“Tables are tricky”

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

Digital History Conference 2024

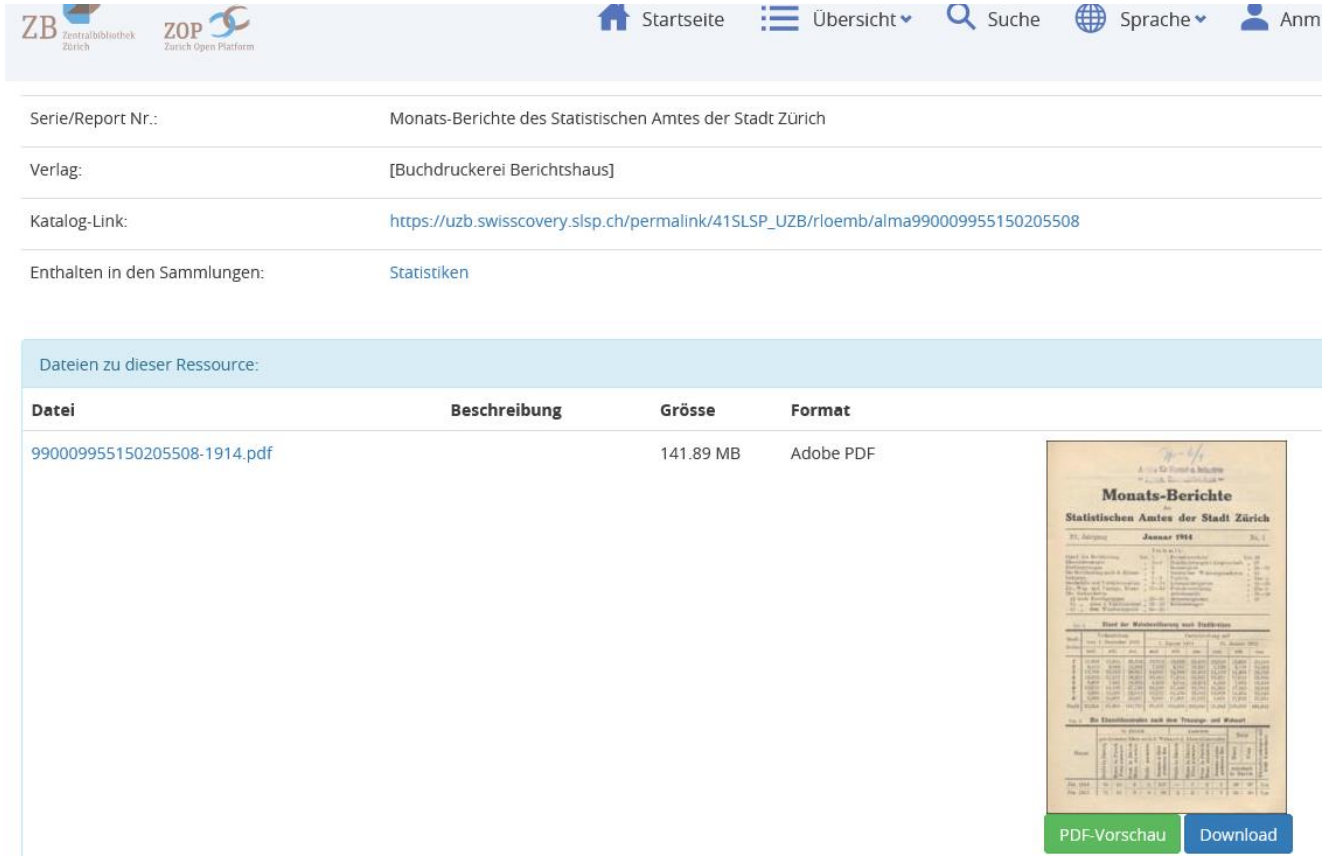
Gabi Wüthrich
Universitätsbibliothek Zürich

Overview

- Project(s) background and inspiration
- TEI in a nutshell
- Original Table and (code) structure of TEI table
- Challenges and conclusion

