"Anyone who would letterspace the lowercase would steal sheep."

Frederick Goudy (1894–1945)

"You cannot *not* communicate"

Paul Watzlawick (1921–)

Lidé komunikují prostřednictvím dokumentů připravovaných elektronicky

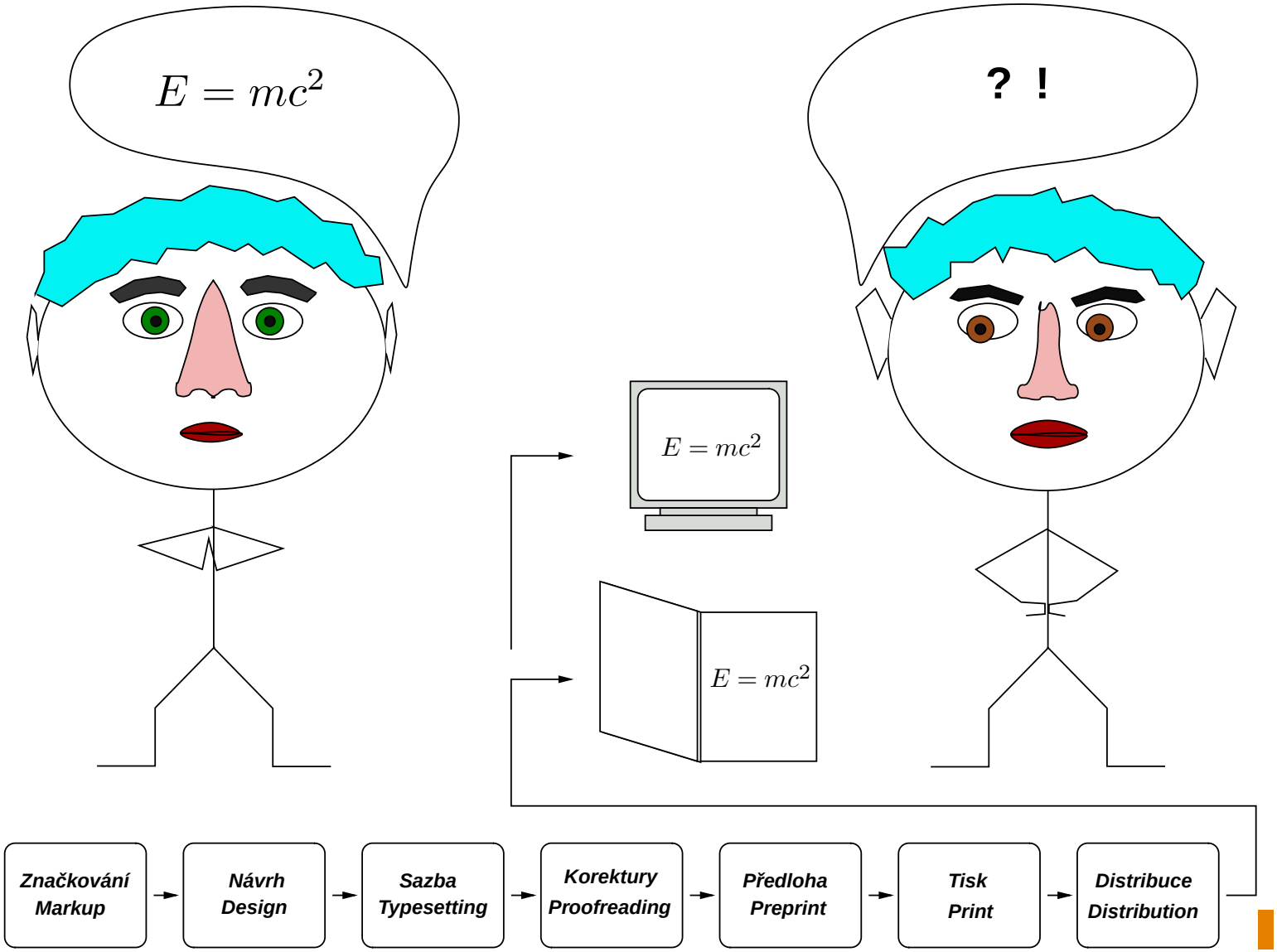
PB029 Elektronická příprava dokumentů

Petr Sojka

Fakulta informatiky
MU, Brno

podzim 2013

Autor a čtenář



Osnova dnešní přednášky

① Agenda (úkoly do cvičení, prémiové úkoly, bakalářské, diplomové či doktorské práce,. . .).

