From paper slips via TUSTEP to TEI - creating a lexicographic information system from the WBÖ database

Jack Bowers
Philipp Stöckle
Omar Siam

ÖAW ÖSTERREICHISCHE AKADEMIE DER WISSENSCHAFTEN

DHA 2017 - Innsbruck - 6/12/2017

acdh austrian centre for digital humanities

ÖAW ÖSTERREICHISCHE AKADEMIE DER WISSENSCHAFTEN

austrian centre for digital humanities

Outline

I. Overview of history of project to the present
II. Discussion of data: conversion; complications; enhancements
III. Future WBÖ work: article writing; online content

ÖAW ÖSTERREICHISCHE AKADEMIE DER WISSENSCHAFTEN

austrian centre for digital humanities

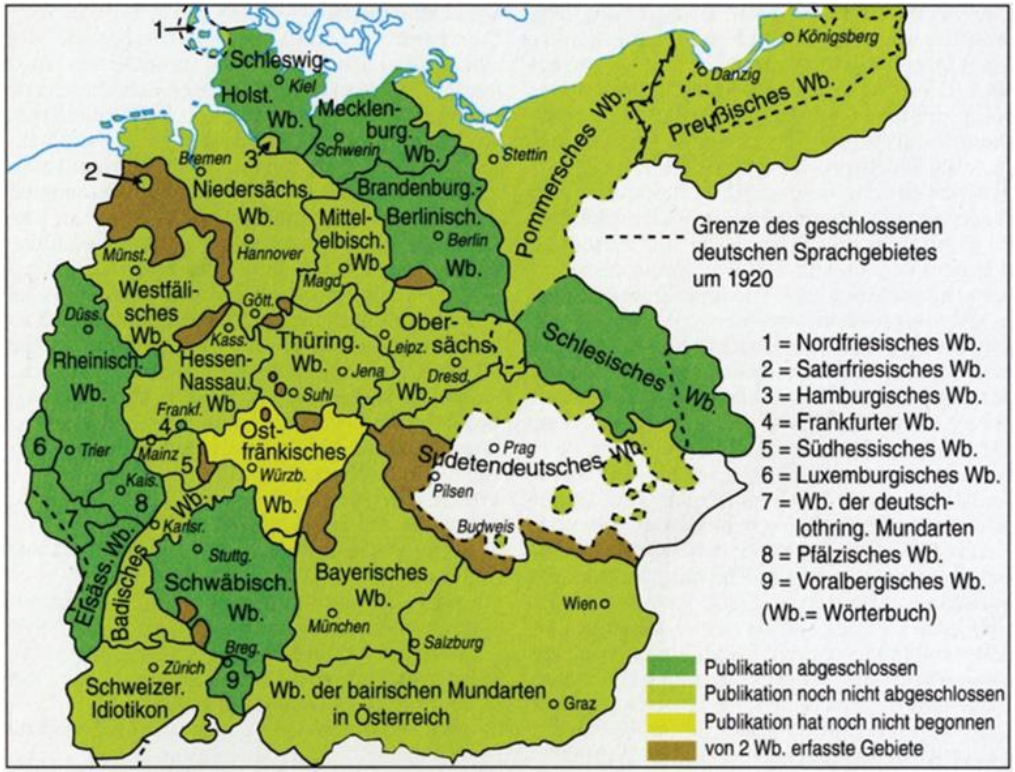
(I) Project Overview: Origins & Timeline

1911 Foundation of research committees in Vienna and Munich with the objective of creating dictionaries of Bavarian dialects
1913-1937 Questionnaire-based surveys, including 109 main questionnaires (1913-1933) and 9 supplementary questionnaires (1927-1937)
1927-1965 Direct data acquisition (so-called "Kundfahrten")

→ Paper slips were collected in the so-called "Hauptkatalog" (main catalogue), which contains approximately 3,6 mio paper slips

1963-2015 Publication of first five volumes of the "Wörterbuch der bairischen Mundarten in Österreich (WBÖ)", including the entries *A–Ezzes*
1993-2011 Creation of a digital version of the database by manually entering the original hand-written paper slip entries into TUSTEP
Since 2015 subsequent conversion into XML/TEI format
Since 2016 Relocation of WBÖ at the department "Variation and Change of German in Austria" at the ACDH

The WBÖ in the context of German-speaking dialect lexicography



König, Werner et al. ([18]2015): *dtv-Atlas deutsche Sprache. Mit 155 Abbildungen in Farbe*. München: Deutscher Taschenbuchverlag, p. 138.