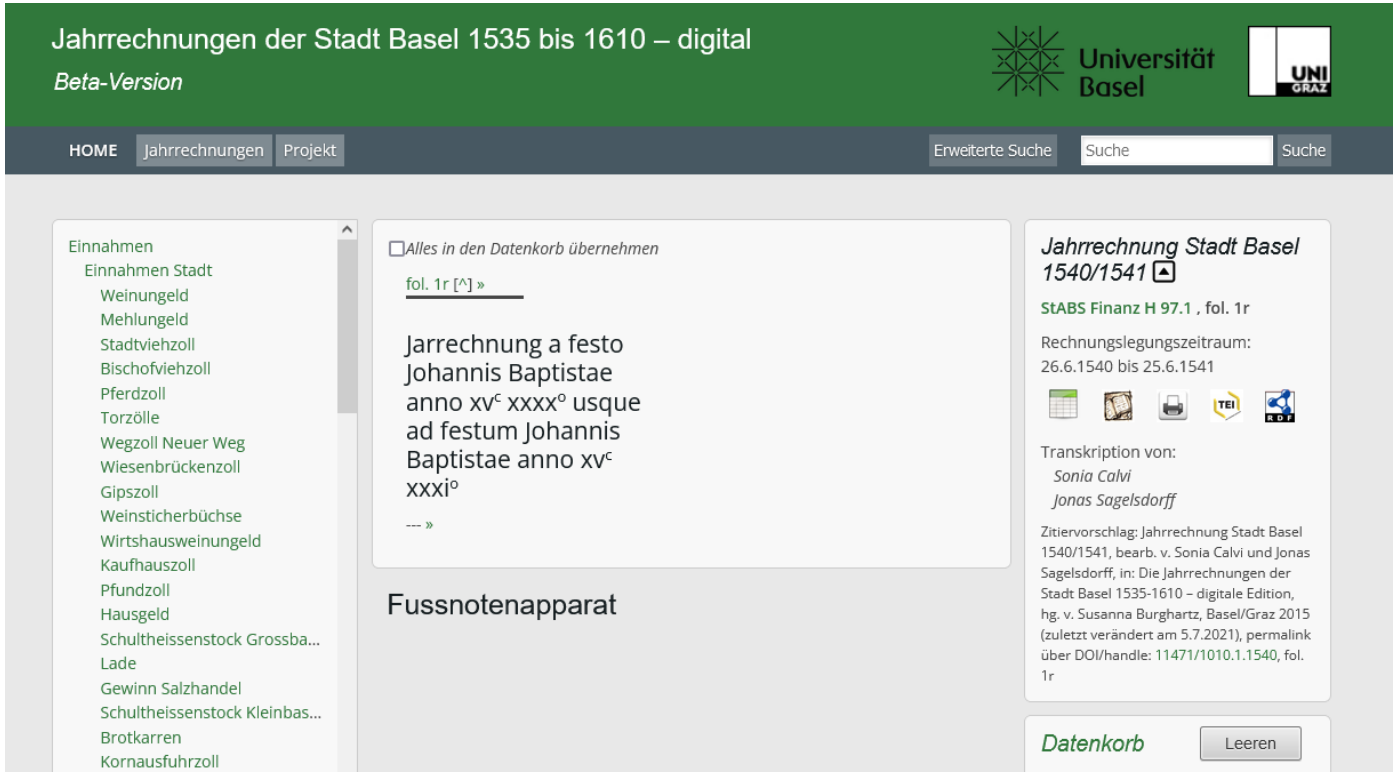
Project Background: Digitising Demographic Statistics of Zurich

- Historical Statistics of Switzerland Online (HSSO)
- Willi Bretscher Fellowship Project by Dr. Joël Floris: Digitizing demographic data from Zurich, 1910-1925, to quantify and contextualize the «Spanish flu»
- Facsimile incl. OCR published on Zurich Open Platform (ZOP) by Central Library (ZB)
- “Pilot” project 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



