"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

Digitální revoluce a její důsledky pro černé řemeslo

- ☞ Digitální revoluce koncem sedmdesátých let.

  "Mathematics books and journals do not look as beautiful as they used to."

  Donald Ervin Knuth

- ☞ Počítač: od imitace psacího stroje po *sázecí* systém DTP.
- ☞ Typografie matematiky, matematika/informatika slouží typografii: místo metalurgie *informatika*, matematický popis křivek (kubické splajny v programu pro návrh písem METAFONT).

Digitální revoluce a její důsledky pro černé řemeslo (cont.)

"Go forth and create *masterpieces of electronic publishing art.*"
(paraphrase of Donald Knuth [TEXbook, p. 303])

Sázecí systémy

"GUIs normally make it simple to accomplish simple actions and impossible to accomplish complex actions."

Doug Gwyn

- **Sazba**: realizace návrhu nad značkovanými daty.
- WYSIWYG (WYSIAWYG): InDesign, Pagemaker (Adobe), QuarkXpress (Quark), 3B2 (Advent Publishing).
- Dávkové systémy: TEX, troff/groff/nroff/runoff, Lout (nutnost kontroly návrhu dokumentů ve *finální* podobě, ne jen na obrazovce).
- Textové procesory Word, AmiPro, ... sem de facto nepatří (nedovedou některé docela zásadní potřeby sazeče – nedělitelná roztažitelná mezera, fixace zlomu pro různá výstupní zařízení apod.).

Programování sazby – TEX

- Algoritmy počítačové sazby v sázecím systému TEX
  1. Zlom řádků.
  2. Zlom stránky.
  3. Dělení slov.
  4. Umisťování obrázků, viditelnost.
  5. Umisťování poznámek pod čarou.
  6. Sazba matematických výrazů:

$$\sqrt{\left(\int_0^\infty \sqrt{\frac{x^2}{2}} + 1\right)}$$

Programování sazby – TEX (cont.)

- ☞ `$$\sqrt{\left(\int_0^\infty \sqrt{\frac{x^2}{2}}+1\right)}$$`
- ☞ (Makro)programování sazby, otevřenost systému.
- ☞ OSS, CSTUG, CTAN, pros & cons.

Co je TEX?

- ☞ Sázecí autorský systém.
- ☞ Programovatelný, t.j. s vlastním makrojazykem (s vyjadřovací silou Turingova stroje – byl v něm například pro zábavu implementován interpret jazyka BASIC); výhoda pro cca 2 % populace.
- ☞ Dávkový: ze vstupu `$c=\sqrt{a^2+b^2}$` dostaneme $c = \sqrt{a^2 + b^2}$: http://tex.mendelu.cz, http://sciencesoft.at/index.jsp?link=latex.
- ☞ Otevřený/rozšiřitelný.
- ☞ Portabilní (od Atari či dvoudisketového PC XT po Cray).
- ☞ Stabilní ($256 za nalezení chyby), verze $\longrightarrow \pi$.
- ☞ Dobře dokumentovaný (vyšel knižně).
- ☞ Volně šiřitelný (vývoj hrazen granty).
- ☞ S výstupem *nezávislým* na výstupním zařízení.
- ☞ Jednoduchý základ pro sazbu: model box, glue, penalty.
- ☞ "A computer program of which a professor of computer science might be proud of." (DEK)

Co TEX není?

- Editor.
- Program na grafiku (na to slouží komplementární program METAFONT přibližně stejné velikosti).
- WYSIWYG (nadstavby jako LyX existují).
- Rychle naučitelný (strmější učicí křivka).

Vznik a vývoj TEXu

- 1977, korektura *The Art of Computer Programming*.
- 1978, první verze; TEX82, METAFONT84; osmibitový TEX 3 (1990); rozšiřování o výstupní formáty (PostScript).
- 1992: zmrazení dalšího vývoje, pevný bod, jen opravy chyb.
- Vznik makrobalíků.
- Téměř žádný marketing, uživatele sdružuje TUG a lokální sdružení uživatelů – LUG.