② XML? XML!

③ Aplikace SGML: HTML, XHTML, WML.

④ ISO/IEC 10646, Unicode.

⑤ Formáty a normy související s XML (XLink, XPointer, RDF, XTM). AJAX. DOM.

⑥ Návrh/design (webový): úvod.

⑦ Webový design prakticky 7. 10. 2013 (Martin Kacvinský).

Domácí úkoly (do cvičení 4)

- ☞ Do cvičení 4: Přinést si neznačkovaný text pro sazbu tištěného dokumentu do čtyř stran A4, t.j. cca osm normostran textu.
- ☞ Dovalidovat český html dokument pomocí nsgmls.

Metajazyk SGML – definice

- ☞ SGML je mezinárodní standard pro popis značkovacích jazyků.
- ☞ Formálněji je to **syntaktický *meta*jazyk** pro definici tříd dokumentů (deskriptivních značkovacích jazyků), nezávislých na abecedě.
- ☞ Formální definice typu dokumentu se nazývá **DTD** – Document Type Definition, **definice typu dokumentu** (gramatika třídy dokumentů). SGML parser (validátor) pak čte na vstupu DTD a kontroluje formální správnost.
- ☞ Instance SGML dokumentu (dále jen SGML dokument) sestává z **deklarace** (pokynů pro parser), **gramatiky** (DTD) a vlastního popisně *označkovaného textu*.

SGML deklarace

- ☞ „Pokyny pro parser“ s definicemi abeced (analogie popisu BNF):
  - ❑ role oddělovačů;
  - ❑ rezervovaná jména;
  - ❑ zakázané znaky, typicky kontrolní ASCII;
  - ❑ pravidla pro pojmenovávání, citlivost na malá a velká písmena;
  - ❑ velikosti délek, značek, . . . ;
  - ❑ komentáře;
  - ❑ parametry chování, které vlastnosti kontrolovat (OMITTAG).

SGML deklarace (cont.)

- ☞ Pokud deklarace v dokumentu není uvedena, používá se implicitně **referenční syntaxe** Reference Concrete Syntax (RCS).

Příklad SGML deklarace

```
<!SGML "ISO 8879:1986"
  CHARSET
  BASESET "ISO 646:1991//CHARSET
                   IRV//ESC 2/8 4/2"
  DESCSET
      0   9 UNUSED
      9   2   9          - TAB, LF -
     11   2 UNUSED
     13   1  13          - CR -
     14  18 UNUSED
     32  95  32
    127   1 UNUSED
  CAPACITY SGMLREF
    TOTALCAP    35000
...
 NAMECASE
        GENERAL YES
        ENTITY  NO
    DELIM
```

Příklad SGML deklarace (pokr.)

```
MDO       "<!"      - markup decl open -

MDC       ">"       - markup decl close -
DSO       "["       - declaration subset open -
DSC       "]"       - declaration subset close -
MSC       "]]"      - marked section close -
COM       "-"     - comment -
RNI       "#"       - reserved name indicator -
LIT       "&#34;"   - literal -
LITA      "’"       - alternative literal -
GRPO      "("       - group open -
GRPC      ")"       - group close -
AND       "&"       - and connector -
OR        "|"       - or connector -
SEQ       ","       - seq connector -
OPT       "?"       - opt occurrence indicator -
REP       "*"       - rep occurrence indicator -
```

Příklad SGML deklarace (pokr.)

```
        PLUS        "+"       - plus occ ind, inclusion -
        MINUS       "-"       - exclusion, omission flag -
        CRO         "&#"      - character reference open -
        ERO         "&"       - entity reference open -
        PERO        "%"       - parameter entity reference open -
        REFC        ";"       - reference close -
        PIO         "<?"      - processing instruction open -
        PIC         ">"       - processing instruction close -
        STAGO       "<"       - start tag open -
        ETAGO       "</"      - end tag open -
        TAGC        ">"       - tag close -
        NET         "/"       - null end-tag -
...
```