FAIR Principles

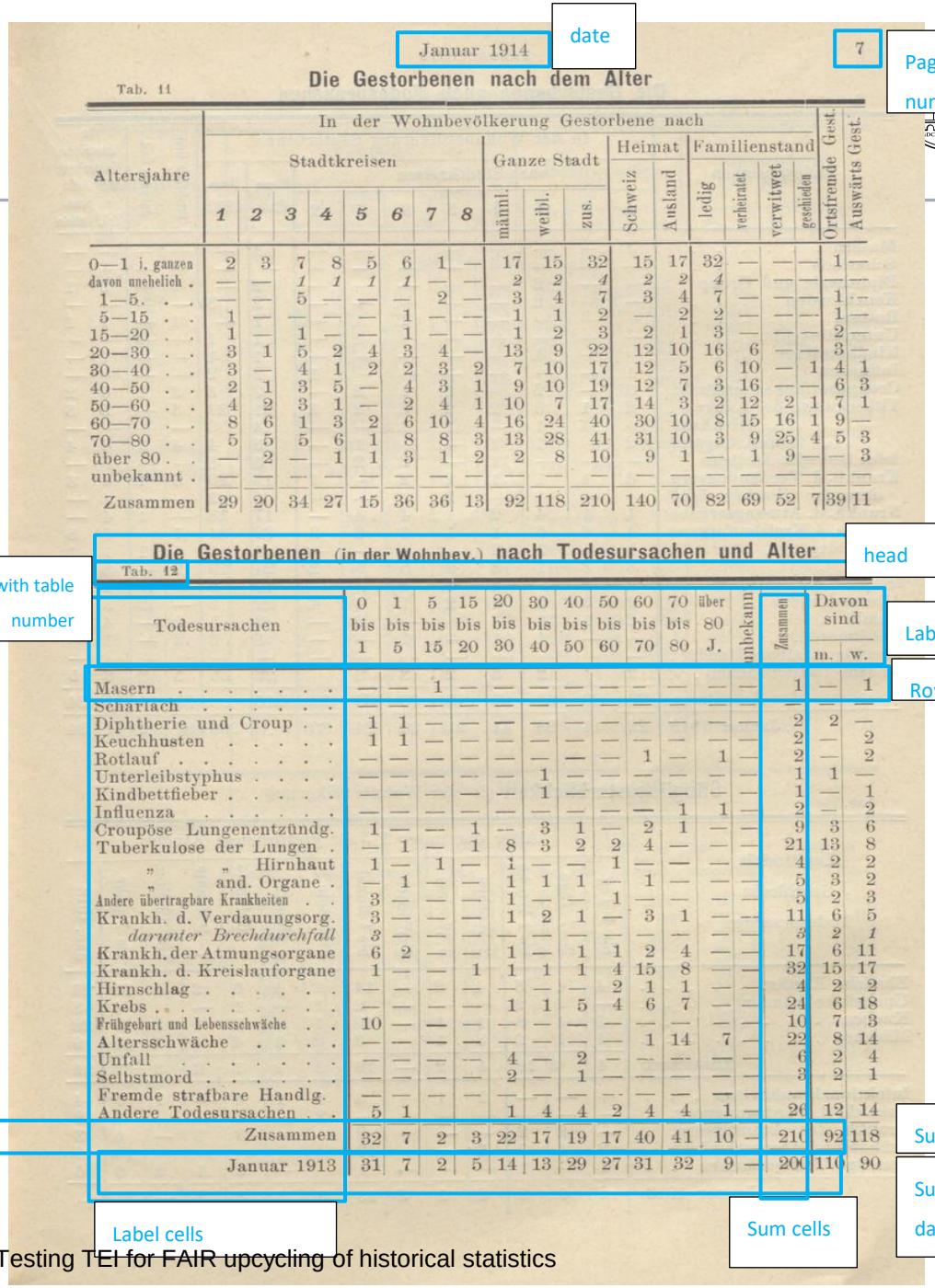
- **F**indable: Persistent Identifier and rich metadata
- **A**ccessible: (Meta)data retrievable (openly) online
- **I**nteroperable: Platform-/OS independent
- **R**eusable: Well-documented (meta)data with clear data usage license

