

Converting data from Czech digital libraries to TEI

Boris Lehečka, lehecka@mzk.cz, boris@daliboris.cz

TEI Conference 2025

16–19 September 2025, Krakau

The paper was funded by the Institutional support for long term conceptual development of a research organization (The Moravian Library) by the Czech Ministry of Culture.



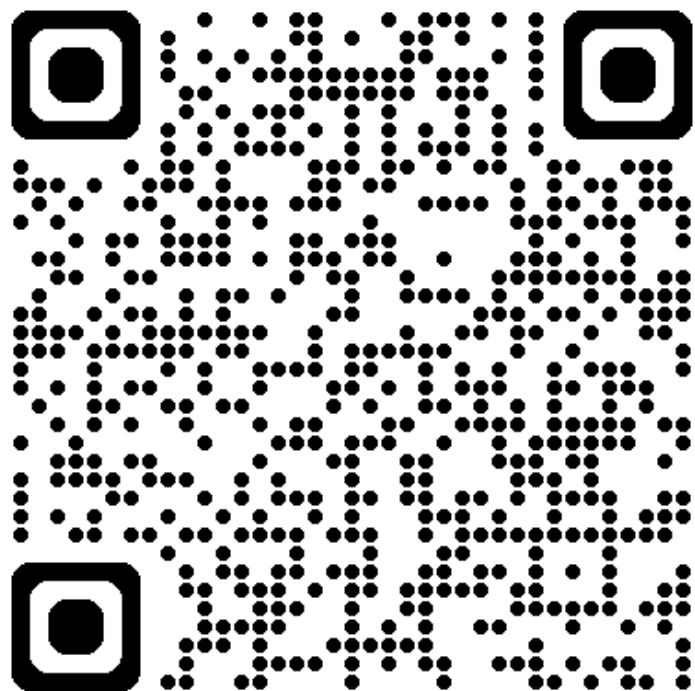
MINISTRY OF CULTURE
CZECH REPUBLIC



Presentation

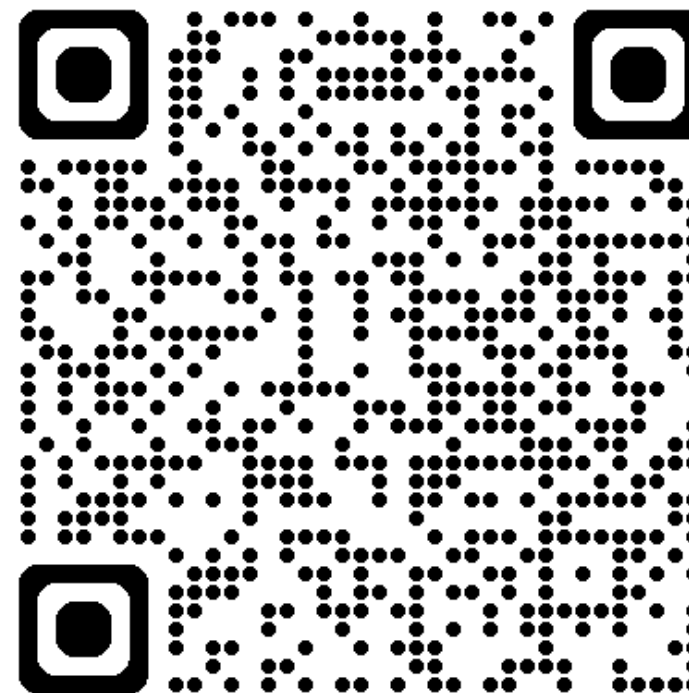
PowerPoint

<https://bit.ly/tei-2025-dl>



PDF

<https://bit.ly/tei-2025-dl-PDF>



Digital libraries

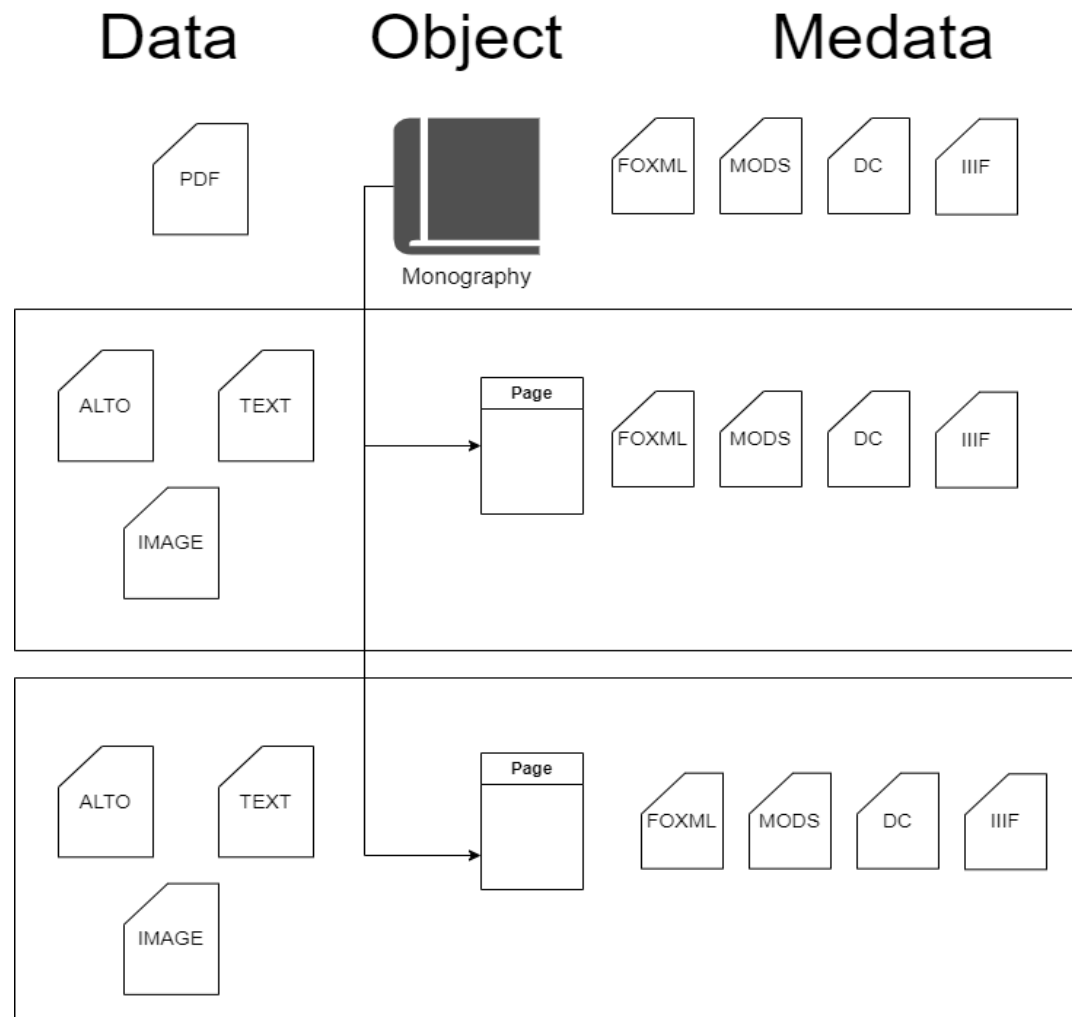
> main goals

- > preserving cultural heritage
- > wide and distant access
- > machine learning (LLM)

> standards

- > [FOXML](#)
 - > Fedora Object XML
- > [MODS](#)
 - > Metadata Object Description Schema
- > [DC](#)
 - > Dublin Core
- > [ALTO](#)
 - > Analyzed Layout and Text Object
- > [IIIF](#)
 - > International Image Interoperability Framework

Digital library objects and usual (meta)data



DL4DH

Digital Libraries for Digital Humanities

- > Development of tools for more effective use and mining of digital library data to strengthen digital humanities research
 - > 2020–2022
 - > supported by the Ministry of Culture of the Czech Republic ([DG20P020VV002](#))
- > making data available for data mining
- > Kramerius
 - > system used by Czech digital libraries
- > set of applications
 - > to enrich textual data (NLP) and convert them into CSV, JSON and TEI format

Pracovní prostory

Nesprávný postoj u dětí a je

https://www.digitalniknihovna.cz/mzk/view/uuid:de87a0e0-643b-11ea-a744-005056827e51?page=uuid:5f552dd5-9d36-4b15-8fc9-056ed06678f4

