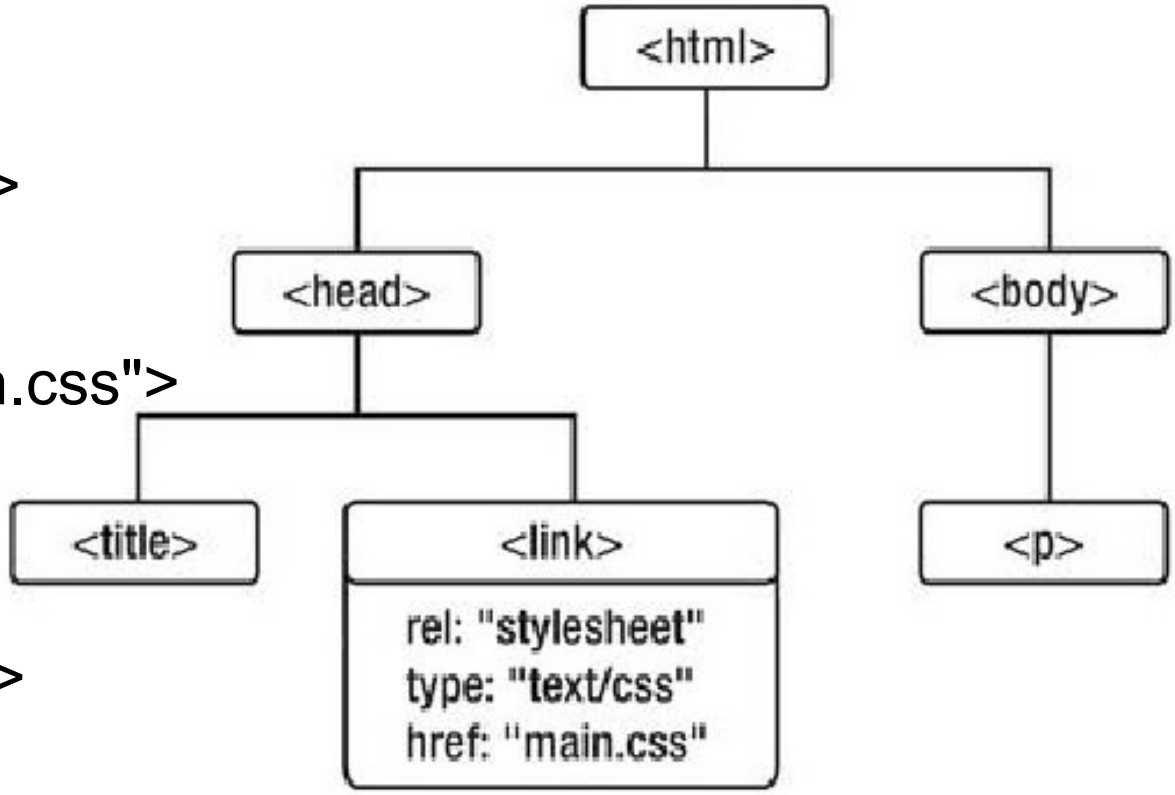
- XHTML – neumožňuje ľubovolné vnáranie
- Nepoužívať staré elementy, atribúty
- Ďalšie obmedzenia: na dáta, ...

Stav (X)HTML(5)

- Počet elementov:
  - HTML:
    - 70 vo verzii 3.2
    - 91 vo verzii 4.01
    - > 100 vo verzii 5
  - XHTML:
    - 89/92/78 vo verzii 1.0
    - 83 vo verzii 1.1
    - 99 vo verzii 2.0
    - >100 vo verzii 5

