

Zásady rekonštrukcie textov

Robert Nádvorník – Róbert Sabo

Materiál nebol publikovaný, je len na interné použitie

Postup pri prevádzaní tlačeneho textu do elektronickej podoby:

naskenovanie publikácie → .tif súbor → rozpoznanie textu → .rtf súbor → kontrola a korekcia textu v textovom editore → .sxw súbor (OpenOffice formát)

Pri kontrole a korekcii textu dodržiavame konvenciu 1 : 1, t. j. v naskenovanom texte opravujeme iba chyby vzniknuté nekorektným rozpoznaním. Preklepy a iné chyby, ktoré sa vyskytujú v pôvodnom tlačennom texte, rekonštruujeme tak, ako boli vytlačené.

Konvencia pomenovania súborov:

!Súbory sa v zásade nepremenovávajú.

Ak je už text opravený pridá sa za pôvodný názov súboru dátum ukončenia korektúry a slovo final.

Príklad:

george.tif → george.rtf → george-2004-06-02-final.sxw

Ak je pôvodný text rozdelený do viacerých súborov, postup je takýto:

gandhi1.tif, gandhi2.tif, ... → gandhi1.rtf, gandhi2.rtf, ... → gandhi-2004-06-02-final

Značky používané pri korekcii textov:

Nadpis: <h1>text nadpisu</h1>

Obrázok: <picture>text</picture>

alebo

<picture>

text

</picture>

Touto značkou sa označujú rôzne obrázky, grafy, fotografie...

Text je text, ktorý bol: v rámci obrázku, názov obrázku pod ním alebo jeho popis.

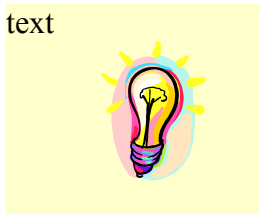
Ak je obrázok bez textu, tak ho jednoducho označíme

<picture></picture>.

Príklady:



text



text



text

Tabuľka: `<table>text</table>`

alebo

```
<table>
text
</table>
```

Touto značkou sa označujú tabuľky, zoznamy...

Príklady:

M. Polák	20 ks
P. Veľký	5 ks
M. Malý	15 ks

Názov		
txt1	txt2	txt3
txt4	txt5	txt6
txt7	txt8	txt9

```
<table>
M. Polák 20 ks
P. Veľký 5 ks
M. Malý 15 ks
</table>
```

```
<table>
Názov
txt1 txt2 txt3
txt4 txt5 txt6
txt7 txt8 txt9
</table>
```

Po oskenovaní sa opravou tabuľky už ďalej nezaobráame. Dôležité je len označenie začiatku a konca tabuľky.

Poznámka pod čiarou: `\footnote{text}`

Príklad:
Kľúčovým slovom je tu blackwhite*.

...
*V spisovnej angličtine black — čierny, white — biely, poznámka prekladateľa.

Kľúčovým slovom je tu blackwhite\footnote{V spisovnej angličtine black — čierny, white — biely, poznámka prekladateľa.}.

Súvislý inojazyčný text: `<lang kód krajiny>text</lang>`

`<lang kód krajiny>`

text
</lang>

kód krajiny:	Angličtina	EN
	Čeština	CS
	Francúzština	FR
	Nemčina	DE

Príklad:
<lang EN>This is continual English text.</lang>

Neidentifikovateľný text: <noise/>
Takto sa označuje poškodený/nečitateľný text.