TEX dnes: pro klasickou publikační činnost

- Zejména pro matematiku a všude tam, kde je možná algoritmizace zpracování ($$).
- Velká nakladatelství technické literatury a časopisů Springer-Verlag, Elsevier Publishers, Kluwer sází v TEXu časopisy, sborníky.
- Databázové publikování: Dopravní podnik města Brna (tabulky jízdních řádů pro zastávky, řidiče, dispečery), rozvrhy FI MU, studijní program FI MU přímo z databáze informačního systému MU.
- Slovníky (LEDA) a první díl encyklopedie (Diderot).
- Jádro TEXu či jeho algoritmy v sázecích systémech 3B2, InDesign, troff, Lout či v textovém procesoru Word.

TEX dnes I: PDF

- Nový formát pro elektronické knihy firmy Adobe.
- Otevřenost TEXu: pdfTEX Hàn Thế Thành.
- Kalkulačka v PDF (+JavaScript): dříve než Adobe.
- AcroTEX.
- PDF není až tak portabilní jak říká název – místo pro DVI3?

TEX dnes II: Internet (HTML/XML/MATHML)

- HyperTEX (Los Alamos archiv).
- LATEX2HTML, konvertor z nejznámějšího makrobalíku do HTML v Perlu.
- TEX4ht, nebo jiné mnohé komerční implementace TEXu s alternativním výstupem do HTML jako TEXpider (MicroPress).
- eeTEX (pro přímou sazbu SGML).
- TTH: konvertor TEXu do HTML.
- IBM techplorer Hypermedia Browser – plug-in do Netscape a Internet Exploreru.
- MathType pro psaní matematiky ve Windows.

TEX dnes II: Internet (HTML/XML/MATHML) (cont.)

- TEXML (Doug Lovell, IBM): XML reprezentace TEXové syntaxe.
- WebEQ: sada programů včetně Javovského apletu pro rastrování matematiky.
- Texi2html: konverze Texinfo formátu do html.
- Word2TeX: konfigurovatelný konvertor z rozšířeného měnícího se formátu do stabilního.
- ASTER systém pro čtení LATEXového zápisu a dokumentů pro nevidomé
- a mnoho dalších . . .

TEX dnes III: Java, Unicode, OpenType

- NTS: reimplementace TEXu bez optimalizací v Javě (Karel Skoupý, FI MU, nyní ÚVT MU), a následný výzkum.
- PDF prohlížeč v Javě → vysoce portabilní prvotřídně vysazené dokumenty s minimem přenosu dat v dohledu.
- Omega: mnohem více než šestnáctibitový TEX; konverze LATEXu do MathML.
- PATLIB, opatgen, upatgen, ...
- XeTEX, viz sobotní přednáška CSTUG.

Perspektivy TEXu a jeho následníků v tomto tisíciletí

- TEX pevný bod v chaosu postmoderního světa?
- Interaktivní dokumenty: otevřenost umožňuje rychle reagovat na nové formáty a standardy.
- Rozšiřování požadavků na **autorské** systémy.
- Bude se opakovat historie s tím, že v TEXové komunitě navržené a vyzkoumané postupy, algoritmy a přístupy se stanou základem většiny komerčních aplikací v oblasti DTP?
- embedded systémy [části] TEXu součástí větších projektů.

Jak se na to připravit?

- Instalovat si TEX a učit se ho:
- učit se, učit se, učit se.
- Stát se členem sdružení CSTUG a účastnit se aktivně jeho akcí a projektů.

Závěr

"We are all apprentices in a craft
where no-one ever becomes a master."
Ernest Hemingway

- Úcta k tradicím, osvěta.
- Další studium: www.sazba.cz (Kapr: Sto a jedna věta o knižní úpravě), www.typo.cz (písma, PostScript, . . .), Beranův Typografický manuál.

"A man who would letterspace lower case would steal sheep."
Frederick Goudy, 1894–1945

"If this wisdom needs updating, it is chiefly to add that
A woman who would . . ."
Robert Bringhurst

Sazba – Algoritmy zlomu a formáty písem

- ☞ Digitální formáty písem
- ☞ Algoritmy zlomu
- ☞ hz rozšíření

Digitální formáty písem

- Formáty METAFONT, PostScript a TrueType (OpenType).
- Princip oddělení metrické informace potřebné při sazbě od informace o tvarech znaků.
- Metrické informace: rozměry jednotlivých znaků, mezipísmenný kerning, ligační programy. Formáty TFM a AFM, program afm2tfm.
- Tvary znaků jsou popsány jako bitmapy (optimalizace znaků pro obrazovku, PostScript Type 3) nebo pomocí křivek obrysů či per. PostScript Type1 používá pro popis obrysů kubické Bézierovy křivky, formát TrueType výpočetně méně náročné kvadratické Bézierovy křivky a METAFONT parametrická pera.