ÖAW ÖSTERREICHISCHE AKADEMIE DER WISSENSCHAFTEN

austrian centre for digital humanities

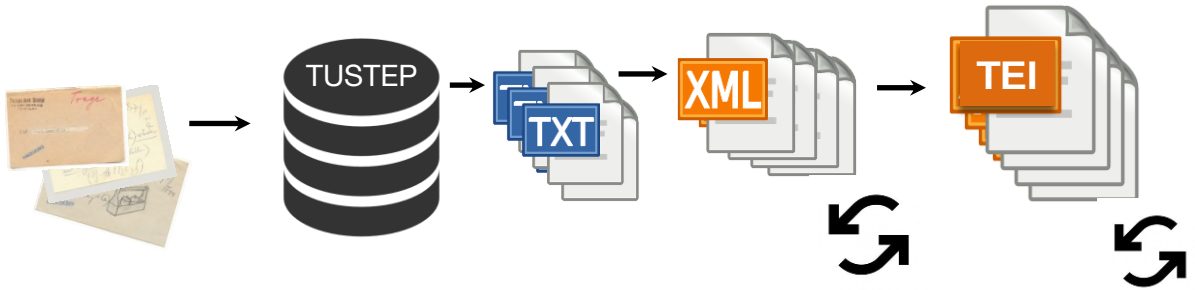
Overview of database contents

- Headwords
- Phonetic representation(s) of dialectal forms
- Grammatical info: pos, number, case, etc.
- Details of word formation
- Etymological information
- Translation (&/or) definition of meaning
- Location data for each record
- Questionnaires used in elicitation of most of dialectal data
- Bibliographic references
- ...

ÖAW ÖSTERREICHISCHE AKADEMIE DER WISSENSCHAFTEN

austrian centre for digital humanities

(II) Stages of Conversion



Transformation of structure from TUSTEP > XML > TEI

ÖAW ÖSTERREICHISCHE AKADEMIE DER WISSENSCHAFTEN

austrian centre for digital humanities

(II) Details & Challenges of Conversion

- Heterogeneous Datasets:
  - no relevant data field was present or accurate in 100% of the entries
  - entry fields had a large degree of irregularity
  - 510 categories in original data! (269 incorrectly labelled!)
- Dialect transcriptions often unreadable to humans
- Invisible characters and characters invalid in Unicode crashed XSL scripts
- Deviations from guidelines in data entries:
  - in dialectal transcriptions
  - in field labeling/contents
- Multiple versions of source TUSTEP database

ÖAW ÖSTERREICHISCHE AKADEMIE DER WISSENSCHAFTEN

austrian centre for digital humanities

(II) TUSTEP > TEI Content

TUSTEP

```
*A* HK 450, k4500502.pas^#95
*HL* k.af;éeln:6
*QU* Obertraun Hallstatt, Hango
*QDB* {5.3a04} söSkgt.:swTraunv.:OÖ *^@ Mtlg.HANGO·
(1927) *O* Hallst. OÖ [2o/2.]
===
*LT1* kafféeln *ANMB* ^#ée^# für: Akut über ^#ee^#
*BD/LT1* Kaffee trinken
**********
```

TEI

```
<entry xmlns="http://www.tei-c.org/ns/1.0"
                    xml:id="k450_qdb-d1e37367"
                    xml:lang="bar">
                    <form type="hauptlemma">
                       <orth type="orig">k.af;éeln</orth>
                       <orth type="normalized">kafféeln</orth>
                    </form>
                    <gramGrp>
                       <pos>Verb</pos>
                    </gramGrp>
                    <form type="lautung" n="1">
                       <pron notation="tustep">kaffe'eln</pron>
                    </form>
                    <sense corresp="this:LT1">
                       <def xml:lang="de">Kaffee trinken</def>
                    </sense>
                    <ref type="archiv">HK 450, k4500502.pas^#91</ref>
                    <ref type="quelle">Obertraun OÖ, Hango</ref>
                    <ref type="quelleBearbeitet">{5.3a06} söSkgt.:swTraunv.:OÖ</ref>
                    <usg type="geo">
                       <placeName type="orig">Obertraun OÖ</placeName>
                       <listPlace ref="sigle:5.3a06">[...]</listPlace>
                    </usg>
</entry>
```