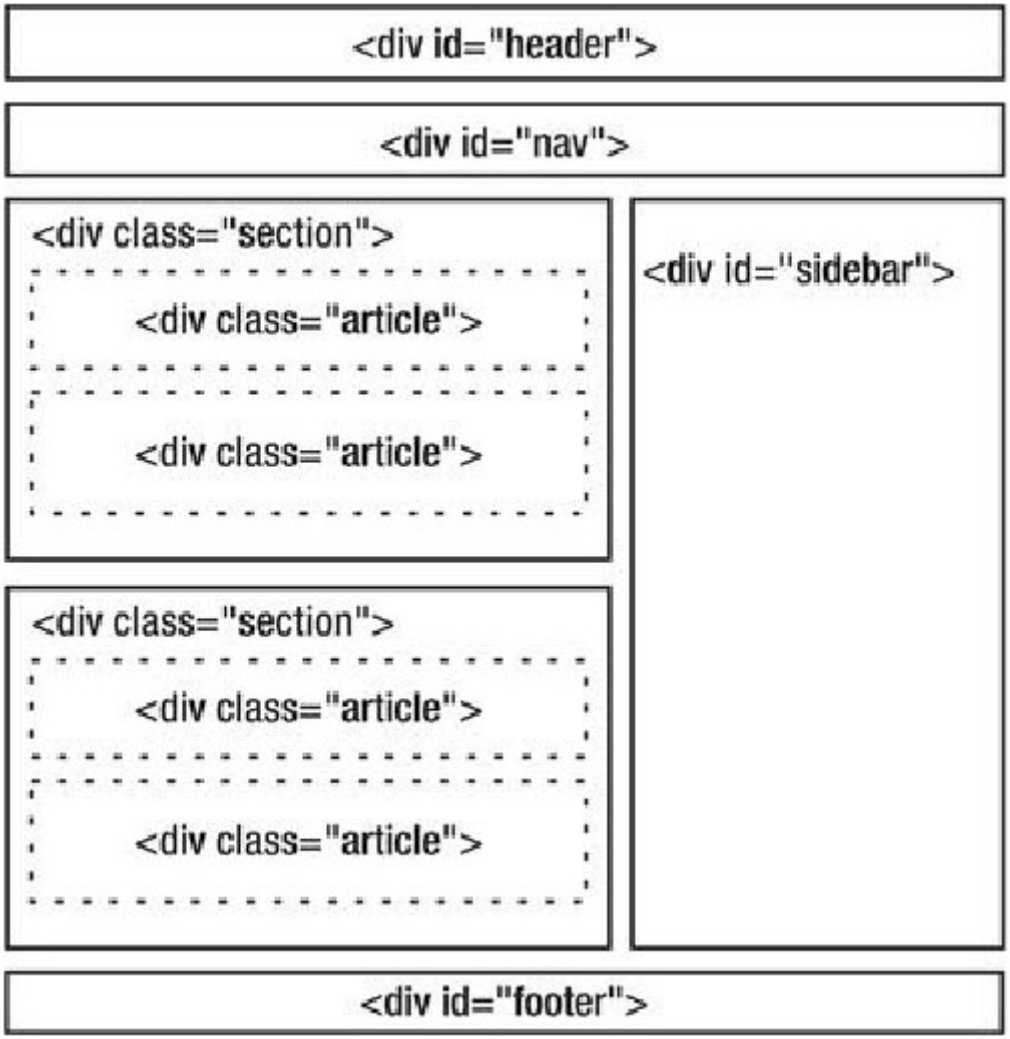
Štruktúra HTML dokumentu

```
<!DOCTYPE html>
<html>
    <head>
    <title>A DOM example</title>
        <link rel="stylesheet"
        type="text/css" href="main.css">
    </head>
    <body>
        <p>Paragraph content</p>
    </body>
</html>
```



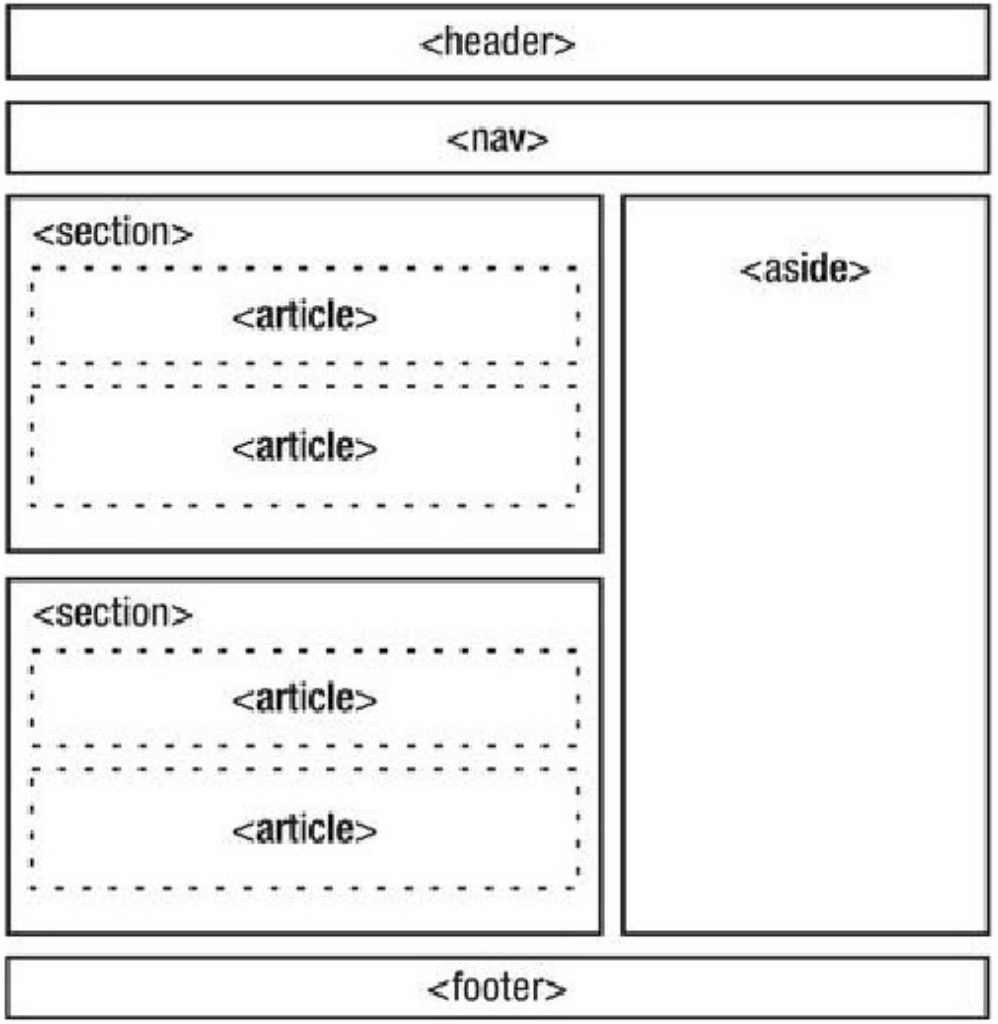
Štruktúra HTML4 dokumentu

```
<!DOCTYPE HTML PUBLIC
"-//W3C//DTD HTML 4.01//EN"
"http://www.w3.org/TR/html4/strict.dtd">
<html>
  <head>
    <title>Sample HTML document structure</title>
  </head>
  <body>
    <div class="section">
      <div class="article">
        <h2>Abstract</h2>
        <p>… first paragraph of main content …</p>
      </div>
      <div class="article">
        <h2>Overview</h2>
        <p>… second paragraph of main content …</p>
      </div>
    </div>
    <div id="footer">
      <p>
      Copyright © 2011 John Smith. All rights reserved.
      </p>
    </div>
  </body>
</html>
```



