



Konverze Zpravodaje ČSTUGu pro Českou digitální matematickou knihovnu (DML-CZ)

Co, proč a jak?

Vítek Novotný

witiko@mail.muni.cz

Valné shromáždění ČSTUGu

6. března 2021 v Brně

Co?

- Česká digitální matematická knihovna (DML-CZ) archivuje československé matematické texty.
- Na výjezdním zasedání v roce 2019 DML-CZ nabídla archivovat [Zpravodaj ČSTUGu](#).
- Za realizaci odpovídá [Vítek Novotný](#) jako člen DML-CZ a technický redaktor ČSTUGu.



Czech Digital Mathematics Library

[About DML-CZ](#) | [FAQ](#) | [Conditions of Use](#) | [Math Archives](#) | [Contact Us](#)

Search

[Advanced Search](#)

Browse

[DML-CZ Home](#) >

Welcome to **DML-CZ** (Czech Digital Mathematics Library), the website offering an open access to the metadata and fulltext of mathematical journals, proceedings and books published throughout history in the Czech lands.

Digitální matematické knihovny [7, sekce 2]

Česká digitální matematická knihovna (DML-CZ)

- DML-CZ je evropský projekt (2005–2009) následujících partnerských organizací:
 1. Matematický ústav a knihovna Akademie věd České republiky (AV ČR)
 2. Ústav výpočetní techniky a Fakulta informatiky Masarykovy univerzity (MUNI)
 3. Matematicko-fyzikální fakulta Univerzity Karlovy (UK)
- Mnozí členové týmu DML-CZ jsou zároveň současnými nebo bývalými členy C5TUGu:

1. Vítek Novotný (MUNI)	3. Michal Růžička (MUNI)	5. Jiří Veselý (UK)
2. Petr Sojka (MUNI)	4. Oldřich Ulrych (UK)	6. ...
- DML-CZ zajišťuje akvizici (AV ČR), digitalizaci (UK, MUNI) [1, 2], archivaci (MUNI) [3] a vyhledávání (MUNI) [4, 5, 6] nad československými matematickými texty.

Evropská digitální matematická knihovna (EuDML)

- EuDML je evropský projekt (2010–2013) MUNI a dalších 11 partnerských organizací.
- EuDML nabízí textové a matematické vyhledávání nad DML-CZ a dalšími 13 knihovnami.
- Matematické vyhledávání vyvíjí skupina Math Information Retrieval (MIR) na MUNI.

Evropská digitální matematická knihovna (EuDML) I

EuDML | The EUROPEAN DIGITAL
MATHEMATICS LIBRARY

čeština (cs) ▾

[Přihlásit se](#) | [Registrovat se](#) | [\(Proč se registrovat?\)](#)

Název, autor, klíčová slova, citace, datum,

Hledat

[Domů](#)

[Rozšířené vyhledávání](#)

[Procházet dle MSC](#)

[Procházet časopisy](#)

[Dohledání referencí](#)

Hledat

Začněte zadáním hledaných pojmů

Název, autor, klíčová slova, citace, datum, ...

Hledat

Rozšířené vyhledávání ➔

`</>` Embed **EuDML Search** on Your Website

Tipy pro vyhledávání

- při vyhledávání se nebere ohled na velikost písmen a diakritiku (**Bézout** = **bezout**)
- slova se hledají přesně tak, jak jsou zapsána (**theorem** ≠ **theorems**)
- sousloví je možné vyhledávat užitím uvozek ("Uniformization theorem" ≠ Uniformization theorem = uniformization **AND** theorem)
- mimo sousloví v uvozkách můžete používat zástupné „žolíkové“ znaky * a ?

EuDML nyní indexuje 268082 položek ve 14 kolekcích. [více statistik](#)

Evropská digitální matematická knihovna (EuDML) II

$E = mc^2$

Formula preview

$$E = mc^2$$

Hledat

Nyní zobrazeny záznamy 1– 20 z 89002

Zobrazit záznamů na stránku

Seřadit podle [relevance](#) | [názvu](#) | [roku vydání](#)

[De \$E = mc^2\$ à l'équation de Dirac : une introduction heuristique aux spineurs](#)

[Luc Rozoy](#)

Séminaire de théorie spectrale et géométrie

[\$E = mc^2\$ aneb dovětek o Einsteinově dovětku](#)

[Jiří Podolský](#) — 2005

Pokroky matematiky, fyziky a astronomie

Jazyky

en	50829
fr	19427
cs	7705
it	2963
de	1830

Více... ➔

Typy dokumentů

Proč?