Digitální formáty písem (cont.)

- ☞ Detailní český popis digitálních formátů písem je k nalezení v DP PřírF MU Pavla Janíka a DP FI MU Jana Vidy.

Algoritmy řádkového a stránkového zlomu

- ☞ Algoritmus First fit: sazba po řádcích, výběr prvního možného zlomu. Nenáročné na výpočetní kapacitu.
- ☞ Algoritmus Best fit: zlom po řádcích, výběr nejlepšího zlomu na řádku bez přihlédnutí k předchozím možnostem. Míra optimality řádku je měřena mírou odchylky od ideální velikosti mezislovní mezery. V TEXu je míra škaredosti (badness) $100 \cdot |r|^3$, kde r je poměr rozdílu přirozené a požadované délky řádku ku součtu roztažitelných a stlačitelných složek lepidla (glue).

Algoritmy řádkového a stránkového zlomu (cont.)

- ☞ Algoritmus Optimum fit: optimalizace zlomu v rámci celého odstavce naráz s využitím modelu ‘box, glue a penalty’. Úloha podobná hledání nejkratší cesty v grafu, každé cestě odpovídá škaredost, která se minimalizuje. Stačí zvažovat pouze možnosti hezčí než zadaná tolerance.

Parametry algoritmu řádkového zlomu TEXu

- ☞ Až tři průchody: první bez dělení slov, limit škaredosti je `\pretolerance`. Pak jsou rozdělena všechna slova v odstavci a proveden druhý pokus s limitem škaredosti `\tolerance`.
- ☞ Pokud je nastaven registr `\emergencystretch` na kladnou délku (zvětšení roztažitelnosti mezislovní mezery) a jde o TEX3.0 či vyšší, provede se třetí pokus o zlom s těmito limity roztažitelností.
- ☞ Další parametry ovliviňující algoritmy zlomu: řádkového: `\adjdemerits`, `\linepenalty`, `\hyphenpenalty`, `\exhyphenpenalty`, `\looseness` a stránkového: `\interlinepenalty`, `\clubpenalty`, `\widowpenalty`, `\brokenpenalty`, ...

Parametry algoritmu řádkového zlomu TEXu (cont.)

- ☞ Ladění parametrů zlomu pro daný projekt (Diderot).
- ☞ Detailní český popis variant algoritmů zlomu je k nalezení v DP FI MU Jana Pazdziory.

hz-algoritmus

- ☞ Motivace: Gutenbergova 42řádková Bible měla 290 písmen abecedy proměnné šířky, stejná mezislovní mezera při sazbě do bloku.
- ☞ Principy: kerning, scaling, composing.
- ☞ URW, Karrow, Zapf, neúspěch (fonty).
- ☞ Implementace v pdfTEXu (Thành).
- ☞ Implementace v InDesignu (Karrow).
- ☞ Detailní český popis modifikací zlomového algoritmu písem je k nalezení v DP Mirky Misákové a disertaci Hàn Thế Thành̀a.

Implementace algoritmu dělení slov

- ☞ Dizertace a program `PatGen` Franka Lianga.
- ☞ Princip soutěživých vzorů, nezávislost na jazyce.
- ☞ Vytváření a generování vzorů.
- ☞ `OPatGen`, knihovna PatLib, upatgen.
- ☞ Detailní popis algoritmu dělení slov je k nalezení v DP FI MU Davida Antoše.

Principy makroprogramování

- Makrojazyk TEXu: složité sázecí postupy skládané z 300 základních příkazů (primitiv).
- (Ne)vhodnost makroprogramování (versus detailní značkování), minimalizace vstupu.
- Input procesor: řádky nezávislé na operačním systému.
- Token procesor.
- Expand procesor.
- Hlavní procesor.
- Detaily v češtině jsou v Olšákově TeXbooku naruby (TBN).

Makrobalíky

- ☞ Znovuobjevování kola většinou není potřeba: tisíce balíků maker na CTAN (Comprehensive TEX Archive Network).
- ☞ Katalog balíků na CTAN.
- ☞ Příklady balíků, kniha LATEX Companion.
- ☞ DVD s distribucí TEXlive 2013.