Štruktúra HTML5 dokumentu

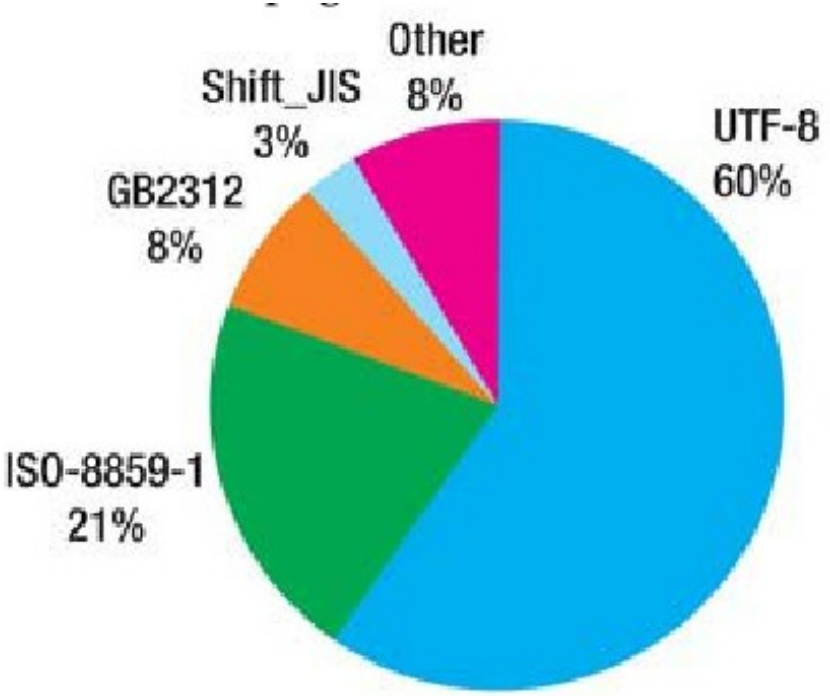
```
<!DOCTYPE html>
<html>
  <head>
    <title>Sample HTML5 document structure</title>
  </head>

  <body>
    <header>
      <h1>Document structure sample</h1>
    </header>
    <section>
      <article>
        <h2>Abstract</h2>
        <p>This sample document demonstrates the structure of
      </article>
      <article>
        <h2>Overview</h2>
        <p>
        HTML5 adds more semantics to the document stucture. In
divisions, it provides meaningful elements.
        </p>
      </article>
    </section>
    <footer>
      <p>
      Copyright © 2011 John Smith. All rights reserved.
      </p>
    </footer>
  </body>
</html>
```