ČSTUG ve službách matematiky

- \TeX vznikl v 70. letech, protože Donald Knuth nebyl spokojený s fotosazbou matematiky.
- O půl století později je \TeX hlavní technologií pro přípravu a digitalizaci matematiky:
 1. Podpora matematických symbolů ve standardu Unicode vychází z maker \LaTeX u. [8]
 2. Lze oboustranně převádět mezi \TeX em a XML formátem MathML, který lze mj. zobrazit na webu, zpracovávat důkazovými systémy a použít k matematickému vyhledávání [9, sekce 2.1].
- ČSTUG financuje vývoj matematických fontů \TeX Gyre polským GUSTem, vzdělává novou generaci matematiků a zajišťuje podporu pro českou matematickou sazbu. [10]

Matematika ve službách ČSTUGu

- Zveřejnění Zpravodaje ČSTUGu v EuDML zajistí dlouhodobou archivaci a evropský dosah.
- Textové a matematické vyhledávání zpřístupní zapomenuté texty českých \TeX istů.

A jak?

- DML-CZ od časopisů přebírá dokumenty a XML metadata ve **standardním formátu**.
- Dokumenty přebírané DML-CZ se řadí do tří kategorií:

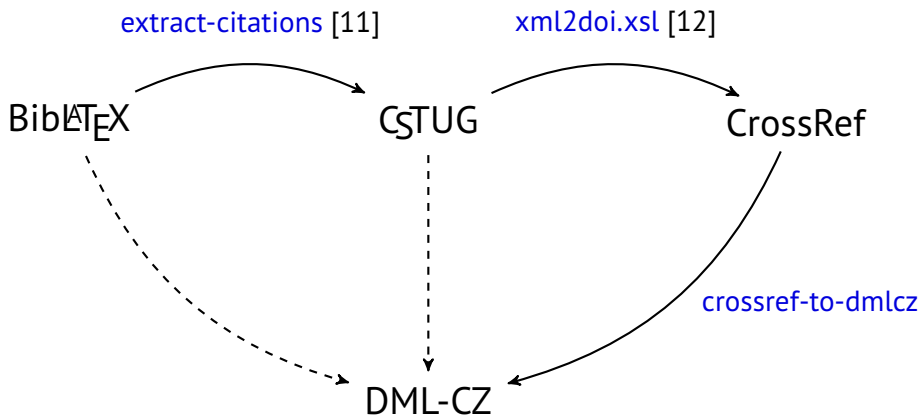
Retro-digitalizované Tištěné dokumenty, které tým DML-CZ **skenuje a digitalizuje**.

Retro-born-digital Digitalizované dokumenty bez zdrojových textů v $\text{T}_{\text{E}}\text{X}$ u.

Born-digital Digitalizované dokumenty včetně zdrojových textů v $\text{T}_{\text{E}}\text{X}$ u.

- Ročníky Zpravodaje 1991 až 2011 jsou retro-born-digital a od 2012 born-digital.
- Pro všechny ročníky Zpravodaje máme dostupná XML metadata:
 - Pro čísla 1991-1 až 2001-1 máme pouze CrossRef XML bez referencí a editorialů.
 - Pro čísla 2001-1 až 2011-1 máme pouze CrossRef XML se surovými referencemi.
 - Pro čísla 2011-1 až 2017-1–2 máme CrossRef XML a ζ TUG XML stále bez editorialů.
 - Od čísla 2017-1–2 máme CrossRef XML, ζ TUG XML a Bib $\text{A}_{\text{T}}\text{E}_{\text{X}}$ XML s editorialem.

