<?xml version="1.0" encoding="utf-8"?> <TEI xmlns="http://www.tei-c.org/ns/1.0"><teiHeader>...<body>

K vybraným projektům digitalizace a publikace jazykových dat

Boris Lehečka, boris@daliboris.cz

Mezioborové kolokvium Historický text a jeho interpretace

Kostelec nad Černými lesy 20. až 22. října 2021

CC BY SA

</body></TEI>

BL

<?xml version="1.0" encoding="utf-8"?> <TEI xmlns="http://www.tei-c.org/ns/1.0"><teiHeader>...<body>

Osnova

- BL
- DJAK
- LeDIIR
- DL4DH

</body></TEI>

CC BY SA

<?xml version="1.0" encoding="utf-8"?> <TEI xmlns="http://www.tei-c.org/ns/1.0"><teiHeader>...<body>

BL: Boris Lehečka

- boris@daliboris.cz
- zaměření na digitalizaci, analýzu a prezentaci jazykových dat
  - textů, slovníků
- od 90. let 20. století práce v oddělení vývoje jazyka ÚJČ
- programátor Vokabuláře webového
- v letech 2016–2019 řešitel *Výzkumné infrastruktury po diachronní bohemistiku* (RIDICS)
- místopředseda České asociace pro digitální humanitní vědy (CzADH)
- 2020–2021 programátor v Syntea, a. s.
- od léta 2021 konzultant a programátor na volné noze (projekty AV ČR)

</body></TEI>

<?xml version="1.0" encoding="utf-8"?> <TEI xmlns="http://www.tei-c.org/ns/1.0"><teiHeader>...</body>

DJAK: Dílo Jana Amose Komenského

- součást výzkumné infrastruktury LINDAT/CLARIAH-CZ
  - projekt č. LM2018101 a CZ.02.1.01/0.0/0.0/16_013/0001781
  - Tomáš Havelka (FlÚ); Kira Droganova (LINDAT), Markéta Jelenová (grafika)
- zveřejnění digitalizovaných svazků Díla Jana Amose Komenského
  - pilotní práce na 3. svazku (Spisy útěšné, publicistické a informační z let 1617–1660; Truchlivý, Listové do nebe, Kšaft, Labyrint ap.)
  - zpřístupnění
    - textu pramene
    - textově-kritického aparátu
    - biblických míst
    - edičních poznámek a komentářů
    - faksimile pramenů
  - nástroje pro výběr pramene
    - na základě metadat o jazyce díla, místa vydání, datu publikace, editorovi ap.

</body></TEI>

CC BY SA

<?xml version="1.0" encoding="utf-8"?> <TEI xmlns="http://www.tei-c.org/ns/1.0"><teiHeader>...<body>