InDesign

- ☞ Komerční program firmy Adobe pro WYSIWYG počítačovou sazbu.
- ☞ Zlomový algoritmus (převzat z TEXu) s pomocí Davida Karrowa (URW).
- ☞ Škálování znaků.
- ☞ Optický proklad (David Kindersley: Optical letter spacing for new printing systems, 1976).
- ☞ Příklady zlomu a chyb.

Jazyky pro popis stránky

- PDL – jazyk pro popis stránky, kombinace textu s grafikou.
- Formát firmy Adobe (Geschke, Warnock, 1982, Apple Laserwriter, Aldus Pagemaker), nyní de facto standard.
- Linotronic 1986, svět osvitových jednotek (vysoké rozlišení).
- Vývoj/podpora – Level1, Level2, (Level3).
- Návazné/odvozené/související produkty/standardy: PDF, písma Type1 (vektorový popis), Type3 (bitmapy).
- Nově OpenType.

PostScript – vlastnosti jazyka

- Nezávislost na zařízení (různé rozlišovací schopnosti výstupních zařízení).
- Souřadné systémy, prostor uživatele, prostor zařízení.
- Interpreter – PostScriptový RIP, hardwarový nebo softwarový (Ghostscript, Ghostview).
- Zásobníkový programovací jazyk, postfixový zápis v reverzní polské notaci.
- Podpora různých barevných prostorů (CMYK, RGB), direktní barvy.
- Nezávislost na platformě – vytváření/generování PostScriptu aplikacemi.

PostScript – příklad

```
 1  %!PS-Adobe-2.0
 2  %%Title: Priklad Postscript programu
 3  /Helvetica-Bold findfont
 4  50 scalefont setfont
 5  0 0 moveto 200 0 lineto
 6  200 100 lineto 0 100 lineto
 7  closepath 0.8 setgray fill
 8  0 setgray
 9  30 40 moveto
10  (FI MU)show
11  showpage
```

PostScript – příklad (cont.)

☞

Struktura PostScriptového souboru

- Záhlaví, komentář Bounding Box, úrovně PostScriptu (Level).
- Definice procedur.
- Nastavení dokumentu.
- Jednotlivé stránky.
- Konec dokumentu.

PostScriptové objekty a jejich zpracování

- ☞ Prostor uživatele a prostor zařízení.
- ☞ Skládání afinních transformací (CTM, operátory pro modifikaci aktuální matice transformace: rotace, translace, zmenšení/zvětšení).
- ☞ Cesty.
- ☞ Konstrukce Bézierovy (kubické) křivky.
- ☞ Detaily v češtině jsou ve Zpravodaji CSTUG 1–3/2001.

PDF – vlastnosti

- *Proprietární* formát firmy Adobe.
- Objektový *formát*, není to programovací jazyk.
- Portabilita je dlužná svému jménu :-(. Neportabilní záložky (závisí na systémovém fontu).
- Binární formát, nelze jednoduše editovat (tabulka xref na konci PDF odkazuje na jednotlivé byty souboru).
- Struktura PDF, základní objekty, struktura souboru, popis stránky, komprese.
- Srovnání s DVI, PS.
- Možnost optimalizovat pro web, možnost vkládat JavaScript přímo do PDF.

Způsoby vytváření PDF

- pdfTEX!
- ghostscript: dávka ps2pdf.
- Adobe Acrobat Distiller.
- PDF Writer (virtuální tiskárna, minimální optimalizace).
- Adobe Acrobat Capture (OCR).
- PStill, ...

Verze PDF

- Úrovně jazyka 1.0 až 1.9 pro Acrobat 1 až Acrobat 10. Aktuální specifikace PDF 1.9 má přes tisíc stran.
- PDF-X: Ořezaná verze jazyka podporovaná ANSI, ISO a firmami v předtiskovém byznysu.
- PDF-X1, PDF-X2, PDF-X3, maximální spolehlivost při DTP a Prepress workflow (povinnost mít vložené fonty, . . .).
- PDF-A verze pro archivaci skenovaných předloh.
- PDF-ISO (upravená verze 1.7 schválená institucí ISO)
- FAQ.