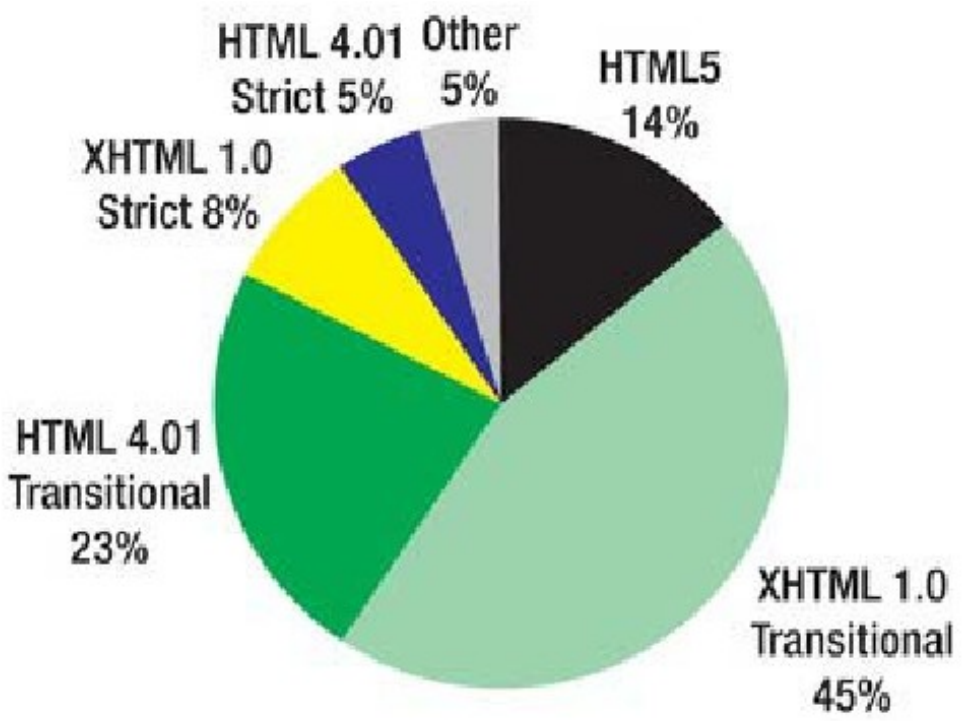


Realita?

Realita...

- 2011: štúdia na 350 najpopulárnejších web-ov
  - 94% zlyhalo na *web standard validation tests*
  - 13 rôznych jazykových sád, 9 (X)HTML standardov





Realita...

- Dnes:
  - ~90% používa HTML5
  - XHTML menej ako 10%
  - HTML 4 menej ako 2%

  - ~94% používa UTF-8

HTML5

<head> Element

- Definovanie nie viditeľných vlastností
- Kontainer pre elementy:
  - <title> - Titulok stránky
  - <style> - CSS štýl
  - <meta> - meta-informácie
  - <link> - definuje vzťah na iný súbor (napr. .css)
  - <script> - vkladá *client-side script* (napr. JavaScript)
  - <base> - základná URL pre relatívne URL cesty

<head> Element

- Meta-informácie: <meta>
  - <meta name="author" content="Peter Kapec">
  - <meta charset="UTF-8">
  - <meta name="description" content="Moja stránka">
  - <meta name="keywords" content="HTML,CSS">

Odkazy

- Zaujímavý odkaz
  - URL - *Uniform Resource Locator* má tvar:

    scheme://host.domain:port/path/filename
  - URL kódovanie – len ASCII kódovanie!
- Ukotvenie v dokumente

  Useful Tips Section

  Odkaz na kotvu v tom istom dokumente

  Odkaz na kotvu v inom dokumente

Zoznamy

- nečíslované

```
<ul>
<li>Coffee</li>
<li>Milk</li>
</ul>
```

- číslované

```
<ol>
<li>Coffee</li>
<li>Milk</li>
</ol>
```

Elementy zoskupovania

- Blokové

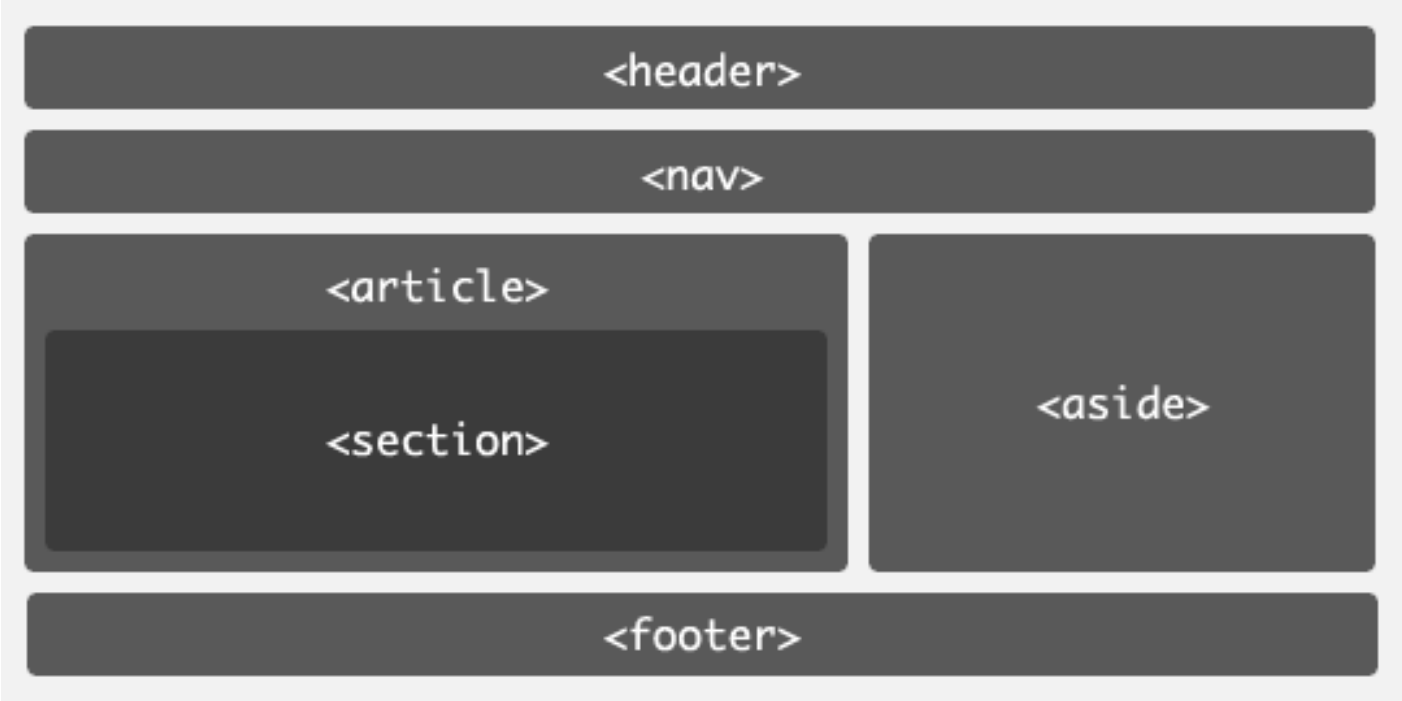
```
<div id=“article“>
<h3>Nadpis</h3>
<p>Paragraf textu. </p>
</div>
```