Moravská zemská knihovna

Hledat v celé digitální knihovně

Sbírký Procházet Informace Přihlásit

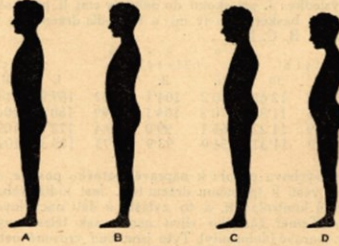
Hledat v dokumentu

Strana 3 z 4



3

zřizování a vedení klinik či ústavů určených jednak k preventivním zákrokům proti nesprávnému držení těla, jednak k nápravě hotového zla. Taková oddělení či kliniky doporučuje se zřídit při veřejných nemocnicích jako ambulatoria, ale hodi se též při jiných ústavech věnovaných péči o dítě. — Konečně dal též úřad pod dozorem dra KLEINA zhotoviti film, který předvádí v první části vliv nesprávného držení na celé tělo, v druhé pak nejdůležitější opatření



Vzory postoje u středostárných hochů.
(Children's Bureau, U. S. Dep. of Labor.)

Výborný postoj: Hlava vzhůru, brada zpět. Hrudník vyklenutý (hrudní kost přechází nejvíce do předu), dolní část břicha ustupuje zpět a jest oploštěná, zakřivení obrysu zad probíhá v normálních hranicích.
B. Dobrý postoj: Hlava málo do předu nakloněná, hrudník málo snížený, dolní část břicha ustupuje, ale není plochá; zakřivení obrysu zad málo zesíleno.
C. Postoj nedostatečný: Hlava nakloněná, hrudník plochý, břicho uvolněné (část trupu přechází nejvíce do předu), zakřivení zad zesíleno.
D. Špatný postoj: Hlava zřejmě nakloněná, hrudník stlačen (zapadlý), břicho úplně uvolněné a předstávající, zakřivení zad nadměrně zesíleno.

k jeho nápravě. Film jest především určen pro lékaře a učitele, ale může býti s užitek předváděn i rodičům a dospívající mládeži. — Při ústavech sloužících naznačenému účelu nutno počítati s dětmi v stáří 5—15 let; nejvíce přijdou k dobru dívkám 7—10 letým. Ústav musí říditi lékař, jemuž k ruce jsou pomocné síly (vyskolené opatrovnice, alespoň jedna na 10 pacientů); dítě se musí dostavit alespoň dvakrát, nejlépe třikrát týdně na jednu až půl druhé hodiny, které se vyplní vhodnými tělesnými cviky, zachováním určitých poloh atd. Výše uvedená publikace podává obrazem i slovně návody k řadě takových pomůcek. Dítě musí pak naučené alespoň třikrát denně doma opakovati.

Nesprávný postoj u dětí a jeho náprava =

Postural defects in children and their correction

...

Zpřístupněno pod licencí

Volná díla

Autor

Matiegka, Jindřich

...

S podporou



Nakladatelské údaje

V Praze: J. Matiegka, 1928

Typ dokumentu

Knihy

Ve sbírce

EOD OPEN

Jazyk

Čeština

Místo uložení

Moravská zemská knihovna v Brně

Signatura: PK-0016.603

100 % 10:08

<https://www.digitalniknihovna.cz/mzk/view/uuid:de87a0e0-643b-11ea-a744-005056827e51?page=uuid:5f552dd5-9d36-4b15-8fc9-056ed06678f4>

Conversion process (pseudocode)

- > for each page
 - > get ALTO
 - > identify paragraphs in ALTO blocks
 - > for each paragraph
 - > extract plain text
 - > analyze plain text with UDPipe
 - > get sentences + part-of-speech, lemma, and morphological categories for each token
 - > returns results in [CoNLL-U](#) format (tab delimited text with structure)
 - > analyze output from UDPipe with NameTag
 - > get named entities (return results in XML format)
 - > convert results from UDPipe and NameTag to TEI
 - > keep metadata about used service (name, link, license, used data model)
 - > merge TEI into one document
 - > convert ALTO blocks to TEI
 - > use TEI paragraphs from previous steps

Conversion process (pseudocode) – cont.

- > use FOXML (Fedora Object XML)
 - > contains metadata for one publication; includes MODS (Metadata Object Description Schema)
 - > convert to <TEI>
 - > convert MODS to <teiHeader>
 - > use metadata about NLP processing from one TEI document from previous conversion
 - > for each page
 - > take TEI from previous conversion

ALTO blocks and paragraphs

> TextBlock in ALTO

- > should represent "A paragraph of text" ([ALTO <Styles> & <Layout> Usage](#))
- > usually represents block of the text with identical typography parameters
 - > alignment (left, right, block), line spaces, left and right indent
 - > may contain
 - > multiple paragraphs
 - > part of one paragraph

ALTO blocks and paragraphs – cont.

