«Tables are tricky»

Testing Text Encoding Initiative (TEI) Guidelines for FAIR upcycling of digitised historical statistics

Gabi Wüthrich, Universitätsbibliothek Zürich

1. Project Background

Digitising Demographic Statistics of Zurich

- Willi Bretscher Fellowship Project at Zentralbibliothek (Central Library, ZB) Zurich by Dr. Joël Floris (2022): «Digitising demographic data from Zurich, 1910-1925, to quantify and contextualize the Spanish flu»
- Facsimile incl. OCR published on Zurich Open Platform (ZOP) by ZB
- This project: Pilot paper «From book to machine. XML as a platform-independent, machine-readable table format» as part of CAS UZH in Data Management and Information Technologies

Inspiration: Annual Fiscal Accounts of Basle

- Digital edition of the Basel annual accounts 1535-1611
- HTML, facsimile, and table view parallelly available
- Based on XML processing in accordance with TEI and RDF (Resource Description Framework) standard
- Early example for FAIR data edition

2. Text Encoding Initiative (TEI)

- Consortium developing and maintaining standards for the digital representation of texts since 1987
- Current version 4.9.0. of the guidelines published Jan 25th
- 24 chapters that explain TEI and define text elements (XML-based), including tables in chapter 15
- Structural elements: TEI root, TEI header, text body
- Advantages
  - Focus on the meaning of words and texts, not on layout
  - Software-independent
  - Supported by an open scientific community
- Disadvantages of tables
  - Layout and arrangement are more important than in continuous text, and convey meaning
  - Table processing in XML notations other than TEI

