"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

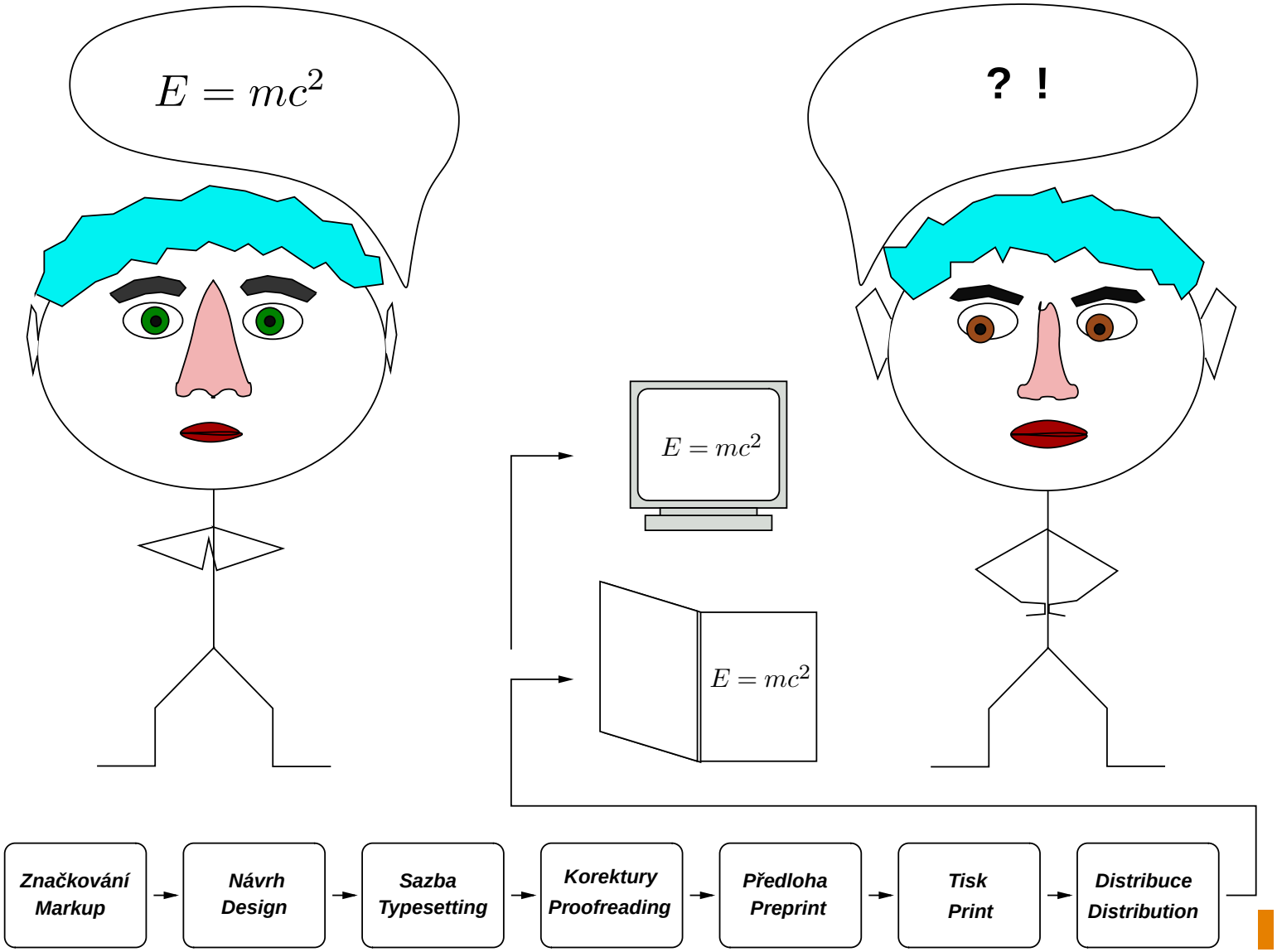
PB029 Elektronická příprava dokumentů

Petr Sojka

Fakulta informatiky
MU, Brno

podzim 2013

Autor a čtenář



Osnova dnešní přednášky

- ☞ Anketa a agenda.
- ☞ Logické a vizuální značkování.
- ☞ Značkovací jazyky.
- ☞ Formální dokumenty a principy značkování – SGML.
- ☞ Značkování v LaTeXu.

Agenda

- ☞ Zhodnocení ankety.
- ☞ Domácí úlohy – logické/vizuální značkování: noste do cvičení.
- ☞ Domácí a prémiové úlohy sledujte pravidelně na webu předmětu!
- ☞ Body a prezence na cvičení je ve vystavených záznamnících na IS (kontrolujte si).