> Solution

- > for each line (TextLine) in the TextBlock
 - > take values of horizontal position and width
 - > compute additional values
 - > number of words, number of spaces, number of characters,
 - > whether line starts with uppercase, whether it ends with hyphen
 - > compare current values for the line with average values for whole block
 - > based on the results of comparison: the line represents beginning or end of a paragraph, or line between
- > merge lines between start and end into one paragraph

§ 4.

Hmotným prostředkem k dosažení spolkového účelu jest spolkové imění, iež se opatřuje členskými příspěvky a výtěžky spolkové činnosti, případně dary.

§ 5.

Členství ve spolku jest pouze činné. Členem se může státí každá svéprávní fysická osoba, která jest přijata vybořem spolku buď na vlastní přihlášku anebo z usnesení výboru.

```
<TextLine BASELINE="528" HEIGHT="38" WIDTH="251" VPOS="498" HPOS="180">
  <String CONTENT="případně" HEIGHT="37" WIDTH="153" VPOS="499" HPOS="180" />
  <SP WIDTH="13" VPOS="498" HPOS="334" />
  <String CONTENT="dary." HEIGHT="37" WIDTH="83" VPOS="498" HPOS="348" />
</TextLine>
<TextLine BASELINE="577" HEIGHT="38" WIDTH="57" VPOS="548" HPOS="773">
  <String CONTENT="§5." HEIGHT="38" WIDTH="57" VPOS="548" HPOS="773" />
</TextLine>
<TextLine BASELINE="647" HEIGHT="43" WIDTH="1171" VPOS="613" HPOS="245">
  <String CONTENT="Členství" HEIGHT="33" WIDTH="139" VPOS="616" HPOS="245" />
  <SP WIDTH="11" VPOS="620" HPOS="385" />
  <String CONTENT="ve" HEIGHT="20" WIDTH="37" VPOS="628" HPOS="397" />
  <SP WIDTH="11" VPOS="628" HPOS="435" />
  <String CONTENT="spolku" HEIGHT="37" WIDTH="111" VPOS="619" HPOS="447" />
  <SP WIDTH="11" VPOS="619" HPOS="559" />
  <String CONTENT="jest" HEIGHT="36" WIDTH="57" VPOS="619" HPOS="571" />
  <SP WIDTH="10" VPOS="625" HPOS="629" />
  <String CONTENT="pouze" HEIGHT="27" WIDTH="100" VPOS="628" HPOS="640" />
  <SP WIDTH="11" VPOS="620" HPOS="741" />
  <String CONTENT="činné." HEIGHT="30" WIDTH="99" VPOS="618" HPOS="753" />
  <SP WIDTH="12" VPOS="613" HPOS="853" />
  <String CONTENT="Členem" HEIGHT="34" WIDTH="129" VPOS="613" HPOS="866" />
  <SP WIDTH="11" VPOS="628" HPOS="996" />
  <String CONTENT="se" HEIGHT="19" WIDTH="32" VPOS="628" HPOS="1008" />
  <SP WIDTH="11" VPOS="628" HPOS="1041" />
  <String CONTENT="může" HEIGHT="28" WIDTH="91" VPOS="619" HPOS="1053" />
  <SP WIDTH="11" VPOS="627" HPOS="1145" />
  <String CONTENT="státí" HEIGHT="29" WIDTH="71" VPOS="618" HPOS="1157" />
  <SP WIDTH="11" VPOS="618" HPOS="1229" />
  <String CONTENT="každá" HEIGHT="28" WIDTH="95" VPOS="619" HPOS="1241" />
  <SP WIDTH="11" VPOS="619" HPOS="1337" />
  <String CONTENT="svě-" HEIGHT="28" WIDTH="67" VPOS="619" HPOS="1349" />
</TextLine>
```

§ 4.

Hmotným prostředkem k dosažení spolkového účelu jest spolkové jmění, jež se opatřuje členskými příspěvky a výtěžky spolkové činnosti, případně dary.

§ 5.