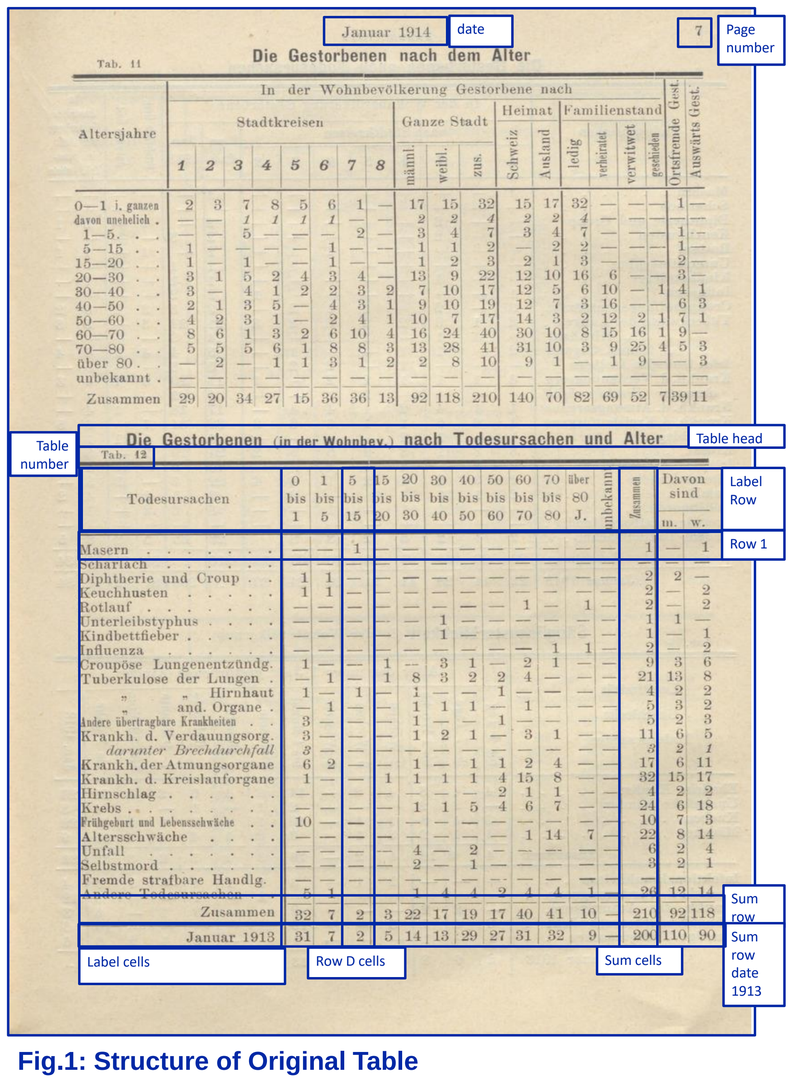


Fig.1: Structure of Original Table

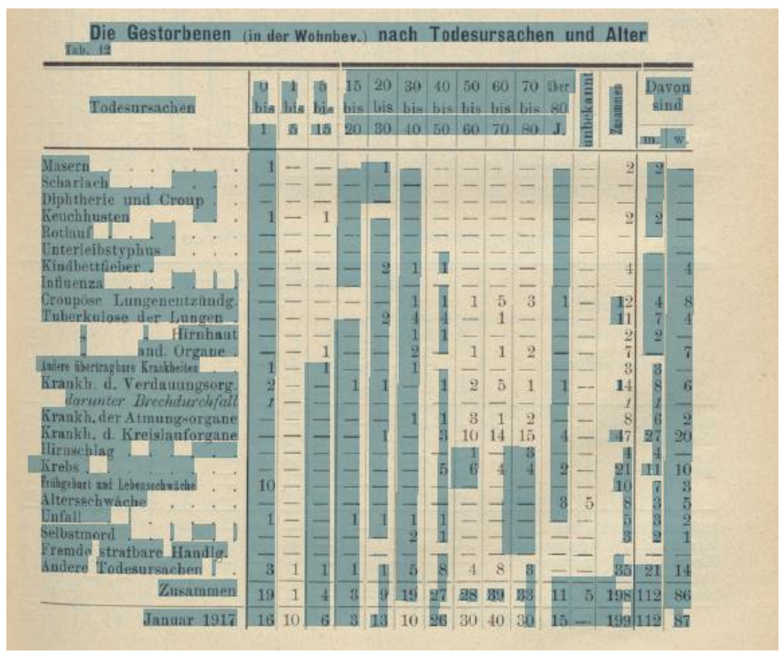
The respective structural elements were included as attributes in the XML code (see section 4), using similar attribute terms like the Basle Account edition wherever possible.

3. Challenges

- OCR recognition in digital copies must be adjusted for tables
- Image quality of digitisation needs to be sufficient
- OCR and TEI are flow text oriented
- Manual transfer of table to TEI structured XML content is prone to errors

→ Table recognition and XML coding are correspondingly deficient for structured text

→ Chicken-and-egg problem: More (XML) table templates for better text recognition supported by LLMs?



Tab.
Er u wu.
0,15 115|20)30|40|50 60 |70 jüber, 5| 5 Davon
Todesursachen bis bis| bis bis bis bis| bis nn bis bis 80'& = sind
1 |515/20|30|40|50 | 5S|ılE ee
Masern . See u 1 — 1
Scharlach ; _— | 1-1 1-2 -|1212)2| —1—-| —
Diphtherie und Croup . d: Hz _ -|-1-1-|-/-|-— .2] 2 —
Keuchhusten T - -|-1-1-1- -)- —— 23-—-| 2
Rotlauf $ — _—|-|1-— Ira 1— 2-2
Unterleibstyphus ; _ —|-111- eis wur. —|— 11 -—
Kindbettfieber . | - -112/=+/-=|-|-| 11 — 1
Influenza : Seesen 1 1—-| 21 -| 2
Croupöse Lungenentzündg. 1] 11773177 = 2 1| —'—- 9 3) 6
Tuberkulose der Lungen .|—| 1— 1| 8 3| 2| 2) 4/— | ut 21] 13) 8
5 wsBrnnaut Eee De 1 41.21 2
ee ieDel en sl 3 2
Andere übertragbare Krankheiten 3—-[—|—| 11-— -| 1 |—| -|=| 5 2 3
Krankh. d. Verdauungsorg. 3 — —|— 1 2/1— 3) 11 — —1165
darunter Brecehäurehfal 83 — —\—|— || — - — Ba een 8.267
Krankh.der Atmungsorgane 6| 2 — — 11-11 2|4 — — 1 611
Krankh. d. BIER 1=)—=| 17.177174 [1578 ef=ej7832[151717
‘Hirnschlag . ö | | — 1-11 1—12|.1.1) — — Al-- 2:52
Krebs .. . | —-1—| 12) 1} 54! 6) 7) —| =} 2724| °°6) 18
Frühgeburt ni Tebensichwäche 10 — || eek er 20l- 7,48
Altersschwäche 1-1 - -—|-|-—- 1 a 71—| 22]. 8) 14
Unfall --/-/-=f4-|2|-/-|-|—— 6| 2 4
Selbstmord . . ——|1— 21—|.11— — 11 —|— .81- 211
Fremde strafbare Handig. er el est | ee en _ —
Andere Todesursachen . DEab 1) 4! 4|.2| 4| 4 1)—| 26| 12) 14
Zusammen 32 7| 2| 322 17 19 17.40 ai 10 — 210] 921118
Januar 1913 31) 7| 2| 5|14|18 29 |27|31 32) 9— 200110 9