SGML DTD: atributová gramatika

- Terminologie teorie formálních jazyků: neterminály (**elementy**), terminály (**entity**).
- Elementy vytvářejí stromovou strukturu, nemohou se navzájem křížit, jeden element je kořenový. Lze použít prázdné elementy. `<br/>`
- Elementy mohou mít **atributy**. O

SGML DTD: atributová gramatika (cont.)

```
<!ELEMENT faktura   (odberatel,dodavatel, polozka+)>
<!ELEMENT odberatel (nazev,adresa,ico,dic)>
<!ELEMENT dodavatel (nazev,adresa,ico,dic)>
<!ELEMENT polozka   (popis?,cena,dph,ks?)>
<!ELEMENT nazev     (#PCDATA)>
<!ELEMENT adresa    (#PCDATA)>
<!ELEMENT ico       (#PCDATA)>
<!ELEMENT dic       (#PCDATA)>
<!ELEMENT popis     (#PCDATA)>
<!ELEMENT cena      (#PCDATA)>
<!ELEMENT dph       (#PCDATA)>
<!ELEMENT ks        (#PCDATA)>
<!ATTLIST faktura
          cislo     CDATA #REQUIRED
          vystaveni CDATA #REQUIRED
          splatnost CDATA #REQUIRED
          vystavil  CDATA #IMPLIED>
<!ATTLIST cena
          mena      CDATA "CZK">
```

Deklarace elementů a atributů

- Sekvence, alternativa, `ANY`, `EMPTY`, `#PCDATA`.
- Opakování: právě jednou, nejvýše jednou (?), alespoň jednou (+), libovolněkrát (*).
- Typy atributů `CDATA`, `NMTOKEN`, `NMTOKENS`, `ID`, `IDREF`, `IDREFS`, `ENTITY`, `ENTITIES`, `výčet`.
- Implicitní hodnoty atributů: `"hodnota"`, `#REQUIRED`, `#IMPLIED`, `#FIXED "hodnota"`.

Entity

- ☞ Vhodné pojmenovat části a fragmenty SGML, některé řetězce (`&TeX;` v IS) a znaky použité jako oddělovače v gramatice: `<`, `&`, `&quot;`.
- ☞ **Entity** interní textové, externí textové, externí binární a parametrické.
- ☞ Příklad množiny interních textových entit: ISO Latin2.

Entity (cont.)

```
<!- Character entity set. Typical invocation:
  <!ENTITY % ISOlat2 PUBLIC
   "ISO 8879:1986//ENTITIES Added Latin 2//EN"> %ISOlat2;
->
<!ENTITY abreve SDATA "[abreve]"-=small a, breve->
<!ENTITY Abreve SDATA "[Abreve]"-=capital A, breve->
<!ENTITY amacr  SDATA "[amacr ]"-=small a, macron->
<!ENTITY Amacr  SDATA "[Amacr ]"-=capital A, macron->
...
```

- ☞ Externí textové entity:

  <!ENTITY *název* SYSTEM "*URI*">

- ☞ Externí binární entity:

  <!ENTITY *název* SYSTEM "*URI*" NDATA "*notace*">

Další příklady DTD

- ☞ DTD diplomové práce O.
- ☞ DTD informací o studijním předmětu O.
- ☞ DTD rozvrhu.

Připojení DTD k instanci dokumentu

- Odkazem na soubor:

```
<!DOCTYPE faktura SYSTEM "faktura.dtd">
<faktura> ...</faktura>
```

- DTD součástí dokumentu:

```
<!DOCTYPE faktura [
 <!ELEMENT faktura (odberatel,
                    dodavatel, polozka+)>
 ...
 ]>
 <faktura> ... </faktura>
```

- **Veřejný identifikátor** `PUBLIC` s URL:

Připojení DTD k instanci dokumentu (cont.)

```
<!DOCTYPE wml PUBLIC "-//WAPFORUM//DTD
                                WML 1.3//EN"
"http://www.wapforum.org/DTD/wml13.xml">
```

V případě veřejného identifikátoru je mapování od řetězce k souboru určeno tzv. **katalogem** (proměnná okolí `SGML_CATALOG_FILES` či `XML_CATALOG_FILES`).