Členství ve spolku jest pouze činné. Členem se může stát každá svépřávná fyzická osoba, která jest přijata výborem spolku buď na vlastní přihlášku anebo z usnesení výboru.

```
<line hpos="241" width="1173" string="11" sp="10" length="56" hyphen="0" uppercase="0" id="" text="7. Udrzuje stálé styky se svým ústředím, jir
<line hpos="180" width="1233" string="12" sp="11" length="55" hyphen="0" uppercase="0" id="" text="k povznesení venkova v Praze, přijímá jeho i
<line hpos="179" width="930" string="9" sp="8" length="45" hyphen="0" uppercase="0" id="" text="úkoly, které mu v rámci těchto stanov ústředí i
<line hpos="768" width="64" string="2" sp="1" length="3" hyphen="0" uppercase="0" id="" text="§ 4."/>
<line hpos="241" width="1175" string="8" sp="7" length="54" hyphen="0" uppercase="1" id="" text="Hmotným prostředkem k dosažení spolkového účel
<line hpos="181" width="1234" string="10" sp="9" length="62" hyphen="0" uppercase="0" id="" text="jmění, jež se opatřuje členskými příspěvky a
<line hpos="180" width="251" string="2" sp="1" length="13" hyphen="0" uppercase="0" id="" text="případně dary." role="end"/>
<line hpos="773" width="57" string="1" sp="0" length="3" hyphen="0" uppercase="0" id="" text="§5."/>
<line hpos="245" width="1171" string="12" sp="11" length="57" hyphen="1" uppercase="1" id="" text="Členství ve spolku jest pouze činné. Členem
<line hpos="182" width="1236" string="11" sp="10" length="60" hyphen="0" uppercase="0" id="" text="právní fyzická osoba, která jest přijata výl
<line hpos="184" width="613" string="5" sp="4" length="30" hyphen="0" uppercase="0" id="" text="přihlášku anebo z usnesení výboru." role="end",
```

NLP enrichment of text

- > external services operated by LINDAT/CLARIAH-CZ
 - > research infrastructure, <https://lindat.cz>
- > [UDPipe](#)
 - > identifies sentences and tokens
 - > assigns a lemma and morphological annotation to the token
 - > for 80 languages
- > [NameTag](#)
 - > tool for named entity recognition (NER)
 - > persons, geographical places, institutions, time and numerical data, etc.
 - > for 20 languages

NLP enrichment of text – cont.

> Problem

- > NameTag and UDPipe input formats
 - > plain text
 - > vertical (1 token per 1 line, empty line denoting end of sentence)
 - > CoNLL-U (1 token per line, empty line denoting end of sentence, extra metadata)
 - > (no XML)
- > NameTag output formats
 - > XML
 - > CoNLL-U
- > UDPipe output formats
 - > plain text
 - > vertical
 - > CoNLL-U
 - > [Matxin](#)
- > tokenization may be different

NLP enrichment of text – solution

- > extract plain text from ALTO
- > use UDPipe for segmentation and morphosyntactic analysis
 - > use language from metadata for analysis
 - > output in CoNLL-U format
 - > convert CoNLL-U to TEI
- > use NameTag
 - > use CoNLL-U output from UDPipe as input + language from metadata for analysis
 - > output in XML format
 - > convert XML to TEI
- > use TEI from NameTag
 - > merge attributes from corresponding tokens from UDPipe