Text Encoding Initiative (TEI)

- Consortium developing and maintaining standards for the digital representation of texts since 1994
- Version 4.8.0. of the guidelines published Sept 2nd
- 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 presentation are more important than in continuous text
  - Table processing in XML notations other than TEI

Structure Original Table

- Month of appearance
- Page number
- Title
- Table No.
- Text and numeric column headers
- Sums both in columns and rows
- Row with sum of same month of previous year
- Zero values with long dash –



Januar 1914

Tab. 11 **Die Gestorbenen nach dem Alter**

Altersjahre	In der Wohnbevölkerung Gestorbene nach: Stadtkreisen 1	2	3	4	5	6	7	8	Ganze Stadt: männl.	weibl.	zus.	Heimat: Schweiz	Ausland	Familienstand: ledig	verheiratet	verwitwet	geschieden	Ortsfremde Gest.	Auswärts Gest.
0—1 i. ganzen	2	3	7	8	5	6	1	—	17	15	32	15	17	32	—	—	—	1	—
davon unehelich	—	—	1	1	1	1	—	—	2	2	4	2	2	4	—	—	—	—	—
1—5.	—	—	5	—	—	—	2	—	3	4	7	3	4	7	—	—	—	1	—
5—15	1	—	—	—	—	1	—	—	1	1	2	—	2	2	—	—	—	1	—
15—20	1	—	1	—	—	1	—	—	1	2	3	2	1	3	—	—	—	2	—
20—30	3	1	5	2	4	3	4	—	13	9	22	12	10	16	6	—	—	3	—
30—40	3	—	4	1	2	2	3	2	7	10	17	12	5	6	10	—	1	4	1
40—50	2	1	3	5	—	4	3	1	9	10	19	12	7	3	16	—	—	6	3
50—60	4	2	3	1	—	2	4	1	10	7	17	14	3	2	12	2	1	7	1
60—70	8	6	1	3	2	6	10	4	16	24	40	30	10	8	15	16	1	9	—
70—80	5	5	5	6	1	8	8	3	13	28	41	31	10	3	9	25	4	5	3
über 80	—	2	—	1	1	3	1	2	2	8	10	9	1	—	1	9	—	—	3
unbekannt	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
Zusammen	29	20	34	27	15	36	36	13	92	118	210	140	70	82	69	52	7	39	11

Die Gestorbenen (in der Wohnbev.) nach Todesursachen und Alter

Tab. 12

Todesursachen	0 bis 1	1 bis 5	5 bis 15	15 bis 20	20 bis 30	30 bis 40	40 bis 50	50 bis 60	60 bis 70	70 bis 80	über 80 J.	unbekannt	Zusammen	Davon sind: m.	w.
Masern	—	—	1	—	—	—	—	—	—	—	—	—	1	—	1
Scharlach	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
Diphtherie und Croup	1	1	—	—	—	—	—	—	—	—	—	—	2	2	—
Keuchhusten	1	1	—	—	—	—	—	—	—	—	—	—	2	—	2
Rotlauf	—	—	—	—	—	—	—	—	1	—	1	—	2	—	2
Unterleibstyphus	—	—	—	—	—	1	—	—	—	—	—	—	1	1	—
Kindbettfieber	—	—	—	—	—	1	—	—	—	—	—	—	1	—	1
Influenza	—	—	—	—	—	—	—	—	—	1	1	—	2	—	2
Croupöse Lungenentzündg.	1	—	—	1	—	3	1	—	2	1	—	—	9	3	6
Tuberkulose der Lungen	—	1	—	1	8	3	2	2	4	—	—	—	21	13	8
„ „ Hirnhaut	1	—	1	—	1	—	—	1	—	—	—	—	4	2	2
„ and. Organe	—	1	—	—	1	1	1	—	1	—	—	—	5	3	2
Andere übertragbare Krankheiten	3	—	—	—	1	—	—	1	—	—	—	—	5	2	3
Krankh. d. Verdauungsorg.	3	—	—	—	1	2	1	—	3	1	—	—	11	6	5
darunter Brechdurchfall	*3*	—	—	—	—	—	—	—	—	—	—	—	*3*	*2*	*1*
Krankh. der Atmungsorgane	6	2	—	—	1	—	1	1	2	4	—	—	17	6	11
Krankh. d. Kreislauforgane	1	—	—	1	1	1	1	4	15	8	—	—	32	15	17
Hirnschlag	—	—	—	—	—	—	—	2	1	1	—	—	4	2	2
Krebs	—	—	—	—	1	1	5	4	6	7	—	—	24	6	18
Frühgeburt und Lebensschwäche	10	—	—	—	—	—	—	—	—	—	—	—	10	7	3
Altersschwäche	—	—	—	—	—	—	—	—	1	14	7	—	22	8	14
Unfall	—	—	—	—	4	—	2	—	—	—	—	—	6	2	4
Selbstmord	—	—	—	—	2	—	1	—	—	—	—	—	3	2	1
Fremde strafbare Handlg.	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
Andere Todesursachen	5	1			1	4	4	2	4	4	1	—	26	12	14
Zusammen	32	7	2	3	22	17	19	17	40	41	10	—	210	92	118
Januar 1913	31	7	2	5	14	13	29	27	31	32	9	—	200	110	90