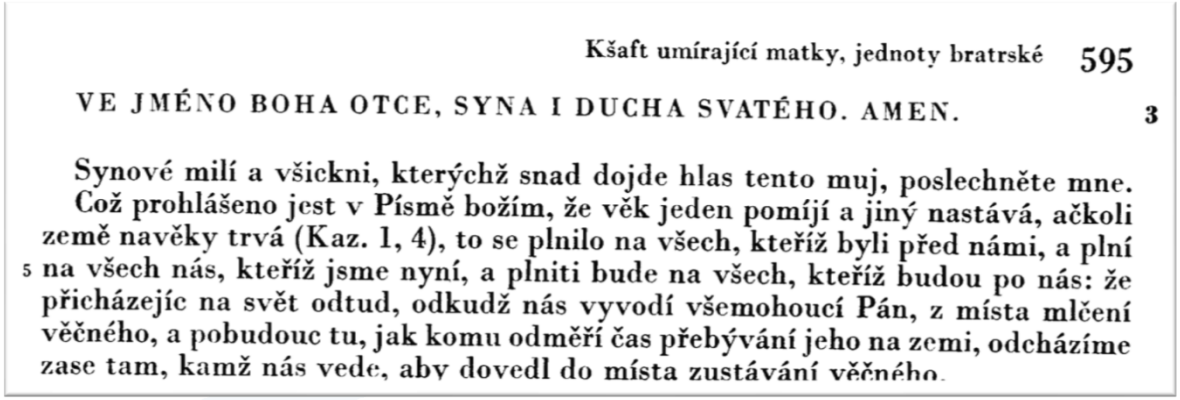
DJAK: Dílo Jana Amose Komenského / II

- použité technologie
  - formát dat TEI P5: *Guidelines for Electronic Text Encoding and Interchange* (https://tei-c.org/release/doc/tei-p5-doc/en/html/index.html)
  - oXygen XML Editor (aplikace pro editaci a validaci XML)
  - platforma pro publikaci: TEI Publisher (https://teipublisher.com)
- poustup prací
  - OCR a kolace s originálem (ještě před začátkem projektu DJAK)
  - poloautomatická konverze z formátu ODT (OpenOffice Writer) do TEI
    - přiřazení různočtení a komentářů k odpovídající pasáži
    - ruční oprava problematických míst

</body></TEI>

CC BY SA

Tisk DJAK



Kšaft umírající matky, jednoty bratrské 595

VE JMÉNO BOHA OTCE, SYNA I DUCHA SVATÉHO. AMEN. 3

Synové milí a všickni, kterýchž snad dojde hlas tento muj, poslechněte mne. Což prohlášeno jest v Písmě božím, že věk jeden pomíjí a jiný nastává, ačkoli země navěky trvá (Kaz. 1, 4), to se plnilo na všech, kteříž byli před námi, a plní na všech nás, kteříž jsme nyní, a plniti bude na všech, kteříž budou po nás: že přicházejíc na svět odtud, odkudž nás vyvodí všemohoucí Pán, z místa mlčení věčného, a pobudouc tu, jak komu odměří čas přebývání jeho na zemi, odcházíme zase tam, kamž nás vede, aby dovedl do místa zustávání věčného.

Textový editor

SIGLA
Škarka vydání Antonína Škarky v: Jan Amos Komenský, Dvojí poselství k českému národu, Praha 1970**<PE_595>|<PO_3>**

VE JMÉNO BOHA OTCE, SYNA I DUCHA SVATÉHO. AMEN.
Synové milí a všickni, kterýchž snad dojde hlas tento muj, poslechněte mne.
Což prohlášeno jest v Písmě božím, že věk jeden pomíjí a jiný nastává, ačkoli země navěky trvá (Kaz. 1, 4), to se plnilo na všech, kteříž byli před námi, a plní na všech nás, kteříž jsme nyní, a plniti bude na všech, kteříž budou po nás: že přicházejíc na svět odtud, odkudž nás vyvodí všemohoucí Pán, z místa mlčení věčného, a pobudouc tu, jak komu odměří čas přebývání jeho na zemi, odcházíme zase tam, kamž nás vede, aby dovedl do místa zustávání věčného

Formát TEI

```
  <head>VE JMÉNO BOHA OTCE, SYNA I DUCHA SVATÉHO. AMEN.</head>
 </div>
 <div>
  <p>Synové milí a všickni, kterýchž snad dojde hlas tento muj, poslechněte mne.</p>
  <p>Což prohlášeno jest v Písmě božím, že věk jeden pomíjí a jiný nastává, ačkoli země navěky trvá (<ref
type="canon" subtype="Bible">Kaz. 1, 4</ref>), to se plnilo na všech, kteříž byli před námi, a plní na všech nás,
teříž jsme nyní, a plniti bude na
    všech, kteříž budou po nás: že přicházejíc na svět odtud, odkudž nás vyvodí všemohoucí Pán, <anchor
ml:id="djak3.ksaft.a-2"/>z místa mlčení věčného,<app from="#djak3.ksaft.a-2">
     <note place="bottom" type="gloss"><label>z místa mlčení věčného </label> Srov. <!-- tady by se taky mohlo
značit biblické místo, i když je to v metatextu --> Ps 115 (114), 17; k tomuto básnickému pojmenování člověkova
východiště před příchodem na svět
      vytváří dále <name>Komenský</name> v duchu biblické dikce také pojmenování protějšku, cíle, kam člověk po
ivotní pouti odchází: do místa zustávání věčného.</note>
    </app> a pobudouc tu, jak komu odměří čas přebývání jeho na zemi, odcházíme zase tam, kamž nás vede, aby
ovedl do místa zustávání věčného.</p>
   <p>A <anchor xml:id="djak3.ksaft.a-3"/>jakýž v tom způsob jest každého člověka<app from="#djak3.ksaft.a-3">
     <note place="bottom" type="gloss"><label>jakýž v tom způsob jest každého člověka </label>jak se v tom směru
ěje každému člověku</note>
    </app>
    <anchor xml:id="djak3.ksaft.a-4"/>v osobě vlastní,<app from="#djak3.ksaft.a-4">
     <note place="bottom" type="gloss"><label>v osobě vlastní </label>jako jednotlivci</note>
    </app> takovýž jest i každé společnosti lidské v domích, městech, královstvích a církvech: že všudy jedno
omíjí, druhé nastává. I nyní co se pod nebem děje, vidíte. Pomíjejí a proměnu berou království některá a v nich
árodové, jazykové, práva, ,<!-- hodnota @break="no" se používá v případě, že je hranice uprostřed slova; tady
de o případ @break="yes" -->
    <pb n="4" ed="T1650" break="no"/> náboženství; proto bez pochyby, že nastává jiný věk. Pomíjejí i církve a
ednoty; proto nepochybně, že obnoviti chce Bůh tvář země své (<ref type="canon" subtype="Bible">žalm 104,
```