Vizuální a logická struktura dokumentu

"Algorithms + Data = Program"
Niklaus Wirth

„Forma + Obsah = Dokument“
Petr Sojka

- ☞ Přenos informace – raison d'être dokumentu
- ☞ Obsah – stylistika, jazyková správnost, koheze
- ☞ Forma
  - ❑ jednotnost

Vizuální a logická struktura dokumentu (cont.)

- ❑ přenos informace
- ❑ struktura

$$\frac{\text{typografie}}{\text{literatura}} = \frac{\text{interpretace}}{\text{kompozice}} \text{ skladby}$$

- ☞ Obsah $\longrightarrow$ Forma; vizuální prvky musí podporovat vnitřní obsah a strukturu a být s ní konzistentní.
- ☞ Značkovací jazyky (**M**arkup **L**anguages): SGML, HTML, WML, XML, LaTeX umožňují oddělit obsah a formu tam, kde je to možné.
- ☞ Značky logické struktury $\times$ vizuální.
- ☞ Někdy oddělit nelze (Trychtýř Christiana Morgensterna).

Z hlavy autora do elektronické podoby

"Data cannot be used at a finer grain
than it is marked up at."
R. Jelliffe

- ☞ Autor, jeho interní model problematiky v hlavě a jeho (neustálá) reorganizace. Lineární zápis v časovém okamžiku formou **textu** (psaní). O
- ☞ Tentýž text může mít více interpretací.
- ☞ Pro uchopení obsahu a automatizaci zpracování je nutný **značkovací jazyk**: text je obohacen a *zjednoznačněn* **značkami**.

Z hlavy autora do elektronické podoby (cont.)

- ☞ Příklady značek: `:-)` (ze slovníčku smileys) či `<vtip>...</vtip>`.
- ☞ **Značka** je kód přidávaný k (elektronicky) vytvářenému textu, který definuje strukturu textu (**logická značka**) nebo formát textu (**vizuální značka**).
- ☞ Značka explicitně určuje interpretaci (víceznačného) textu. Dříve sazeč určoval interpretaci implicitně z kontextu a sémantiky textu (holý text v přirozeném jazyce ve strojopise je víceznačný).
- ☞ **Procedurální** (jména pro zpracování) versus **deskriptivní značkování** (jména pro kategorizaci).

Z hlavy autora do elektronické podoby (cont.)

- ☞ Výhody deskriptivního značkování:
  - ❑ Nezávislé zpracování označených dat různými způsoby a programy. Tedy například umožňuje generování různých výstupů z jednoho zdroje (databáze, dobře označkovaný text) pro elektronickou (XML, HTML, PDF, Hypercard, . . .) nebo tištěnou verzi (PS, PDF z LaTeXu).
  - ❑ Oddělení obsahu a formy (ohledně formy má rozhodující slovo nakladatel).
  - ❑ Je snazší výměna a komunikace obsahu: komunikace mezi (spolu)autory, redakcí (přes Internet).

Z hlavy autora do elektronické podoby (cont.)

- Datová nezávislost: textové soubory, dlouhodobá archivace trvalých hodnot [papír (500 let) versus bity (k nové verzi programu)].

→ *lingua franca* značkovacích jazyků?

Jaký značkovací jazyk?

- ☞ Nevýhody *proprietárních* formátů (Word, WordPerfect): účelově se mění, jsou nestabilní. Hrozí babylonské *zmatení* jazyků nebo *monopol* nevhodného formátu.
- ☞ Organizace **ISO** (International Standards Organization), vydává známé normy jako ISO 8859-2, 10646-1 (Unicode), . . .
- ☞ Norma ISO 8879:1986 Information processing – Text and office systems – **Standard Generalized Markup Language** (SGML), vydaná 15. 10. 1986 po dlouhém procesu standardizace na základě návrhu Charlese Goldfarba a jeho GML. Počátky již na konci šedesátých let při návrhu informačního systému právnických textů v IBM.

Jaký značkovací jazyk? (cont.)

- ☞ Celá sada standardů ISO: kromě SGML, DSSSL (Document Style Semantics and Specification Language, ISO/IEC DIS 10179.2:1994), SPDL (Standard Page Description Language, ISO/IEC DIS 10180:1991) a HyTime (Hypermedia/Time-based Structuring Language, ISO/IEC 10744:1992). O
- ☞ S odstupem času: nesporný úspěch SGML, základ pro další značkovací jazyky (XML), ale například neúspěch SPDL oproti proprietárnímu PostScriptu.