- Inline v texte

```
<p>Obloha je
<span
style="color:blue">
modrá</span>.
</p>
```

Sémantické elementy

- Pridávajú sémantický význam častiam HTML dokumentu
- <header>
- <nav>
- <section>
- <article>
- <aside>
- <footer>
- <figcaption>
- <figure>



Čo ďalej?

- Štúdium elementov, atribútov, pravidiel, ...
  - http://www.w3schools.com/html
  - http://www.w3schools.com/tags
- HTML5 Reference
  - http://www.w3.org/TR/html5/
    - Kap. 4 „The elements of HTML“

- !!! Skontrolujte si svoje stránky !!!
  - http://validator.w3.org/ - HMTL
  - https://jigsaw.w3.org/css-validator/ - CSS

Jazyky opisujúce stránky dokumentov

Page Description Languages

- Opisujú zalomenie dokumentu
  - Nezávislé od výstupného zariadenia
  - Podpora textu a grafiky
- 1985: PostScript (PS)
  - Level 1,2,3, Encapsulated PostScript
- 1993: Portable Document Format (PDF)
  - Verzie 1.3 (JavaScript), 1.5 JPEG 2000, 1.7 (ISO), PDF/A (long-term archiving), ...

PostScript

- Sekvencia kresliacich inštrukcií (aj pre znaky)
- High-level programming language
  - PostScript interpreter
  - Rasterizácia (vykreslenie)
- 7-bit ASCII
- % - začiatok komentára
- %% - štrukturálna informácia

PostScript

- Stranovo orientovaný (príkaz *showpage*)
- Umiestňovanie prvkov – ako „maľovanie“:
  - Grafické prvky
    - Rovné čiary, oblúky, krivky, obrázky a text
  - Geometrické operácie
    - škálovanie, translácia a rotácia
  - Atribúty čiar
    - šírka, prerušované, koncovky
  - Atribúty fontov
  - Font, typ, veľkosť
  - Farba
  - Cesty
    - Sekvencia grafických prvkov a atribútov
  - Vykresľovanie
    - odtiene šedej, farebné
  - Orezávanie

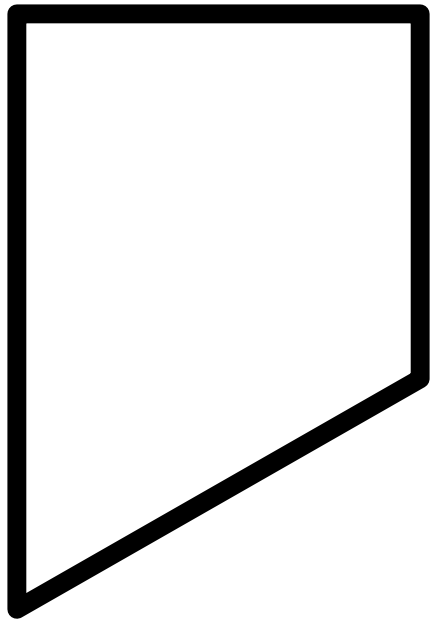
PostScript príklad - text

Welcome Haere mai Wilkommen Bienvenue Akwäba

```
%!PS-Adobe-3.0
/Helvetica findfont
14 scalefont
setfont
10 10 moveto
(Welcome ) show
(Haere mai ) show
(Wilkommen ) show
(Bienvenue ) show
(Akw) show 2 0 rmoveto (\310) show -6 0 rmoveto (aba ) show
showpage
```

PostScript príklad - grafika

```
%!PS
matrix currentmatrix /originmat exch def
/umatrix {originmat matrix concatmatrix setmatrix} def
[28.3465 0 0 28.3465 10.5 100.0] umatrix
0.1 setlinewidth
2 2 newpath moveto
3 3 lineto
3 4 lineto
2 4 lineto
closepath
0 setgray
stroke
showpage
```



PostScript

- Výhody
  - Programovací jazyk
- Nevýhody
  - Nebol navrhnutý pre zobrazovanie na monitoroch → limitujúce
  - Podpora softvéru
  - Zložitá extrakcia textu pre indexovanie