Konverze mezi metadatovými formáty XML



Závěr

- Vyvinul jsem v jazyce Python nástroj [crossref-to-dmlcz](#).
- Nástrojem jsem [zkonvertoval](#) ročník 1991 z CrossRef XML do DML-CZ XML.
- Reference jsem doplnil ručně, u novějších ročníků lze využít bohatší metadata.
- Editorialy jsem doplnil ručně, od čísla 2017-1–2 lze využít bohatší metadata.
- Bez doplnění referencí a editorialů lze zkonvertovat za den, poctivě za člověkotýden.

vyhledávání | statistika | autoři | verifikace | násověda | správa witiko (odhlásit)

Metadatový editor

ME / serial / Zpravodaj Československého sdružení uživatelů TeXu / Volume 001 /

Svazek 001 (1991) << | < | > | >>

[upravit svazek](#) | [správa](#)

Status: completed
Číslo: 001
Datum: 1991

- [Číslo 1 \[1\]](#) (1991)
- [Číslo 2 \[2\]](#) (1991)
- [Číslo 3 \[3\]](#) (1991)
- [Číslo 4 \[4\]](#) (1991)

● Zpravodaj Československého sdružení uživatelů TeXu (1991)

[upravit metadata](#) | [správa](#)

<< | < | > | >>

Status: completed

Pořadové číslo vydání: 1

Jméno: 1

Datum: 1991

Název: Zpravodaj Československého sdružení uživatelů TeXu (cze)

Paper size: 148x210mm



Články čísla

● #1 Několik slov úvodem...	1-1
● #2 Co nového v „TeXworldu“...	2-2
● #3 Budoucnost TeXu a METAFONTu	2-3
● #4 Komentář k budoucnosti TeXu a METAFONTu	3-10
● #5 ASCII Editor pro TeX	10-12
● #6 Textový editor CSED	12-16
● #7 Popis textového editoru PragoText	16-18
● #8 DANTE	18-20
● #9 Co je nového s šířenou verzí TeXu pro češtinu?	20-23
● #10 Stanovy Československého sdružení uživatelů TeXu	23-27

	<p>Název: Několik slov úvodem... Autor: Veselý, Jiří Jazyk: cze Typ článku: editorial Stránky: 1-1</p>
	<p>Název: Co nového v „TeXworldu“... Autor: Veselý, Jiří</p>

Komentář k budoucnosti TeXu a METAFONTu

<< | < | > | >>

[upravit](#) | [správa](#) << | < | > | >>**Název:** Komentář k budoucnosti TeXu a METAFONTu (cze)**Status:** completed**Autor:** Beebe, Nelson H. F.**Jazyk:** cze**Rok:****Zdroj:****Informace o publikování:****Copyright:****Typ článku:** Informatics**Typ práce:****Externí URL:****Externí poznámka:****Stránky:** 3-10**idMR:****idZBL:****idJFM:****DOI:** 10.5300/1991-1/3**Odkaz:****Odkazy:****Odkazy na literaturu:**

- [1] Barnhart, Elizabeth: *TeX in Commercial Environment Setting Multi-Column Output*, TUGboat, 8(2):185-189, July 1987.
- [2] Becker, Zeev; Berry, Daniel: *Triroff, an Adaptation of the Device-Independent Troff for Formatting Tri-Directional Text*, Electronic Publishing—Origination, Dissemination, and Design, 2(3):119-142, October 1989.
- [3] Beebe, Nelson H. F.: *TeX and Graphics: The State of the Problem*, Cahiers GUTenberg, 1(2):13-53, March 1989. Pretended to: Congrès GUTenberg, Paris, France, 16-17 May 1989.
- [4] Ferguson, Michael: *A Multilingual TêX*, TUGboat, 7(1):16, March 1986.
- [5] Ferguson, Michael: *A (Hopefully) Final Extension of Multilingual TêX*, TUGboat, 8(2):102-103, July 1987.
- [6] Ferguson, Michael: *Coordination of Non-English Use of TeX*, TUGboat, 11(1):8-9, April 1990.
- [7] Ferguson, Michael: *A Multilingual TêX*, TUGboat, 6(2):57-58, July 1985.
- [8] Hobby, John D.: *A METAFONT-Like System with PostScript Output*, TUGboat, 10(4):505-512, December 1989.
- [9] Knuth, Donald E.; MacKay, Pierre: *Mixing Right-to-Left Texts with Left-to-Right Texts*, TUGboat, 8(1):14-25, April 1987.
- [10] Lamport, Leslie: *LaTeX—A Document Preparation System—User's Guide and Reference Manual*, Addison-Wesley, 1985.
- [11] Liang, Franklin Mark: *Word Hy-phen-ation by Com-pu-ter*, PhD thesis, Stanford University, August 1983.
- [12] Mittelbach, Frank: *E-TeX: Guidelines for Future TeX Extensions*. TUGboat. 11(3):337-345. September 1990.