Fig. 2 and 3: Optical Character Recognition (OCR) output as shown in Acrobat Reader (above) and OCR copied from pdf in table format (below).

Tesseract was used for the OCR, prioritising speed over precision. However, the Central Library’s digitisation center has received feedback from researchers that OCR quality becomes more and more important due to rising demand in the context of Linked Open Data and Large Language Models (LLMs).

Okular, a free OCR tool in which table structuring can be predetermined, was used to capture the table content in the end, but the output was still full of errors and missing values due to the insufficient OCR quality.

Until recently, AI tools struggled to extract information from complex tables such as the one chosen for the pilot project, yielding incomplete and/or false results. OpenAI's new o4-mini reasoning model now seems to interpret such tables relatively accurately when guided by tailored prompts.

Check the QR-Code for the prompting and its output in ChatGPT.

4. Structure of Table in TEI-XML

TEI Header Elements

- «Metadata» of original table
- Some elements are strictly required by TEI guidelines:
  - Title
  - Edition Statement
  - Publication Statement
  - Bibliographic Information on original source
- This minimal TEI header is specifically adjusted to the pilot project

TEI Text Body

- XML version of table content (see Fig. 1):
  - Title
  - Table number as note
  - Date of table
- Table organised row by row, cell after cell
- Cell content can be text or numeric or empty
- Lines and long dashes ignored for swifter implementation, unlike in standard TEI. Easy to include and then remove for statistical analysis