UDPipe CoNLL-U

```
# generator = UDPipe 2, https://lindat.mff.cuni.cz/services/udpipe
# udpipes_model = czech-pdt-ud-2.15-241121
# udpipes_model_licence = CC BY-NC-SA
# newdoc
# newpar
# sent_id = 1
# text = 7.
1 7 7 NUM C=----- NumForm=Digit|NumType=Card 0 root _ SpaceAfter=No
2 . . PUNCT Z:----- _ 1 punct _ _

# sent_id = 2
# text = Udrzuje stález styky se svým ústředím, jímž jest spolek Sdružení k povznesení venkova v Praze, přijímá jeho rady a pokyny a provádí úkoly, kte
1 Udrzuje udržovat VERB VB-S---3P-AAI-- Aspect=Imp|Mood=Ind|Number=Sing|Person=3|Polarity=Pos|Tense=Pres|VerbForm=Fin|Voice=Act 0 root _ _
2 stález stález ADJ AAIP4-----1A---- Animacy=Inan|Case=Acc|Degree=Pos|Gender=Masc|Number=Plur|Polarity=Pos 3 amod _ _
3 styky styk NOUN NNIP4-----A---- Animacy=Inan|Case=Acc|Gender=Masc|Number=Plur 1 obj _ _
4 se s ADP RV--7----- AdpType=Voc|Case=Ins 6 case _ _
5 svým svůj DET P8ZS7----- Case=Ins|Gender=Masc,Neut|Number=Sing|Poss=Yes|PronType=Prs|Reflex=Yes 6 det _ _
6 ústředím ústředí NOUN NNNS7-----A---- Case=Ins|Gender=Neut|Number=Sing 3 nmod _ SpaceAfter=No
7 , , PUNCT Z:----- _ 8 punct _ _
8 jímž jenž PRON P4ZS7----- Case=Ins|Gender=Masc,Neut|Number=Sing|PrepCase=Npr|PronType=Rel 6 acl:relcl _ _
9 jest být AUX VB-S---3P-AAI-2 Aspect=Imp|Mood=Ind|Number=Sing|Person=3|Polarity=Pos|Tense=Pres|VerbForm=Fin|Voice=Act 8 cop _ _
10 spolek spolek NOUN NNIS1-----A---- Animacy=Inan|Case=Nom|Gender=Masc|Number=Sing 8 nsubj _ _
11 Sdružení sdružení NOUN NNNS1-----A---- Case=Nom|Gender=Neut|Number=Sing|VerbForm=Vnoun 10 nmod _ SpacesAfter=\r
12 k k NOUN Q3----- Abbr=Yes 13 nmod _ _
13 povznesení povznesení NOUN NNNS2-----A---- Case=Gen|Gender=Neut|Number=Sing|VerbForm=Vnoun 11 nmod _ _
14 venkova venkov NOUN NNIS2-----A---- Animacy=Inan|Case=Gen|Gender=Masc|Number=Sing 13 nmod _ _
15 v v ADP RR--6----- AdpType=Prep|Case=Loc 16 case _ _
16 Praze Praha PROPN NNFS6-----A---- Case=Loc|Gender=Fem|NameType=Geo|Number=Sing 11 nmod _ SpaceAfter=No
```


UDPipe CoNLL-U to TEI

```
<s n="1">
  <w n="1" lemma="7" pos="NUM" msd="NumForm=Digit|NumType=Card">7</w>
  <pc n="2" lemma="." pos="PUNCT" msd="_">.</pc>
</s>
<s n="2">
  <w n="1" lemma="udržovat" pos="VERB" msd="Aspect=Imp|Mood=Ind|Number=Sing|Person=3|Polarity=Pos|Tense=Pres|VerbForm=Fin|Voice=Act">Udržuje</w>
  <w n="2" lemma="stálý" pos="ADJ" msd="Animacy=Inan|Case=Acc|Degree=Pos|Gender=Masc|Number=Plur|Polarity=Pos">stálé</w>
  <w n="3" lemma="styk" pos="NOUN" msd="Animacy=Inan|Case=Acc|Gender=Masc|Number=Plur">styky</w>
  <w n="4" lemma="s" pos="ADP" msd="AdpType=Voc|Case=Ins">se</w>
  <w n="5" lemma="svůj" pos="DET" msd="Case=Ins|Gender=Masc,Neut|Number=Sing|Poss=Yes|PronType=Prs|Reflex=Yes">svým</w>
  <w n="6" lemma="ústředí" pos="NOUN" msd="Case=Ins|Gender=Neut|Number=Sing">ústředím</w>
  <pc n="7" lemma="," pos="PUNCT" msd="_">,</pc>
  <w n="8" lemma="jenž" pos="PRON" msd="Case=Ins|Gender=Masc,Neut|Number=Sing|PrepCase=Npr|PronType=Rel">jímž</w>
  <w n="9" lemma="být" pos="AUX" msd="Aspect=Imp|Mood=Ind|Number=Sing|Person=3|Polarity=Pos|Tense=Pres|VerbForm=Fin|Voice=Act">jest</w>
  <w n="10" lemma="spolek" pos="NOUN" msd="Animacy=Inan|Case=Nom|Gender=Masc|Number=Sing">spolek</w>
  <w n="11" lemma="sdružení" pos="NOUN" msd="Case=Nom|Gender=Neut|Number=Sing|VerbForm=Vnoun">Sdružení</w>
  <w n="12" lemma="κ" pos="NOUN" msd="Abbr=Yes">κ</w>
  <w n="13" lemma="povznesení" pos="NOUN" msd="Case=Gen|Gender=Neut|Number=Sing|VerbForm=Vnoun">povznesení</w>
  <w n="14" lemma="venkov" pos="NOUN" msd="Animacy=Inan|Case=Gen|Gender=Masc|Number=Sing">venkova</w>
  <w n="15" lemma="v" pos="ADP" msd="AdpType=Prep|Case=Loc">v</w>
  <w n="16" lemma="Praha" pos="PROPN" msd="Case=Loc|Gender=Fem|NameType=Geo|Number=Sing">Praze</w>
```