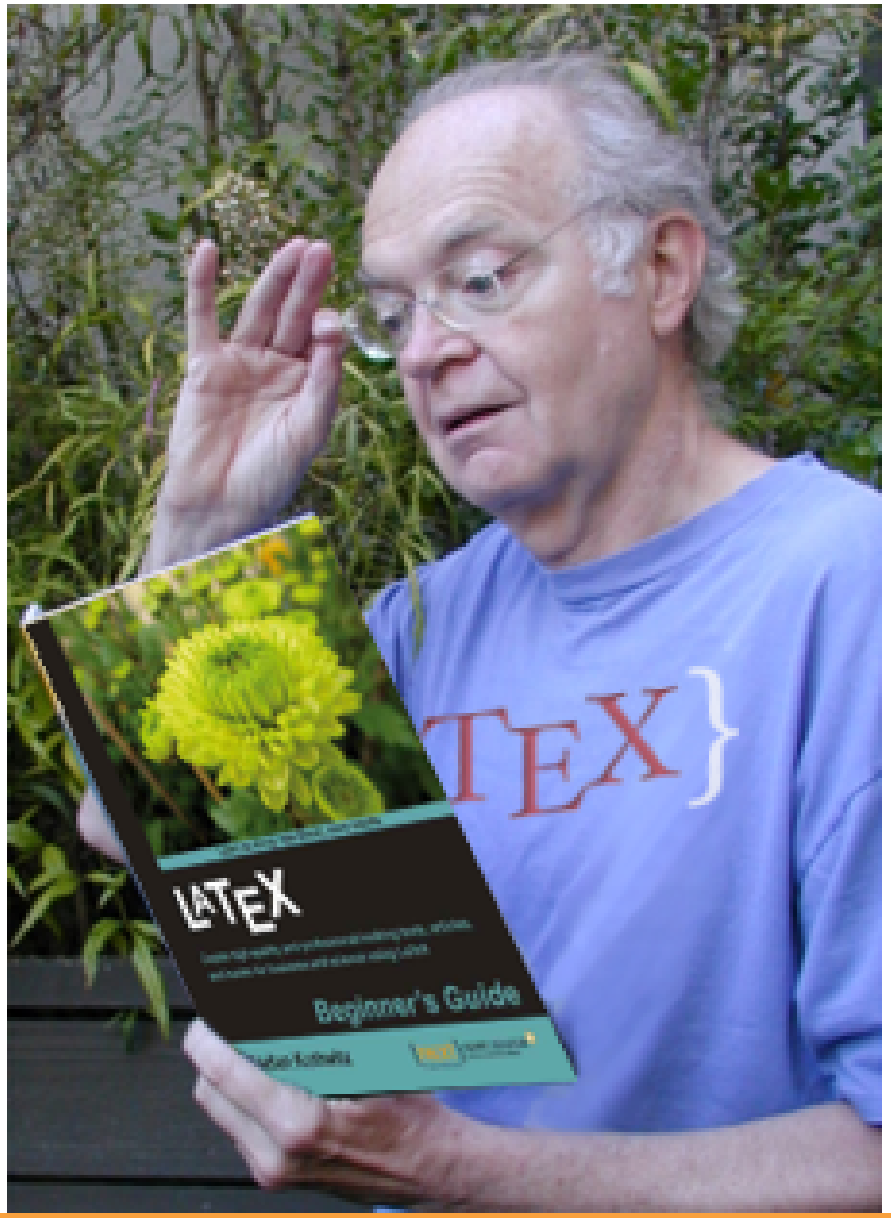
SGML – historie a motivace

- ☞ Výhody: znovupoužívání částí dokumentů, kvalita a otevřenost systému, obecnost, nezávislost na konkrétní formě (WWW konsorcium), rozšiřitelnost dle technologických možností (hypertext), snížení nákladů, možnost validace (ověření korektnosti dokumentu na základě formální definice jazyka dokumentu).
- ☞ Nevýhody: za obecnost se platí složitostí, i po dekádách užívání jsou nejlepší systémy na plné SGML drahé, změna s rozšířeními HTML, přesto přechod k XML (složitost).