→ Human and machine readable but error prone…

```
<?xml version="1.0" encoding="UTF-8"?>
<TEI xsi:schemaLocation="http://www.tei-c.org/ns/1.0 AdjSchema.xsd"
xmlns="http://www.tei-c.org/ns/1.0"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">
  <teiHeader>
    <fileDesc>
      <titleStmt> <!-- Titel dieser XML-Datei CHECK -->
        <title>Todesursachen nach Alter im Januar 1914, Tabelle 12</title>
      </titleStmt>
      <editionStmt> <!-- Die dieser XML-Datei zugrundeliegende Edition -->
        <edition>Digitalisierte pdf-edition der historischen
Statistiken</edition>
        <respStmt>
          <resp>Digitalisiert durch</resp>
          <orgName>Zurich Open Platform Zentralbibliothek
Zürich</orgName>
        </respStmt>
      </editionStmt>
      <publicationStmt> <!-- Art dieser digitalen "Edition" -->
        <p>Tabelle als XML in TEI für Projektarbeit</p>
      </publicationStmt>
      <sourceDesc>
        <biblFull> <!-- Bibliografische Erfassung -->
          <titleStmt>
            <title>Monats-Berichte des Statistischen Amtes der Stadt Zürich
1914</title>
            <author>Statistisches Amt der Stadt Zürich</author>
          </titleStmt>
          <extent>S.7-8</extent>
          <publicationStmt>
            <publisher>Buchdruckerei Berichtshaus</publisher>
            <date>1915</date>
          </publicationStmt>
          <notesStmt>
            <relatedItem>
              <ref>
                <idno type="DOI">https://doi.org/10.20384/zop-869</idno>
              </ref>
            </relatedItem>
          </notesStmt>
        </biblFull>
      </sourceDesc>
    </fileDesc>
    <!-- Hier eventuell xenodata ergänzen -->
  </teiHeader>
  <!-- Ende TEI Header und Start Textbody, d.h. Start der in der Tabellen
erhaltenen Informationen-->
```

```
<text>
    <body>
      <div>
        <head>
          <title>Die Gestorbenen (in der Wohnbevölkerung) nach
Todesursachen und Alter</title>
          <note n="Tabelle 12"></note>
          <date> Januar 1914 </date>
        </head>
        <table>
          <row role="label">
            <cell role="label">Todesursachen</cell> <!-- entspricht Zelle
in der Spalte A, markiert mit Attributrolle Label -->
            <cell>0 bis 1</cell> <!-- entspricht Zelle in der Spalte B -->
            <cell>1 bis 5</cell> <!-- entspricht Zelle in der Spalte C -->
            <cell>5 bis 15</cell> <!-- entspricht Zelle in der Spalte D -->
            <cell>15 bis 20</cell> <!-- entspricht Zelle in der Spalte E -->
            <cell>20 bis 30</cell> <!-- entspricht Zelle in der Spalte F -->
            <cell>30 bis 40</cell> <!-- entspricht Zelle in der Spalte G -->
            <cell>40 bis 50</cell> <!-- entspricht Zelle in der Spalte H -->
            <cell>50 bis 60</cell> <!-- entspricht Zelle in der Spalte I -->
            <cell>60 bis 70</cell> <!-- entspricht Zelle in der Spalte J -->
            <cell>70 bis 80</cell> <!-- entspricht Zelle in der Spalte K -->
            <cell>über 80 J.</cell> <!-- entspricht Zelle in der Spalte L -->
            <cell>unbekannt</cell> <!-- entspricht Zelle in der Spalte M --
>
            <cell ana="#sum">Zusammen</cell> <!-- entspricht Zelle in
der Spalte N, markiert mit Attributname #sum für Summenzelle der
entsprechenden Zeile -->
            <cell>Davon sind m</cell> <!-- entspricht Zelle in der Spalte
O -->
            <cell>Davon sind w</cell> <!-- entspricht Zelle in der Spalte
P -->
          </row>
          <row role="data">
            <cell role="label">Masern</cell> <!-- A -->
            <cell></cell> <!-- B -->
            <cell></cell> <!-- C -->
            <cell>1</cell> <!-- D -->
            <cell></cell> <!-- E -->
            <cell></cell> <!-- F -->
            <cell></cell> <!-- G -->
            <cell></cell> <!-- H -->
            <cell></cell> <!-- I -->
            <cell></cell> <!-- J -->
            <cell></cell> <!-- K -->
            <cell></cell> <!-- L -->
            <cell></cell> <!-- M -->
            <cell ana="#sum">1</cell> <!-- N -->
            <cell></cell> <!-- O -->
            <cell>1</cell> <!-- P -->
          </row>
```

5. Conclusion

- Upcycling tables according to TEI guidelines is generally possible
- Comprehensive TEI text elements require training time for practical implementation in XML
- TEI offers tools that can be used to capture complex structures of historical tables in XML format
- TEI can be used to track changes in table structures
- TEI enables platform- and software-independent processing of table data – so do .csv and .txt
- TEI generally suitable for structuring continuous text, less suitable for texts where structure determines meaning
- Workflow becomes easier with improving AI tools and LLMs

Contact

Gabi Wüthrich
gabi.wuethrich@ub.uzh.ch
www.ub.uzh.ch

Paper

10.5281/
zenodo.
13903990

Key References

- Zentralbibliothek Zürich, Zurich Open Platform, https://zop.zb.uzh.ch/home
- Floris, Joël (2023), Die Spanische Grippe in Zürich erzählen, https://www.zb.uzh.ch/de/zuerich/die-spanische-grippe-zuerich-erzaehlen
- TEI-L mailing list, TEI for tables, https://listserv.brown.edu/cgi-bin/wa?A1=ind2206&L=TEI-L#20 [accessed on 17th of June 2022]
- Vogeler, Georg (2015). Warum werden mittelalterliche und frühneuzeitliche Rechnungsbücher eigentlich nicht digital ediert?. In: Baum, Constanze und Stäcker, Thomas (Hg.). Grenzen und Möglichkeiten der Digital Humanities. (Sonderband der Zeitschrift für digitale Geisteswissenschaften, 1). <10.17175/sb001_007> [accessed on 14th of June 2022].