Děkuji vám za pozornost.

Bibliografie I

- [1] Tomáš Mudrák. „Digitalizace matematických textů“. Dipl. Fakulta informatiky Masarykovy univerzity, 2006. URL: <https://is.muni.cz/th/n991g> (cit. 05.03.2021).
- [2] Tomáš Panák. „Digitalizace matematických textů“. Dipl. Fakulta informatiky Masarykovy univerzity, 2006. URL: <https://is.muni.cz/th/pspz5> (cit. 05.03.2021).
- [3] Vlastimil Krejčíř. „Building the Czech Digital Mathematics Library upon DSpace System“. eng. In: *DML 2008 : Towards Digital Mathematics Library*. Ed. Petr Sojka. Brno: Masaryk University, 2008, s. 117–126. ISBN: 978-80-210-4658-0.

Bibliografie II

- [4] Radim Řehůřek a Petr Sojka. „Software Framework for Topic Modelling with Large Corpora“. In: *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Valletta, Malta: ELRA, květ. 2010, s. 45–50. ISBN: 2-9517408-6-7. URL: <http://is.muni.cz/publication/884893/en> (cit. 05.03.2021).
- [5] Petr Sojka, Michal Růžička a Vít Novotný. „MlaS: Math-Aware Retrieval in Digital Mathematical Libraries“. In: *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. CIKM '18. Torino, Italy: Association for Computing Machinery, 2018, s. 1923–1926. ISBN: 9781450360142. DOI: 10.1145/3269206.3269233.
- [6] Vít Novotný. *Math-Aware Search Engine in a Single Line of Code*. 18. ún. 2021. URL: <https://is.muni.cz/el/fi/podzim2020/PV212/index.qwarmp?prejit=6501377> (cit. 05.03.2021).

Bibliografie III

- [7] Vít Novotný. „Document Representations for Fast and Accurate Retrieval of Mathematical Information“. Rigorózní práce. Brno: Fakulta informatiky Masarykovy univerzity, 2020. URL: <https://is.muni.cz/th/x86jd> (cit. 05.03.2021).
- [8] Barbara Beeton, Asmus Freytag a Murray Sargent III. *Unicode Support for Mathematics*. Tech. zpr. URL: <http://www.unicode.org/reports/tr25/> (cit. 05.03.2021).
- [9] Vít Novotný et al. „Three is Better than One: Ensembling Math Information Retrieval Systems“. eng. In: *CEUR Workshop Proceedings 2020* (2020). ISSN: 1613-0073. URL: http://ceur-ws.org/Vol-2696/paper_235.pdf (cit. 05.03.2021).
- [10] David Nečas. *Yetiho typografický bestiář*. 19. květ. 2003. URL: <https://www.physics.muni.cz/~yeti/tex/bestiary.pdf> (cit. 05.03.2021).

Bibliografie IV

- [11] Vít Novotný. „Příprava Zpravodaje C_STUG“. In: *Zpravodaj C_STUG* 28.1–4 (2018), s. 1–10. ISSN: 1211-6661. DOI: 10.5300/2018-1-4/1.
- [12] Zdeněk Wagner. „Zkušenosti ze sazby Zpravodaje Československého sdružení uživatelů T_EXu“. In: *Zpravodaj C_STUG* 10.4 (2000), s. 113–122. ISSN: 1211-6661. DOI: 10.5300/2000-4/113.

MUNI

FAKULTA

INFORMATIKY