<?xml version="1.0" encoding="utf-8"?> <TEI xmlns="http://www.tei-c.org/ns/1.0"><teiHeader>...<body>

</body></TEI>

CC BY SA

<?xml version="1.0" encoding="utf-8"?> <TEI xmlns="http://www.tei-c.org/ns/1.0"><teiHeader>...<body>

DJAK: Dílo Jana Amose Komenského / III

- postup prací
  - problematická místa (⇒ ruční opravy)
    - nejednoznačný začátek a konec komentovaného úseku
    - odlišné ediční zásady
    - marginálie (více řádků)
  - další pomůcky
    - nástroje pro generování čtení jednoho pramene
    - vygenerování seznamu digitálních kopií a jejich přiřazení k paginaci v edici
  - následující fáze
    - spolupráce s grafičkou
    - implementace nezbytných funkcí v TEI Publisheru
- další plány
  - testovací verze ke konci roku 2021
  - po vyřešení autorských práv zpřístupnění veřejnosti

</body></TEI>

CC BY SA

LeDIIR: Elektronická lexikální databáze indoíránských jazyků. Pilotní modul perština.

- projekt TA ČR reg. č. TL03000369
  - hlavní řešitelka Darina Vystrčilová (Sociologický ústav), Mona Khademi; Ľubomír Novák, Zuzana Kříhová (FF UK)
- předpokládaný uživatel
  - od laika (turisty) přes překladatele po pedagogy
- principy
  - výběrový dokladový slovník založený na korpusech
  - vychází z Česko-perského slovníku D. Vystrčilové
  - slovník vzniká v programu FLEx (FieldWorks Language Explorer; (https://software.sil.org/fieldworks/)
  - transformace z formátu LIFT do formátu TEI Lex-0
  - publikace na webu s využitím TEI Publisheru (https://teipublisher.com)
  - mobilní aplikace pro Android a iPhone (https://software.sil.org/dictionaryappbuilder/)

CC BY SA

<?xml version="1.0" encoding="utf-8"?> <TEI xmlns="http://www.tei-c.org/ns/1.0"><teiHeader>...<body>

LeDIIR: Elektronická lexikální databáze indoíránských jazyků. Pilotní modul perština. / II

- FieldWorks Language Explorer (FLEx)
  - aplikace Windows pro zpracování lexikografických dat
  - primárně určeno pro minimálně zdokumentované jazyky
  - nízké náklady (zdarma)
  - vhodné pro spolupráci více autorů (sdílení přes internet, lokální síť, USB)
    - ne vždy je synchronizace bezproblémová
  - podpora velkého množství jazyků (písma, abecední řazení)
  - velké množství slovníkových částí
    - varianty, výslovnost, doklady, morfologická stavba, sémantické kategorie
  - velké množství funkcí
    - hromadné úpravy, vlastní kategorie, export do různých formátů, tiskový výstup

</body></TEI>

CC BY SA

LeDIIR: Elektronická lexikální databáze indoíránských jazyků. Pilotní modul perština. / III