NameTag XML and TEI

```
<?xml version="1.0" encoding="utf-8"?><TEI xmlns="http://www.tei-c.org/ns/1.0"><teiHeader>...<body>
  <s>
    <num ana="#nametag-ni">
      <w>7</w>
    </num>
    <pc>.</pc>
  </s>
  <s>
    <w>Udrzuje</w>
    <w>stále</w>
    <w>styky</w>
    <w>se</w>
    <w>svým</w>
    <w>ústředím</w>
    <pc>,</pc>
    <w>jímž</w>
    <w>jest</w>
    <w>spolek</w>
    <orgName ana="#nametag-ic">
      <w>Sdružení</w>
      <w>k</w>
      <w>povznesení</w>
      <w>venkova</w>
      <w>v</w>
      <placeName ana="#nametag-gu">
        <settlement>
          <w>Praze</w>
        </settlement>
      </placeName>
    </orgName>
  </s>
</body></TEI>
```

```
<sentence>
  <ne type="ni">
    <token>7</token>
  </ne>
  <token>.</token>
</sentence>
<sentence>
  <token>Udrzuje</token>
  <token>stále</token>
  <token>styky</token>
  <token>se</token>
  <token>svým</token>
  <token>ústředím</token>
  <token>,</token>
  <token>jímž</token>
  <token>jest</token>
  <token>spolek</token>
  <ne type="ic">
    <token>Sdružení</token>
    <token>k</token>
    <token>povznesení</token>
    <token>venkova</token>
    <token>v</token>
    <ne type="gu">
      <token>Praze</token>
    </ne>
  </ne>
</sentence>
```

Merge of TEI from NameTag and UDPipe

```
<s>
  <num ana="#nametag-ni">
    <w n="1" lemma="7" pos="NUM" msd="NumForm=Digit|NumType=Card">7</w>
  </num>
  <pc n="2" lemma="." pos="PUNCT" msd="_">.</pc>
</s>
<s>
  <w n="1" lemma="udržovat" pos="VERB" msd="Aspect=Imp|Mood=Ind|Number=Sing|Person=3|Polarity=Pos|Tense=Pres|VerbForm=Fin|Voice=Act">Udržuje</w>
  <w n="2" lemma="stálý" pos="ADJ" msd="Animacy=Inan|Case=Acc|Degree=Pos|Gender=Masc|Number=Plur|Polarity=Pos">stálé</w>
  <w n="3" lemma="styk" pos="NOUN" msd="Animacy=Inan|Case=Acc|Gender=Masc|Number=Plur">styk</w>
  <w n="4" lemma="s" pos="ADP" msd="AdpType=Voc|Case=Ins">se</w>
  <w n="5" lemma="svůj" pos="DET" msd="Case=Ins|Gender=Masc, Neut|Number=Sing|Poss=Yes|PronType=Prs|Reflex=Yes">svým</w>
  <w n="6" lemma="ústředí" pos="NOUN" msd="Case=Ins|Gender=Neut|Number=Sing">ústředím</w>
  <pc n="7" lemma="," pos="PUNCT" msd="_">,</pc>
  <w n="8" lemma="jenž" pos="PRON" msd="Case=Ins|Gender=Masc, Neut|Number=Sing|PrepCase=Npr|PronType=Rel">jímž</w>
  <w n="9" lemma="být" pos="AUX" msd="Aspect=Imp|Mood=Ind|Number=Sing|Person=3|Polarity=Pos|Tense=Pres|VerbForm=Fin|Voice=Act">jest</w>
  <w n="10" lemma="spolek" pos="NOUN" msd="Animacy=Inan|Case=Nom|Gender=Masc|Number=Sing">spolek</w>
  <orgName ana="#nametag-ic">
    <w n="11" lemma="sdružení" pos="NOUN" msd="Case=Nom|Gender=Neut|Number=Sing|VerbForm=Vnoun">Sdružení</w>
    <w n="12" lemma="k" pos="NOUN" msd="Abbr=Yes">k</w>
    <w n="13" lemma="povznesení" pos="NOUN" msd="Case=Gen|Gender=Neut|Number=Sing|VerbForm=Vnoun">povznesení</w>
    <w n="14" lemma="venkov" pos="NOUN" msd="Animacy=Inan|Case=Gen|Gender=Masc|Number=Sing">venkova</w>
    <w n="15" lemma="v" pos="ADP" msd="AdpType=Prep|Case=Loc">v</w>
    <placeName ana="#nametag-gu">
      <settlement>
        <w n="16" lemma="Praha" pos="PROPN" msd="Case=Loc|Gender=Fem|NameType=Geo|Number=Sing">Praze</w>
      </settlement>
    </placeName>
  </orgName>
</s>
```