Ďalšie poznámky:

- V opravovanom dokumente nezáleží na veľkosti písma a type fontu, t. j. časť textu môže byť v Times Roman a časť v Arial a podobne.
- Netreba dodržiavať zarovnanie textu (vľavo, na stred...)
- Treba dávať pozor na automatické odrážky ako: 1.
2.
alebo •. Nesmie to byť automatická odrážka. Rozlíšiť sa to dá tak, že je podfarbená šedo alebo sa nedá predňu presunúť kurzor.
- Zvýraznené písmo (kurzíva, tučné, podčiarknutie) sa ponecháva v súlade s originálom. Ak je zvýraznený text napríklad nadpisom, tak zvýraznený bude len text nadpisu, nie značky <h1></h1>. Napr. <h1>**1. kapitola**</h1>
- Treba rozoznávať rôzne typy úvodzoviek: “ ” „ “ ” ‘ ’ ‚ ‚ ‘ ‚ ‘ a pod.
- Na zápis znakov gréckej, ruskej... abecedy a iných znakov sa pokiaľ možno používajú Unicode znaky.
- Treba rozoznávať pomlčku a spojovník: spojovník –
pomlčka — (U+2014)
Tieto znaky opravíme na také, aké boli v pôvodnom texte.
- V texte sa nesmie vyskytovať znak tabelátora.
- Matematické (chemické a iné) vzorce nahradíme tagom <equation/>.
-
- Horné a dolné indexy:

Ak je horný alebo dolný index číselný, treba používať zodpovedajúce Unicode znaky, teda ⁰¹²³⁴₀₁₂₃₄. Možnosť nastavenia indexov cez formátovanie písma v textovom editore sa používa len pre nečíselné indexy, napr. „Kapitola I^{vá}“

Skenovanie periodík Sme, Smena, Práca, Pravda

Záhlavie jednotlivých článkov:

ZDROJ: Názov periodika; číslo strany
DATUM: 19941231 (teda RokMesiacDen)
RUBRIKA: napr. Titulka
Spravodajstvo
Zo sveta
Kultúra
Šport
Názory
Ekonomika
Publicistika

(názov rubriky je uvedený v paginácii každej strany)

NAZOV: Názov článku

AUTOR: Meno autora alebo jeho skratka

Tabuľky sa nerobia, do tagov <table></table> sa uvedie len názov tabuľky

Podrubriky a podnadpisy sa označujú hviezdičkou na začiatku riadku.

Príklad:

RUBRIKA: Spravodajstvo
*Stalo sa

!!!Každý článok sa ukladá do samostatného súboru s menom:

ZDROJ_DATUM_RUBRIKA_Cislo strany_Poradove cislo clanku na strane_Srdiečka

Jedno srdiečko a
Dve srdiečka b
Na nedeľu c
Ani jedno srdiečko d

Príklad:

ZDROJ: SME; s. 3

DATUM: 19990301

RUBRIKA: Spravodajstvo

NAZOV: Bulharských skialpinistov v Tatrách strhla lavína, lyžiarom robil problémy ťažký sneh na zjazdovkách

AUTOR: SME

Názov súboru bude: SME_19990301_Spravodajstvo_03_01_a.sxw

REKONŠTRUKCIA DOKUMENTOV PÍSANÝCH PÍSCÍM STROJOM

Každý dokument sa ukladá ako samostatný súbor. Napr. každá jedna zápisnica je jeden súbor. Ak je k nej priložená aj prezenčná listina, uvedie sa na konci ako tabuľka. Napr.:

<h1>Zápisnica</h1>

{text zápisnice}

<h1>Prezenčná listina</h1>

<table></table>

Ak je v dokumente text dodatočne opravovaný perom, túto opravu ignorujeme a držíme sa textu, ktorý je písaný písacím strojom. Ak je ale text opravený písacím strojom, teda ak je zlé písmenko nahradené (resp. preklepnuté) správnym, tak zachytávame správnu, upravenú podobu.

Podpisy a pečiatky ignorujeme.

Ak je napríklad nadpis zvýraznený nadpis tak, že sú medzi jednotlivými písmenami medzery, tieto odstraňujeme. Napr. text

Z Á P I S N I C A

rekonštruujeme ako: <h1>ZÁPISNICA</h1>

Ak je text podčiarknutý podčiarknutie dodržíme.

Tam kde je prečítaná číslovka 1 ako l, opravíme ju na číslo 1 podobne je často prečítaná 0 ako O, tu opravujeme tiež na správny znak..

V celom texte používame značku spojovníka - .