date

7

Page number

head

Note with table number

Label row

Row 1

Sum row

Sum row date 1913

Label cells

Sum cells

Metadata of Original Table in TEI-XML

TEI Header Elements

- Title
- Edition Statement
- Publication Statement
- Bibliographic Information on original source

```
<?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-->
```

Structure Original Table in TEI-XML

TEI Text Body

- Title
- Table no. 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

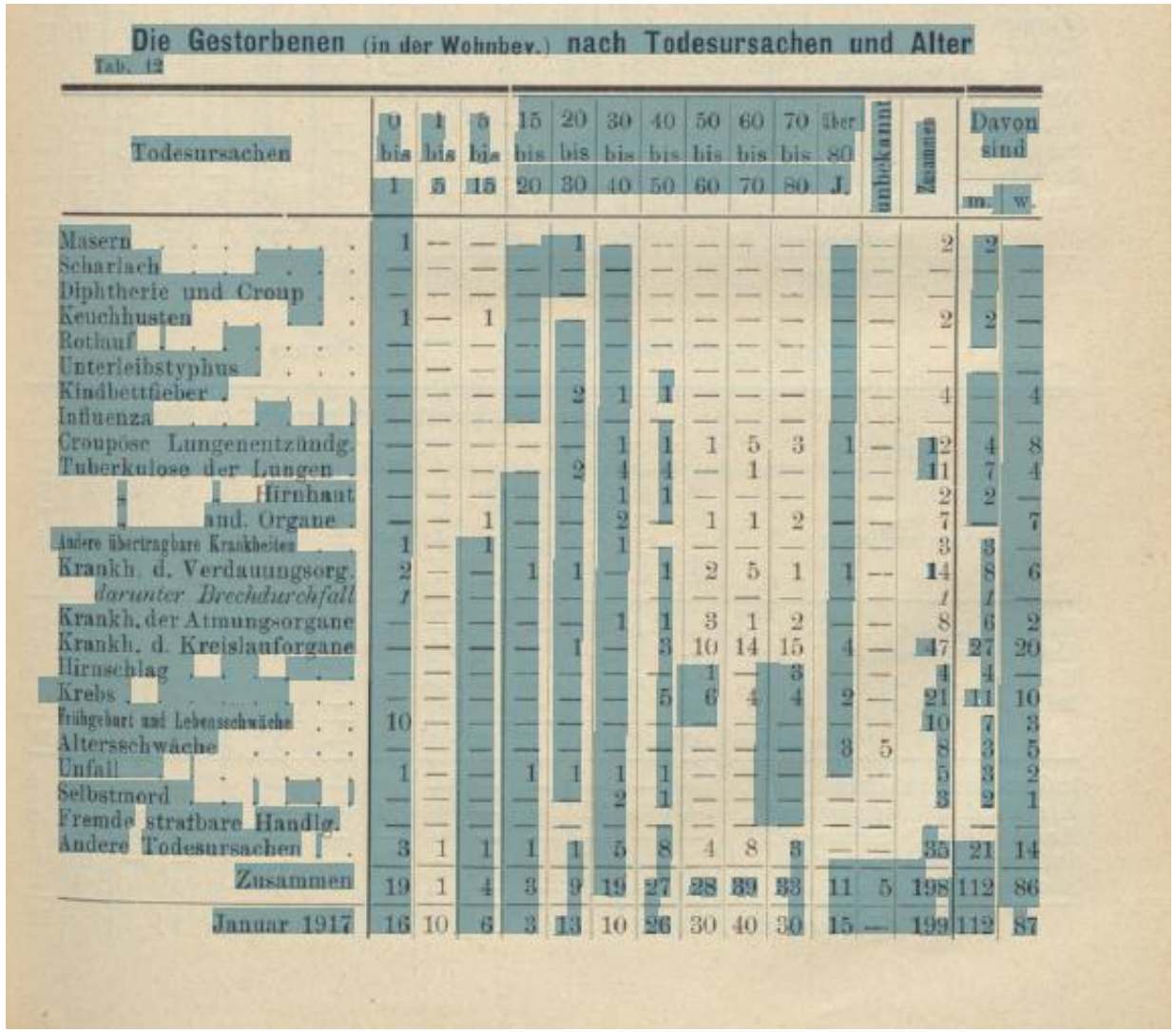
→ Human and machine readable but error prone…

```
<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>
```