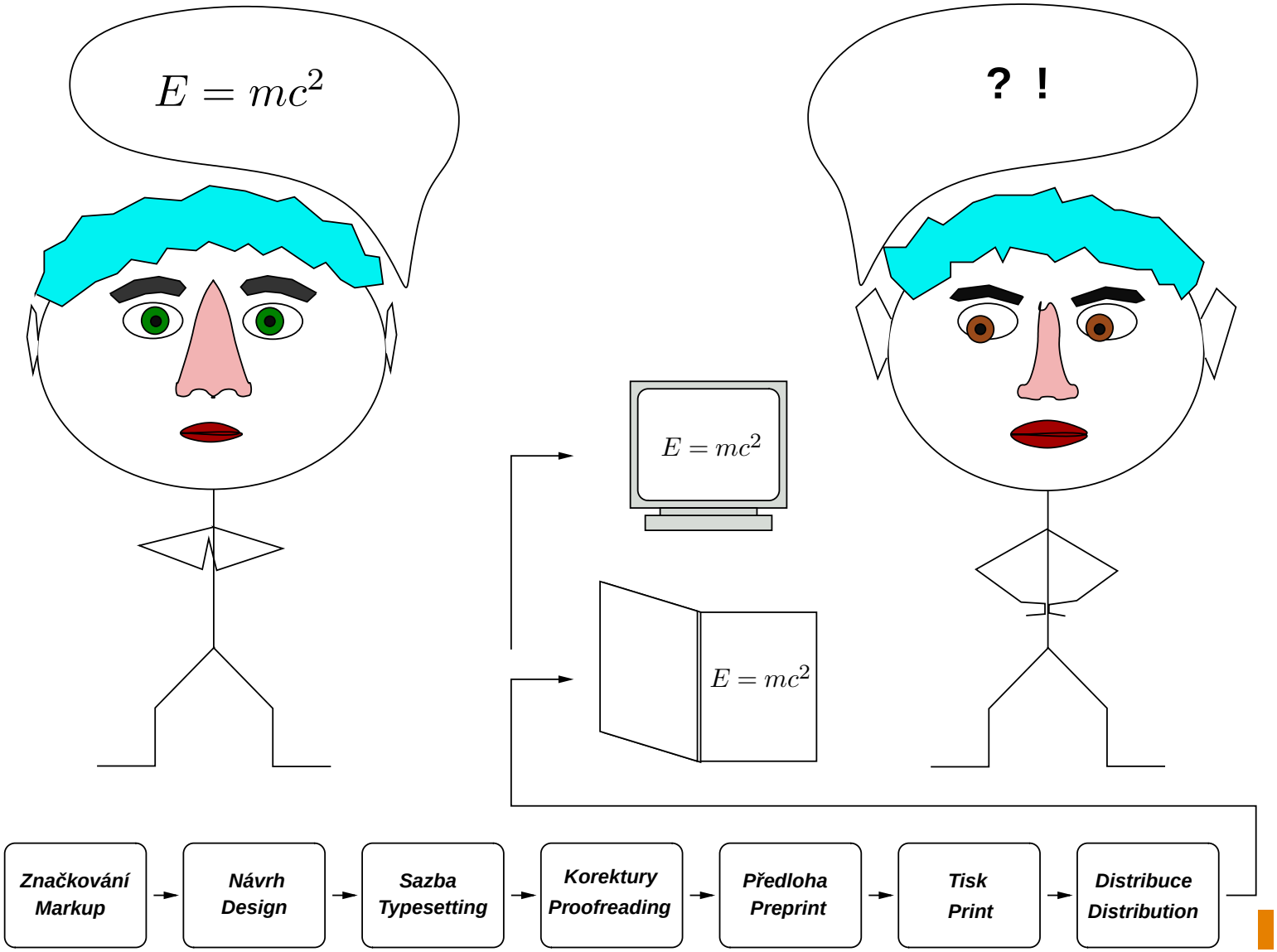
SGML – historie a motivace (cont.)

- ☞ Uživatelé SGML (zpočátku velké instituce, elektronický a letecký průmysl, nakladatelé): AAP, OUP, Boeing, Lufthansa, British Patent Office, Association of German editors, TEI, McGraw-Hill, LDC, . . . viz stránky Organization for the Advancement of Structured Information Standards.
- ☞ Dnešní motivací je přesnější a *relevantnější* vyhledávání na Internetu.

LaTeX jako značkovací jazyk TeXu



Autor a čtenář



WYSIWYG versus dávkové zpracování dokumentů

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

Doug Gwyn

- **Sazba**: realizace návrhu nad značkovanými daty.
- Krok **návrh**u u WYSIWYG často chybí.
- Podstatou je účel, míra interakce.
- Krok návrhu u WYSIWYG často chybí.
- WYSIWYG (WYSIAWYG): InDesign, Pagemaker (Adobe), QuarkXpress (Quark), 3B2 (Advent Publishing).

WYSIWYG versus dávkové zpracování dokumentů (cont.)

- ☞ 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).

Co je TEX? (cont.)

- ☞ 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)
- ☞ Optimalizovaný, vysoce efektivní, využívající nejrychlejší algoritmy své doby – při vývoji nalezeny nové informatické metody a datové struktury (trie).

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.