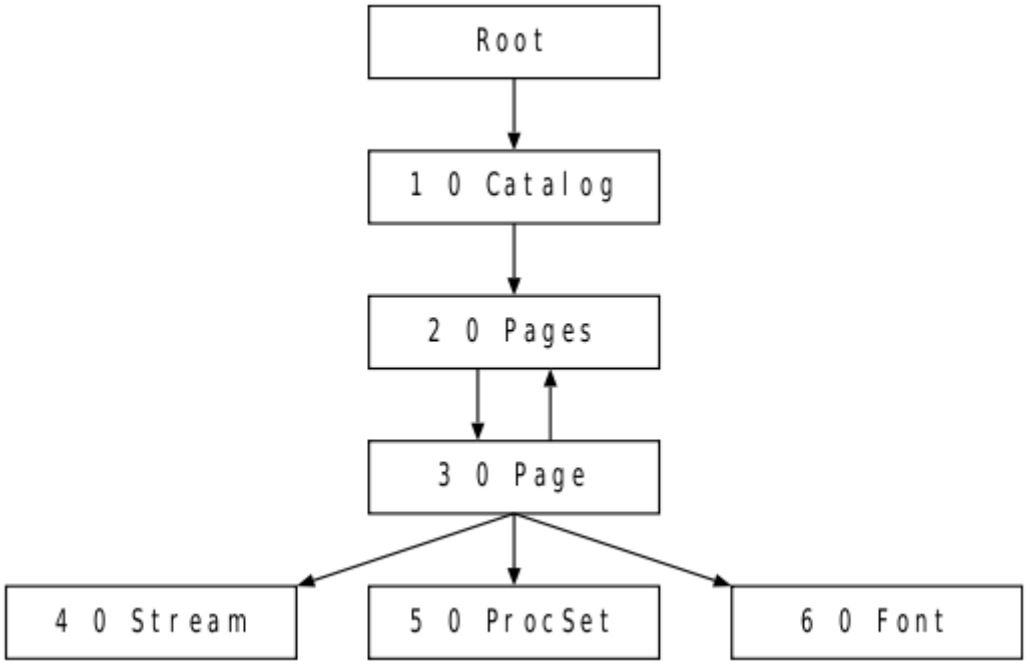
Portable Document Format

- Nasledovník PostScript-u
  - Nie je programovací jazyk
  - Kompresia a šifrovanie – zabudované
  - Pridaná podpora interaktívnych prvkov
  - Štruktúra „preddefinovaná“
    - Náhodný prístup, hierarchické štrukturovanie, navigácia
  - hyperlinky

```
%PDF-1.3
1 0 obj
<< /Type /Catalog
   /Pages 2 0 R
>>
endobj

2 0 obj
<< /Type /Pages
   /Kids [3 0 R]
   /Count 1
>>
endobj

3 0 obj
<< /Type /Page
   /Parent 2 0 R
   /MediaBox [0 0 612 792]
   /Contents 4 0 R
   /Resources << /ProcSet 5 0 R
                 /Font << /F1 6 0 R >>
              >>
>>
endobj

4 0 obj
<< /Length 118 >>
stream
BT
  /F1 14 Tf
  10 10 Td
  (Welcome ) Tj
  (Haere mai ) Tj
  (Wilkommen ) Tj
  (Bienvenue ) Tj
  (Akw\344ba ) Tj
ET
endstream
endobj

5 0 obj
[ /PDF /Text ]
endobj

6 0 obj
<< /Type /Font
   /Subtype /Type1
   /Name /F1
   /BaseFont /Helvetica
   /Encoding /WinAnsiEncoding
>>
endobj

xref
0 7
0000000000 65535 f
0000000009 00000 n
0000000062 00000 n
0000000126 00000 n
0000000311 00000 n
0000000480 00000 n
0000000511 00000 n

trailer
<< /Size 7
   /Root 1 0 R
>>
startxref
631
%%EOF
```

PDF - štruktúra

- Časti: header, objects, cross-references, trailer
- Sieť objektov – hierarchická grafová štruktúra
  - Koreň definovaný v *trailer*



Word-Processor Documents

- Špeciálne navrhnuté pre editovanie
  - Rich Text Format (RTF)
  - natívny Word
  - Microsoft Office Open XML (OOXML)
  - Open Document for Office Applications (ODF)
  - ...