- TEI Lex-0
  - https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html
  - TEI Lex-0 = odnož TEI věnovaná specificky slovníků
  - snaha o co největší kompatibilitu a sdílení dat
  - zjednoznačnění některý pravidel
  - některé úpravy přecházejí z TEI Lex-0 do TEI
  - stále se vyvíjí, je možno jej usměrňovat (na základě příkladů)
- konverze do formátu TEI
  - pomocí programovacího jazyka XProc 3.0
    - programovací jazyk (převážně) pro XML zapsaný v XML
    - MorganaXProc-III (https://www.xml-project.com/morganaxproc-iii/) pro spuštění programu
    - kontrola formální správnosti, odstranění odkazu na nerevidovaná hesla, seskupení, seřazení

CC BY SA

<?xml version="1.0" encoding="utf-8"?> <TEI xmlns="http://www.tei-c.org/ns/1.0"><teiHeader>...<body>

LeDIIR: Elektronická lexikální databáze indoíránských jazyků. Pilotní modul perština. / IV

- ukázka uživatelského rozhraní
- vybrané problémy
  - perština
    - abecední řazení, znaky s diakritickými znaménky, mezera s nulovou šířkou
    - použití interpunkce (tečka z latinky a perštiny)
  - kombinace čtení zleva doprava a zleva doprava
  - variantní podoby heslových slov
    - generují se automaticky, před uživatelem skryté, ale nalezitelné
  - redakční opomenutí
    - chybějící ekvivalent k dokladu

</body></TEI>

CC BY SA

<?xml version="1.0" encoding="utf-8"?> <TEI xmlns="http://www.tei-c.org/ns/1.0"><teiHeader>...<body>

DL4DH: Digital Libraries for Digital Humanities

- projekt NAKI, reg. č. DG20P02OVV002
  - Knihovna AV ČR, Moravská zemská knihovna v Brně, Národní knihovna ČR
  - Martin Lhoták, Petr Žabička, Pavel Straňák, Zdenko Vozár, David Novák, Radim Hladík…
  - InQool (https://inqool.cz; implementace)
- cíl projektu
  - obohacení textových dat a přístup pro jejich vytěžování (uživatelské i programové rozhraní)
  - data, která jsou součástí projektu Kramerius (http://www.digitalniknihovna.cz)
- dostupná data
  - údaje o digitalizaci (kdy, co, čím [skener, OCR], velikost obrázku, rozlišení atp.)
  - text digitalizátu (OCR) včetně jeho grafických kvalit (záhlaví, kurzíva, odstavce)

</body></TEI>

DL4DH: Digital Libraries for Digital Humanities / II

- formát výstupních dat
  - TEI
  - JSON
  - TSV/CSV (hodnoty oddělené tabulátorem/čárkou)
- obohacení dat (prostřednictvím služeb LINDAT/CLARIAH-CZ)
  - UDPipe (http://lindat.mff.cuni.cz/services/udpipe/)
    - lemmata, morfologické kategorie tokenů
  - NameTag (http://lindat.mff.cuni.cz/services/nametag/)
    - rozpoznání entit (jména osob, geografické názvy, instituce, adresy, čísla aj.)
- vaše náměty na využití dat?

CC BY SA

<?xml version="1.0" encoding="utf-8"?> <TEI xmlns="http://www.tei-corg/ns/1.0"><teiHeader>...<body>

DL4DH: Digital Libraries for Digital Humanities / III

- ukázka obohacení dat
  - v programovacím jazyce XProc 3.0, MorganaXProc-III
  - stažení OCR jednotlivých stran, zaslání textu na služby UDPipe a NameTag, převod výstupu do formátu TEI, spojení stran do jednoho souboru, vytvoření hlavičky TEI z metadat o dokumentu
- ukázka obohacení dat
  - obohacení dat pomocí testovací aplikace „Kramerius+“ (pracovní název)
- vybrané problémy
  - odlišnost dat ve formátu TEI a JSON
  - citlivost externích nástrojů (NameTag, UDPipe) na vstup (kontinuální text, vertikála)

</body></TEI>

CC BY SA

<?xml version="1.0" encoding="utf-8"?> <TEI xmlns="http://www.tei-c.org/ns/1.0"><teiHeader>...<body>

Děkuju za pozornost, komentáře, dotazy a práci

- Boris Lehečka
  - boris@daliboris.cz
  - konzultace a programování
  - volná noha

</body></TEI>