Challenges

- Comprehensive TEI text elements require training time for practical implementation in XML
- Manual transfer of table content is prone to errors
- TEI's XML generates long code file for only one table
- tbc…

Challenge: Original OCR Output



Die Gestorbenen (in der Wohnbev.) nach Todesursachen und Alter

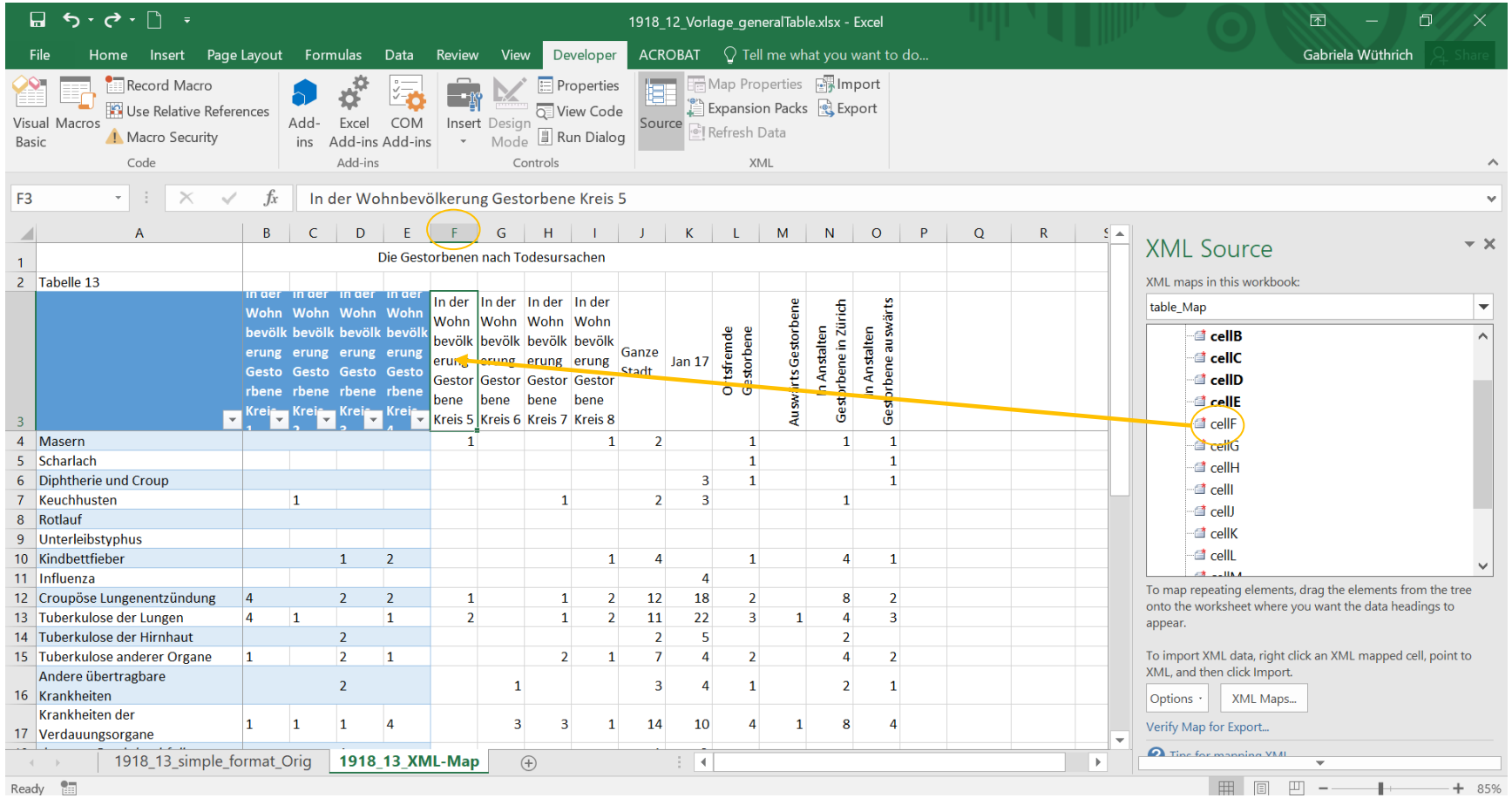
Tab. 12

Todesursachen	U bis 1	1 bis 5	5 bis 15	15 bis 20	20 bis 30	30 bis 40	40 bis 50	50 bis 60	60 bis 70	70 bis 80	über 80 J.	unbekannt	Zusammen	Davon sind m.	Davon sind w.
Masern	1	—	—		1								2	2	
Scharlach	—	—	—										—		
Diphtherie und Croup	—	—	—												
Keuchhusten	1	—	1										2	2	
Rotlauf	—	—	—												
Unterleibstyphus	—	—	—												
Kindbettfieber	—	—	—		2	1	1						4		4
Influenza	—	—	—												
Croupöse Lungenentzündg.	—	—	—			1	1	1	5	3	1	—	12	4	8
Tuberkulose der Lungen	—	—	—		2	4	4	—	1	—			11	7	4
Hirnhaut	—	—	—			1	1						2	2	