Příklad SGML dokumentu

```
<!DOCTYPE faktura SYSTEM "faktura.dtd">
<faktura vystaveni="6.10.2000"
         splatnost="20.10.2000">
  <odberatel>
    <nazev>Ferda Pistorius</nazev>
    <adresa>Boubín 77</adresa>
    <ico>2862667777</ico> <dic>291-2862667777</dic>
  </odberatel>
  <dodavatel>
    <nazev>Hynek Bedna</nazev>
    <adresa>Máchovo jezero 78</adresa>
    <ico>2862467777</ico> <dic>291-2862467777</dic>
  </odberatel>
  <polozka>
    <popis>XML editor</popis>
    <cena mena="Kč">500</cena>
  </polozka>
</faktura>
```

Pohledy na SGML

- **Lexikální pohled**: SGML dokument je řetěz znaků, z nichž některé jsou **data** a některé **oddělovače**, *značky*.

  ```
  <article> <title>Úvod do SGML</title> <section>SGML: hlavní
  principy</section> <subsection>Zobecněné značkování</subsect
  <p>Základem ... ... </article>
  ```

- **Syntaktický pohled**: SGML dokument má tři části: SGML deklaraci (instrukce pro SGML parser), DTD (gramatiku jazyka) a vlastní instanci dokumentu (v tomto jazyce). Instance dokumentu obsahuje *data* a *značky* vyhovující DTD. Tento pohled vytváří parser z lexikálního pohledu.

Pohledy na SGML (cont.)

- ☞ **Pohled hierarchie objektů**: SGML dokument je uspořádaná hierarchie (obvykle stromová struktura) objektů s obsahem (**elementů**). Elementy mají atributy, obsah a další vlastnosti. O
- ☞ **Entitně-strukturní pohled**: SGML dokument je vnořená kolekce **entit**. Většina entit je pojmenovaných. **Textové** entity mohou odkazovat na další entity, zatímco **datové** entity obsahují libovolná data.
- ☞ **Aplikační pohled**: SGML dokument je cokoli, co autor chce aby jím bylo. *Data belongs to whoever creates it, and you get to decide what's important about your own data.*

SGML – validace a parsing

- ☞ Validace. Parsery. SP a sgmls/nsgmls Jamese Clarka.
- ☞ Výstup parseru: **Element Structure Information Set** (ESIS) formát.
- ☞ Ukázky validace, chyb, ladění, ESIS.
- ☞ Třídy dokumentů. CATALOG. `/packages/share/sgml-catalogs/`
- ☞ Další příklady DTD – informace o studijním předmětu. O

Vytváření značkovaných dokumentů

- ☞ Běžný ascii editor (`emacs`, `joe` nebo i `notepad`).
- ☞ PSGML mód emacsu: příklad stránky předmětu.
- ☞ Komerční systémy: Arbortext (Epic), SoftQuad, Corel XMetal . . .
- ☞ (X)HTML svět: Mozilla/Firefox, MSIE, Amaya, HomeSite, Netscape, FrontPage, . . .

Značkovací jazyky na Internetu – vývoj

- ☞ 3/1989: návrh projektu World Wide Web, Tim Berners-Lee, CERN
- ☞ 12/1990: návrh HTML DTD, první Web software pod NExT
- ☞ 1991: první WWW prohlížeč pro omezené užití
- ☞ 1992: CERN začíná propagovat WWW projekt
- ☞ 2/1993: NCSA zveřejňuje alfa verzi prohlížeče Mosaic/X Marca Andreesena

Značkovací jazyky na Internetu – vývoj (cont.)

- 7/1993: HTML 1.0 specifikace (Hypertext Markup Language) (RFC 1866) jako Internet draft pracovní skupinou IETF/IIR (Internet Engineering Task Force Internet Information Resources): dokumentový jazyk definovaný pomocí SGML užívaný na WWW.
- 9/1993: Mosaic pro PC, MAC a X-Window
- 11/1993: Dave Raggett (HP) navrhuje HTML specifikaci s formuláři, tabulkami a rovnicemi
- 4/1994: HTML DTD test suite (Dan Connolly)
- 6/1994: MIT/CERN vytváří organizaci W3

Značkovací jazyky na Internetu – vývoj (cont.)

- 6/1994: IETF vytváří pracovní skupinu HTML a HTML 2.0 specifikaci
- 11/1994: volně šířený prohlížeč Netscape 1.0
- 4/1995: Netscape Navigator 1.1
- 5/1995: Netscape a Sun se dohodli na podpoře Javy
- . . . World Wide Web Consortium (W3C).
- XML/XHTML, MathML.