Rich Text Format (RTF)

- Spätné lomítko – začiatok formátovacieho príkazu
- \yr2001 - \yr príkaz s parametrom 2001
- {\title Welcome example} – log. zoskupenia
  - Hierarchická štruktúra – efekty a formátovacie inštrukcie majú lexikálny rozsah platnosti

Rich Text Format (RTF) - príklad

{\rtf1\ansi\deff0{\fonttbl{\f0\froman Times;}{\f1\fswiss Helvetica;}}

{\info{\title Welcome example}{\creatim\yr2001\mo8\dy10}{\
nofpages1}}

\pard\plain\f1\fs28\uc0

Welcome

Haere mai

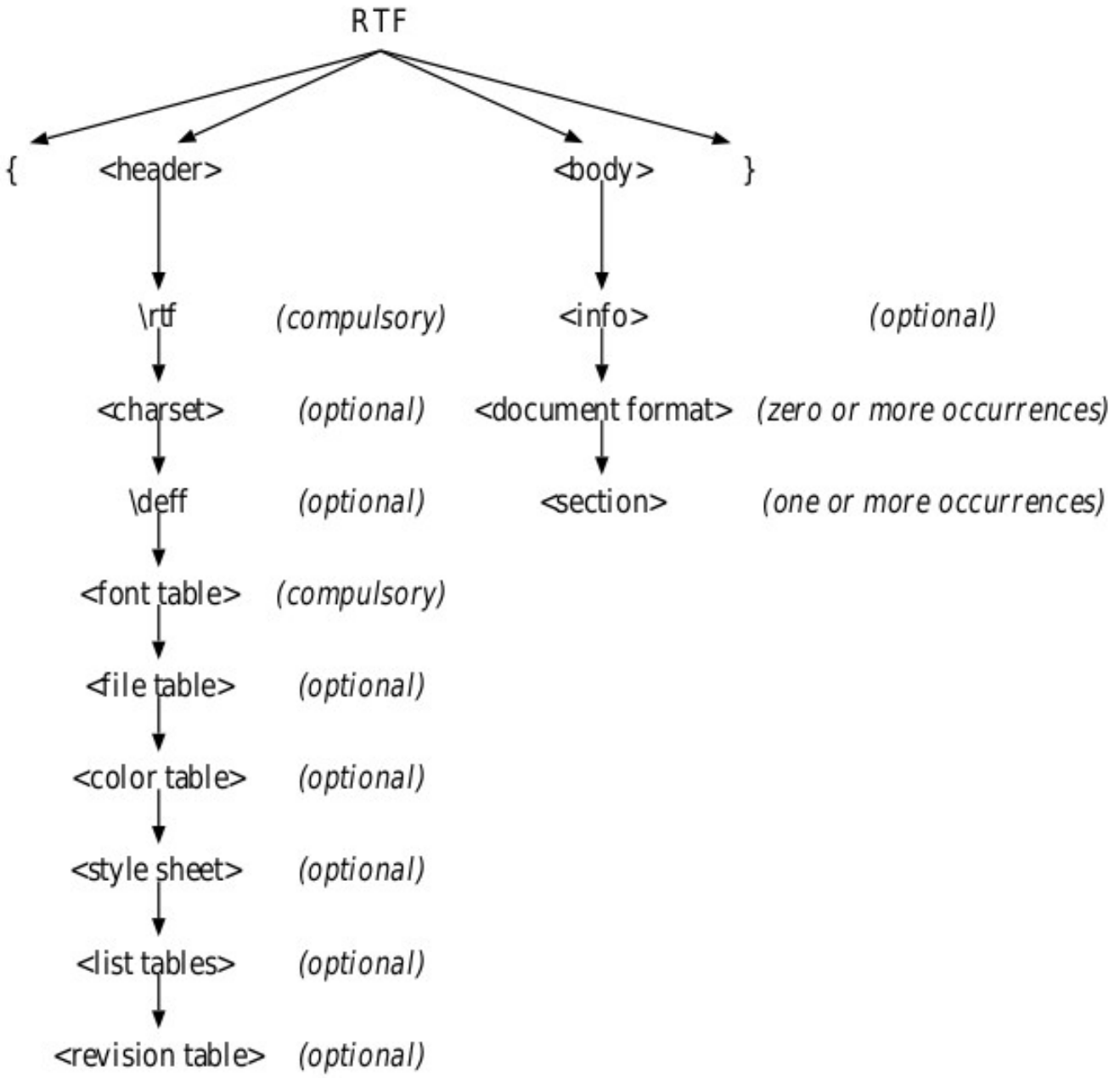
Wilkommen

Bienvenue

Akw\u228ba

\par}

RTF – štruktúra súboru



Microsoft Office Open XML (OOXML)

- Navrhnutý špeciálne pre
  - podporu Microsoft Office
  - spätne kompatibilný s binárnym formátom
- Kritika:
  - Štandard kontrolovaný komerčnou firmou
  - Nepoužíva zaužívané štandardy pre dátumy, grafiku, mat. formuly
  - ne-XML formátovacie kódy – nečitateľné pre XML parsery
  - Komponenty OOXML – previazané s Win aplikáciami

Open Document format (ODF)

- Formát pre
  - Textové dokumenty, tabuľky, prezentácie, kresby, grafiku, obrázky, grafy, mat. formule ...
  - a ich kombinácie
- XML súbor (alebo ZIP obsahujúci XML súbory)

