Modeling letters using TEI

Corpus Linguistics 2013 – pre conference workshop

Peter Stadler
@ps_tadler

22nd June 2013

Material online

- http://goo.gl/8ZogvF
- https://github.com/peterstadler/Emigrant-Letters

What TEI is

- international standard for representation of texts in digital form
- XML based
- scholarly initiative
- institution: TEI consortium
- coding schema: available as RelaxNG, W3C Schema, DTD
- guidelines

What TEI has to offer

- more than 20 years of experience with text encoding
- no major gaps in the conceptual model
- active community
- active development
- several modules for distinct tasks: dictionaries, drama, feature structures, critical apparatus, . . .
- several SIGs: linguists, correspondence, . . .
- variety of tools for processing TEI

TEI showcases

- William Godwin’s Diary
- Carl-Maria-von-Weber-Gesamtausgabe (WeGA)

What TEI does not *out of the box*

TEI is a framework. Every edition is different.

- interoperability/interconnectivity
- linked (open) data
- taxonomies

But it *facilitates* all of the above!

How the TEI works

- one file consisting of meta description (catalogue data) *and* text
- `teiHeader`
- `text`
- `facsimile`

TEI for emigrant letters

Aim: Identify and model the idiosyncrasies of *emigrant letters*
Presupposition: Start with the meta data

- sender
- addressee
- date (of shipment/receipt)
- place (of sender/addressee)
- incipit(?)
- relation(s) of sender and addressee
- repository

Linguistic annotation with TEI

- tag every word as `<w>`
- tag phrases as `<phr>`
- tag sentences as `<s>`
- encode lemmata with `@lemma`
- encode parts of speech with `@type` (or `@function`?)