ÖAW ÖSTERREICHISCHE AKADEMIE DER WISSENSCHAFTEN

austrian centre for digital humanities

(II) TUSTEP, TEI and DBÖ Contents

- TUSTEP is a great word processing suite with macro capabilities and its own programming language and the data is encoded in it’s own way
- TEI is open source very well documented and the programs you need are all used for any generic XML therefore well known
- TEI vocabulary is able to accommodate all content and structures inherited from TUSTEP and improves structural efficiency and clarity
- TEI is used in the community, data more compatible with partner projects

ÖAW ÖSTERREICHISCHE AKADEMIE DER WISSENSCHAFTEN

austrian centre for digital humanities

(II) Further Enhancements

- Re-export and convert TUSTEP Teuthonista encoding to actual Unicode characters (and IPA conversion in conjunction w/ Verba Alpina project)
- *Siglen* -> Add multiple <listPlace> structures for detailed geographic information -> can be used with geo information systems and displayed
- Normalization of Hauptlemma and store their decoded segmentation as XML
- Enhance and normalize also the sense content
- Scan the notecards and link these pictures to entries
- Continue to fix erroneous miscellaneous contents
- Add missing content

ÖAW ÖSTERREICHISCHE AKADEMIE DER WISSENSCHAFTEN

austrian centre for digital humanities

Re-export & Teuthonista Conversion

Currently the majority of the transcriptions are stored in a code invented because TUSTEP system still could not represent enough characters. We will re-export the contents with the Teuthonista transcriptions converted to Unicode characters - (Derkits)

d-.es -is |A diN dA)u¡nm-.e%))glixkaid

↓

dēs īs ɐ diŋ dɐ ŭⁿmēgliχkaid

Kontext aus HK 157, d157^#142.1 Ding

i h;a¡n m.e¡c¡h >s;a¡n fodíNd

↓

i hąⁿ mėᶜʰ šąⁿ fodíŋd

Kontext aus HK 157, d157^#910.1 ferdingen

ÖAW ÖSTERREICHISCHE AKADEMIE DER WISSENSCHAFTEN

austrian centre for digital humanities

Refinement Example - Decoding Sigles

- Sigles contain a hierarchical concept of regions in Austria and STir
- Can be expressed in TEI with a <listPlace> structure
- Could be referenced with <region ref=”sigle:3.1k”>
    - May not be searchable fast enough due to lookup

ÖAW ÖSTERREICHISCHE AKADEMIE DER WISSENSCHAFTEN

austrian centre for digital humanities

Refinement Example - Decoding Sigles II

```
<listPlace ref="sigle:5.3a06">
                    <place type="Bundesland">
                    <placeName>OÖ</placeName>
                    <idno>5</idno>
                    <listPlace>
                    <place type="Großregion">
          <placeName>Traunv.</placeName>
          <idno>5.3</idno>
          <listPlace>
            <place type="Kleinregion">
              <placeName>söSkgt.</placeName>
              <idno>5.3a</idno>
              <listPlace>
                <place type="Gemeinde">
                  <placeName>Obertraun</placeName>
                  <idno/>
                    <listPlace>
                     <place type="Ort">
                        <placeName>Obertraun</placeName>
                        <idno>5.3a06</idno>
                     </place>
```

ÖAW ÖSTERREICHISCHE AKADEMIE DER WISSENSCHAFTEN

austrian centre for digital humanities

Refinement Example - Normalizing

- Word segments are encoded using []{}()-
- Encoding is bad for full text search
- Added a normalized form by removing them
- Check and correct the normalized form (e. g. F -> V)
- Use TEI vocabulary to express segment

ÖAW ÖSTERREICHISCHE AKADEMIE DER WISSENSCHAFTEN

austrian centre for digital humanities

Refinement Example - Normalizing II

```
<form type="hauptlemma">
        <orth type="parsed">
            <seg>
            <seg>Amts</seg>
            <seg>pflicht</seg>
            </seg>
            <seg>for</seg>
            <seg>halt</seg>
        </orth>
        <orth type="orig">(Amts–pflicht)for–halt</orth>
        <orth type="normalized">Amtspflichtforhalt</orth>
</form>
```

ÖAW ÖSTERREICHISCHE AKADEMIE DER WISSENSCHAFTEN

austrian centre for digital humanities

Challenges in Implementing Enhancements

Editing and making these enhancements to the DBÖ data is not a straightforward process when working with BaseX (and other XML) databases.

