Document-centred corpora

Why we need TEITOK at LINDAT

Pavel Straňák, “virtual Hejnice” seminar, 24 September 2020

Authors' corpora / "scholarly digital editions"

- **J. A. Komenský** – FLU AV, currently being prepared in XML editor
- **T. Aquinas** – UFAL, parallel corpus cs-en-lat, plaintext, alligned
- **Jan Patočka** – FLU AV, digitised most text (plaintext, no notes?) and audio. HTML version online
- **T. G. Masaryk** – MUA AV, many books in electronic formats for press (InDesign); grant application for corpus + dictionary

Other corpora with documents

- **EHRI Editions** – MUA AV, documents from holocaust era, facsimiles + transcriptions; complete editions online, but limited functionality (we could do more)
- **Early English Books Online (EEBO)** – Oxford, transcriptions with metadata in TEI, no facsimiles (copyright)
- **Corpus of Middle English** – dep. of English, faculty of Arts; creation in progress (O. Fusik)
- **Historical corpus at Institute for Czech Language** – transcriptions into TEI in progress
- **ParCzech** – Czech Parliamentary Corpus @UFAL. Matyáš Kopp downloading, converting and annotating transcripts + audio
  - http://quest.ms.mff.cuni.cz/parczech/teitok/parczech/index.php?action=browser (beta)
- **PDT** (discourse visualisation?) and other corpora?

Need for the right tools

- **Visualise** documents
  - with facsimiles, if they exist
  - combine with advanced corpus search
  - ability for advanced views like linked maps or dictionaries, text re-use, etc.
- **Annotate** documents and correct errors
  - Annotation tools and “live corpus” approach
- **TEITOK as a solution.** Handing over to Maarten Janssen.