ODF - príklad

```
<?xml version="1.0" encoding="UTF-8"?>
<office:document
xmlns:office="urn:oasis:names:tc:opendocument:xmlns:office:1.0"
xmlns:meta="urn:oasis:names:tc:opendocument:xmlns:meta:1.0"
xmlns:dc="http://purl.org/dc/elements/1.1/"
xmlns:text="urn:oasis:names:tc:opendocument:xmlns:text:1.0"
office:version="1.0">
<office:meta>
   <dc:title>Welcome Example</dc:title>
   <dc:creator>David Bainbridge</dc:creator>
</office:meta>
<office:body>
      <office:text>
         <text:p>
            Welcome Haere mai Wilkommen Bienvenue Akw&#228;ba
         </text:p>
    </office:text>
</office:body>
</office:document>
```

ODF - príklad

- **Definovanie typu písma**

```
<text:p text:style-name="Normal">
  Welcome Haere mai Wilkommen Bienvenue Akw&#228;ba
</text:p>
```

- **Definovanie štýlu**

```
<style:style style:name="Normal">
  <style:text-properties style:font-name="Helvetica"/>
</style:style>
```

ODF - súbory

- Časti: content, metadata, styles, settings
  - content.xml,meta.xml, settings.xml, styles.xml
- Importované do hlavného dokumentu
  - <office:document-content>
  - <office:document-meta>
  - <office:document-styles>
  - <office:document-settings>

Vedecké dokumenty

Vedecké dokumenty – (La)TeX

- 1982 - Donald E. Knuth
  - Sádzanie textu a matematických výrazov
- LaTeX
  - Sada makier pre TeX pre najvyššiu typografickú kvalitu – preddefinované profesionálne layout-y
- Author → Book Designer → Typesetter
- Author → LaTeX → TeX

LaTeX

- Príkazy
  - Začínajú \ napr.: \maketitle
  - S parametrami: \section{Názov sekcie}
  - Niektoré sú párové:

    \begin{document}

    …

    \end{document}
  - Používateľom definované
    - \newcommand{\Author}{Jožko Mrkvička}

LaTeX - príklad

```
\documentclass[a4paper,11pt]{article}
% This is a comment
\author{I. H. Witten and D. Bainbridge}
\title{Welcome example}
\date{10 August 2001}
\begin{document}
\maketitle
\section{Introduction}
% This is another comment.
Welcome, Haere mai, Wilkommen, Bienvenue, Akw\"aba
\section{Syntax}
LaTeX syntax is a little bit like RTF. It uses the $\backslash$
character for special formatting commands: what you see as the end result is certainly \
emph{not} what you type. One important difference from RTF is that it is designed to be
generated by people, not automatically generated by computer. This means that a written
file can be more liberal with its use    of white space and this does not affect the overall
prose.
```

LaTeX - príklad

If you really need extra spaces you need to do it \ \ like \ \ this.

Special symbols include: \{ \} \% _ \# \&. Speech marks are done “like this”.

A blank line is used to separate paragraphs. It supports all the usual document structures:

\begin{itemize}

\item bullet point list

\item enumerated list

\item tables and figures

\item drawn graphics

\item \ldots

\end{itemize}

In particular Latex has a powerful maths mode capable of expressing complex equations. A rudimentary example is:

\begin{displaymath}

x \geq \sum_{i=0}^{\infty}\frac{1}{i^2\pi}

\end{displaymath}

\end{document}

LaTeX – príklad - výstup

Welcome example

I. H. Witten and D. Bainbridge

10 August 2001

1 Introduction

Welcome, Haere mai, Wilkommen, Bienvenue, Akwäba

2 Syntax

LaTeX syntax is a little bit like RTF. It uses the \ character for special formatting commands: what you see as the end result is certainly *not* what you type. One important difference from RTF is that it is designed to be generated by people, not automatically generated by computer. This means that a written file can be more liberal with its use of white space and this does not affect the overall prose. If you really need extra spaces you need to do it like this.

Special symbols include: { } % _ # &. Speech marks are done "like this".

LaTeX – príklad - výstup

A blank line is used to separate paragraphs. It supports all the usual document structures:

- bullet point list
- enumerated list
- tables and figures
- drawn graphics
- ...

In particular Latex has a powerful maths mode capable of expressing complex equations. A rudimentary example is:

$$x \geq \sum_{i=0}^{\infty} \frac{1}{i^2 \pi}$$

1

Zhrnutie

- Klúčové poznatky z prednášky
  - Reprezentácia textu a textové dokumenty
  - Kódovanie textu
  - Jednoduchý a štruktúrovaný text
  - Značkovacie jazyky
  - HTML
  - Jazyky opisujúce stránky dokumentov
  - Vedecké dokumenty

Nabudúce

- Reprezentácia textu a textové dokumenty
  - Analógovo-digitálny prevod
  - Digitálna reprezentácia zvukových dát
  - Zvukové dáta a kompresia
  - HTML5 a audio
  - CSS

1. kontrolný bod

Odovzdať

- Dokument s opisom projektu, návrhom vzhľadu stránok a návrhom navigačného dizajnu
  - Stručný opis o čom je projekt, podľa MUDPY:
    - Concept, Goals, Requirements, ...
  - *Wireframe* návrh pre všetky pod-stránky
  - Ilustrácia navigačnej mapy
- Dokument do AIS vo formáte PDF **podľa pokynov v mieste odovzdania**
  - Termín: **11.10. do 23:59**

Ďakujem za pozornosť