NLP enrichment metadata

```
<?xml version="1.0" encoding="utf-8"?><TEI xmlns="http://www.tei-c.org/ns/1.0"><teiHeader>...<body><encodingDesc><appInfo><application xml:id="NameTag" ident="NameTag" version="2"><label>NameTag, model: nametag3-czech-cnec2.0-240830</label></application><application xml:id="UDPipe" ident="UDPipe" version="2"><label>UDPipe, model: czech-pdt-ud-2.15-241121</label></application></appInfo></encodingDesc><profileDesc><textClass><classCode scheme="https://ufal.mff.cuni.cz/nametag/2/models"><interpGrp><interp xml:id="nametag-P">Komplexní jméno osoby</interp><interp xml:id="nametag-T">Komplexní časový údaj</interp><interp xml:id="nametag-A">Komplexní adresa</interp><interp xml:id="nametag-C">Komplexní bibliografická položka</interp><interp xml:id="nametag-a">ČÍSLA JAKO SOUČÁSTI ADRES</interp><interp xml:id="nametag-ah">Číslo popisné</interp><interp xml:id="nametag-at">telefon, fax</interp></interpGrp></classCode></textClass></profileDesc></body></TEI>
```

ALTO, TEI and images

> ALTO

- > describes areas with text or non-text content
- > with position on the image

> TEI

- > has <facsimile> + <surface> + <zone> elements
- > [alto2tei.xsl](#) can be used

> Problem

- > ALTO treats punctuation as a part of the token (<String>)
- > UDPipe and NameTag treats punctuation as separate token (<pc> in TEI)
- > merging surfaces/zones and NLP analysis needs more research

Resources

- > DL4DH TEI Converter

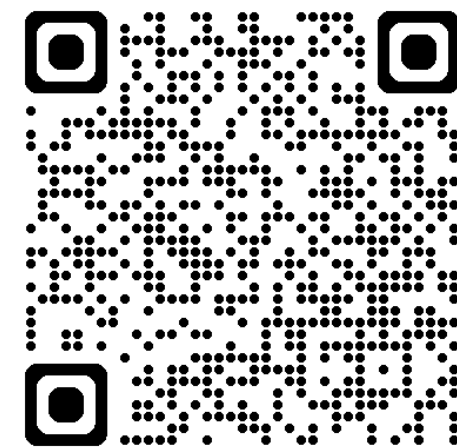
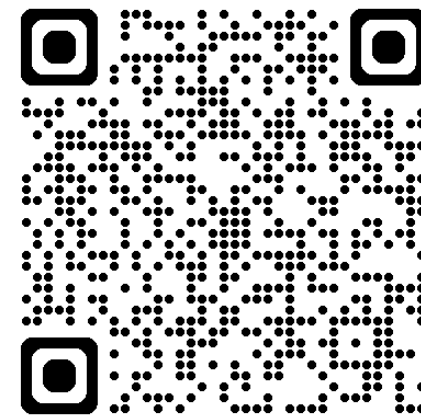
- > <https://github.com/LIBCAS/DL4DH-TEI-Converter>
- > converts enriched data (server version connected with Kramerius instance)

- > Libri augmentati

- > <https://github.com/moravianlibrary/libri-augmentati>
- > grabs (meta)data from digital library, enriches them and converts to TEI

- > Enriching textual data from digital libraries

- > demonstration of both of the above tools
- > 2025-09-19, 10:15—10:30 AM, room 402



Thank you!

> Boris Lehečka

- > lehecka@mzk.cz, boris@daliboris.cz
- > Moravian Library
- > Freelancer