Výměna dokumentů na Internetu

- Identifikace dokumentů na Internetu: Uniform Resource Locator/Identifier/Name (URL/URI/URN).
- URN: URI, kdy organizace zajišťuje trvalost odkazů (typicky překladovou službou).
- `metoda://server[:port]/cesta/soubor[#kotva]`
- Http, https, mailto, news, file, ftp, gopher, rlogin, telnet, tn3270, wais.
- Klient/server; http, https server.
- Klienti: Google Chrome, Firefox/Mozilla/Netscape Navigator, MS Internet Explorer, Safari, Opera, Galeon, Konqueror, Lynx, Mosaic, Amaya, HotJava, ...

Výměna dokumentů na Internetu (cont.)

- ☞ Různé módy renderování v posledních verzích prohlížečů: zpětně kompatibilní (*quirk mode*) a standardy W3C dodržující (*standard mode*).
- ☞ Servery: Apache, MS IIS, Netscape Commerce, NCSA.

Dokumentové jazyky založené na SGML

- ☞ Nejrozšířenější je HTML.
- ☞ Vývoj HTML – různá *fixní* DTD (jedné třídy dokumentů).
- ☞ Koordinace W3C, velké nekompatibilní odchylky velkých firem, problémy s validací.
- ☞ DocBook DTD: svět Linuxu, dokumentace, knihy (O'Reilly, Kosek, Safari online).
- ☞ Další rozšířené dokumentové jazyky: TEI (Text Encoding Initiative) DTD.
- ☞ Rainbow DTD: formalizace RTF pro konverze z Wordu.
- ☞ WML (WAP), CALS, MATHML, TEXML, ...
- ☞ Nyní HTML5 viz rozdíly HTML4 a HTML5.

Proč XML? Desatero cílů.

XML (**E**xtensible **M**arkup **L**anguage) – zjednodušená verze SGML optimalizovaná pro použití na Internetu vyvíjená konsorciem W3C.
Desatero cílů:

- ☞ Přímočaré použití na Internetu.
- ☞ Široké spektrum použití/aplikací.
- ☞ Kompatibilita s SGML.
- ☞ Snadnost vytváření programů pro práci s XML.
- ☞ Absolutní minimum či absence volitelných rysů XML.
- ☞ Čitelnost a jasnost.
- ☞ Rychlost návrhu.

Proč XML? Desatero cílů. (cont.)

- ☞ Formální popis a návrh.
- ☞ Snadnost vytváření XML dokumentů.
- ☞ Úsečnost zápisu není důležitá.

XML? XML! Nikdy není pozdě!

Co je to XML?

- Doporučení W3C: rozšiřitelný značkovací *(meta)*jazyk.
- Univerzální a otevřený formát pro reprezentaci (a výměnu) téměř libovolné datové struktury – dat i dokumentů. Násobné a opakované použití jako u SGML.
- Celá sada technologií a formátů s XML souvisejících (XPointer, XLink, XSL) pro elektronickou výměnu dat (a dokumentů).

Vymezení XML

- ☞ Odlišení od HTML: rozšiřitelnost, možnost změn sémantiky značek. XML nespecifikuje ani sémantiku, ani množinu značek, umožňuje však značky definovat a definovat jejich strukturální závislosti. Sémantiku určují aplikace (webové prohlížeče jsou jedny z nich).
- ☞ Odlišení od SGML: zjednodušeně řečeno XML je SGML s restrikcemi. Odchylky jsou minimální (chování mezery).
- ☞ Technický úvod do XML Normana Walshe.

Syntaxe XML

- ☞ Slabší pojem než validita: **správná strukturovanost** (**well-formed**).
- ☞ Správně strukturovaný dokument by měla zpracovat každá XML aplikace.
- ☞ Validátory SGML umí obvykle i XML.
- ☞ Parsery: nsgmls, msxml, xerces, xmllint, . . .
- ☞ Sekce CDATA pro pohodlnější psaní části dokumentů obsahujících významné znaky:

Syntaxe XML (cont.)

```
<moudrost>
<![CDATA[
 (1 < 2) & 2 = 1000 x 1
]]>
</moudrost>
```

XML deklarace

```
<?xml version="1.0" encoding="iso-8859-2"
                                standalone="no"?>
```

- ☞ Musí být na prvním řádku dokumentu.
- ☞ Při `standalone` *yes* není třeba číst externí DTD.