- 2.2 Mio <tei:entry> with various child nodes is challenging for XML database systems (BaseX, exist-db)
- Finding an entry in an XML database of above 3 GB size needs indexes
- Searching is quite fast in BaseX as there are various indexes to really speed up searches

ÖAW ÖSTERREICHISCHE AKADEMIE DER WISSENSCHAFTEN

austrian centre for digital humanities

Challenges in Implementing Enhancements - Indexes

- It is easy to confuse BaseX 8.6 so it does not use the indexes
- Changing a single character will either delete all the indexes or
- The indexes need to be rebuild after the change and that takes **very** long
- While writing XML data is in progress the data is inaccessible

ÖAW ÖSTERREICHISCHE AKADEMIE DER WISSENSCHAFTEN

austrian centre for digital humanities

Challenges in Implementing Enhancements - Data Splitting

- In BaseX the solution we pursued is splitting the data into 700 small databases
- For searching we hide this fragmentation behind an API that automatically queries each database and then presents the results

ÖAW ÖSTERREICHISCHE AKADEMIE DER WISSENSCHAFTEN

austrian centre for digital humanities

Challenges in Implementing Enhancements - Parallel Work

- We can query several DBs in parallel to speed up the search
- While there is an update only one out of those 700 DBs is inaccessible
- Index generation is per database so it is fast
    - This is in part possible due to the fact that the source itself is split into 700 drawers

ÖAW ÖSTERREICHISCHE AKADEMIE DER WISSENSCHAFTEN

austrian centre for digital humanities

(III) Next stages in WBÖ project

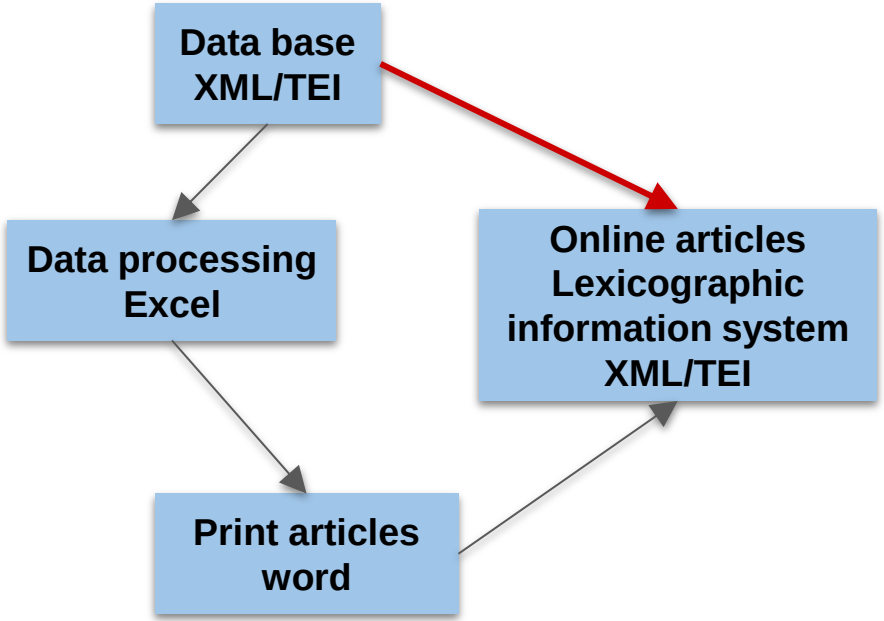
- Continue to make further enhancements to source (DBÖ) data
- Build lexicographic editor tool
- Begin writing articles for print and online versions
- Build online platform

ÖAW ÖSTERREICHISCHE AKADEMIE DER WISSENSCHAFTEN

austrian centre for digital humanities

Article Writing

- Formulation of lexicographic guidelines for article writing
- Articles will be written directly in TEI as a dually born digital and print resource output
- A separate XML based system will be used - still evaluating solutions



ÖAW ÖSTERREICHISCHE AKADEMIE DER WISSENSCHAFTEN

austrian centre for digital humanities

Online Lexicographic Information System

- Contains platform for visualizing articles as well as access to data base
- Geo information as maps <-> entries by selected region
- Source links into the lexicographic information system for reproducibility
- Link database entries with scans of paper slips
- Link database entries with other materials (such as scans of questionnaires)
- Link database entries and articles with other projects and dictionaries

ÖAW ÖSTERREICHISCHE AKADEMIE DER WISSENSCHAFTEN

austrian centre for digital humanities

Conclusion

- Self invented data structures and field names are not sustainable → TEI provides a solid framework
- Data is best kept and edited in database solutions not in spreadsheets or text files
- … collecting, storing, digitizing, converting data within a time span of more than 100 years can create various sources of errors → require careful revisions