Instrukce pro zpracování

Pokyny (procesní instrukce) pro specifické aplikace:

```
<?xml-stylesheet href="epd.css" type="text/css"?>
```

nebo

```
<?LaTeX \pagebreak?>
```

či

```
<datum>
 <?php echo Date "d.m.Y"?>
</datum>
```

XHTML? XHTML!

- ☞ Postupný přechod z HTML: XHTML, reformulace HTML jako modulární XML aplikace.
- ☞ XHTML čtou všechny XML-kompatibilní aplikace, a zároveň je možno psát **dobře zformované** (well-formed) dokumenty již nyní.
- ☞ Příklad XHTML dokumentu:

XHTML? XHTML! (cont.)

```
<!DOCTYPE html
  PUBLIC "-//W3C//DTD XHTML 1.0 Strict//EN"
  "http://www.w3.org/TR/xhtml1/DTD/xhtml1-strict.dtd">
  <html>
   <head>
    <title>simple document</title>
   </head>
   <body>
    <p>a simple paragraph</p>
   </body>
  </html>
```

MathML: značkování matematiky na Webu

- MATHML 3.0 (Mathematical Markup Language) – doporučení W3C. XML slovník/aplikace pro popis matematiky (struktury formulí i prezentačních forem).
- Příklad: kubickou křivku formalizovat pro další zpracování (vykreslení křivky, integrace, ...) na základě XML kódu:

```
<math xmlns='http://www.w3.org/1998/Math/MathML'>
  <msup>
    <mi>x</mi>
    <mn>3</mn>
  </msup>
</math>
```

MathML: značkování matematiky na Webu (cont.)

- ☞ Použití jak pro sazbu, tak pro výpočet v systémech jako je Mathematica: přímočaré odlišení víceznačností: diferenciál x od proměnné dx a součinu proměnných dx.
- ☞ Nativní podpora MathML od Mozilly 1.1. Renderovací stroj MathML pro MSIE: MathPlayer. Techexplorer IBM: plug-in pro Navigator a MSIE pro renderování TEXu, LATEXu a MathML.
- ☞ Možnosti výpočtů, validace, renderování na MathMLcentral. Možnost copy&paste.
- ☞ Značkování matematiky v (Tagged) PDF.
- ☞ Indexování MathML na `http://eudml.org`. Ukázka hledání včetně formulí.

MathML: značkování matematiky na Webu (cont.)

- ☞ Přednáška o Digitálních matematických knihovnách na Informatickém kolokviu 8. 10. 2013 v D2: jste zváni!

WML

- ☞ WML (Wireless Markup Language) je jazyk dokumentů, které jsou zobrazitelné na displejích mobilních telefonů: jsou dostupná DTD.
- ☞ Průmyslová asociace WAPFORUM. (1997, Nokia, Ericsson, Motorola, Unwired Planet).
- ☞ WAP (Wireless Application Protocol) – komunikační protokol pro přenos informací z Internetu na mobilní zařízení. resp. `/packages/share/sgml-catalogs/WML`.
- ☞ WAP browsery: CCWAP, Nokia browser, WINWAP, ...
- ☞ Více paměti, silnější procesory: přechod na XHTML?

Příklad WML dokumentu

```
<?xml version="1.0"?>
<!DOCTYPE wml PUBLIC
 "-//WAPFORUM//DTD WML 1.1//EN"
 "http://www.wapforum.org/DTD/wml_1.1.xml">
 <wml>
   <card id="Card1" title="ccWAP WML ">
   <p>
   <!- I am learning WML example ->
   I am learning the basics of WML.
   </p>
   </card>
 </wml>
```

WML – pokračování

- Konfigurace http serveru pro WML: do souboru `~/.htaccess` je třeba přidat:

```
addtype text/vnd.wap.wml Wml
addtype Application/vnd.wap.wmlc Wmlc
addtype text/vnd.wap.wmlscript Wmls
addtype Application/vnd.wap.wmlscriptc Wmlsc
addtype image/vnd.wap.wbmp wbmp
```

- Prohlížení většinou pomocí apletů v prohlížeči, vývojová prostředí např. EasyPad Waptor.