and. Organe	—	—	1			2		1	1	2			7		7
Andere übertragbare Krankheiten	1	—	1			1							3	3	
Krankh. d. Verdauungsorg.	2	—	—	1	1		1	2	5	1	1	—	14	8	6
darunter Brechdurchfall	*1*	—	—										*1*	*1*	
Krankh. der Atmungsorgane	—	—	—			1	1	3	1	2			8	6	2
Krankh. d. Kreislauforgane	—	—	—	—	1	—	3	10	14	15	4	—	47	27	20
Hirnschlag	—	—	—					1	—	3			4	4	
Krebs	—	—	—				5	6	4	4	2	—	21	11	10
Frühgeburt und Lebensschwäche	10	—	—										10	7	3
Altersschwäche	—	—	—					—	—	—	3	5	8	3	5
Unfall	1	—	—	1	1	1	1	—	—	—			5	3	2
Selbstmord	—	—	—			2	1	—	—	—	—	—	3	2	1
Fremde strafbare Handlg.	—	—	—	—	—		—	—	—	—	—	—	—	—	
Andere Todesursachen	3	1	1	1	1	5	8	4	8	3	—	—	35	21	14
Zusammen	19	1	4	3	9	19	27	28	39	33	11	5	198	112	86
Januar 1917	16	10	6	3	13	10	26	30	40	30	15	—	199	112	87

OCR text copied from pdf in table format

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

Challenge: Implementing Simple XML Table Structure in Excel



Challenges

- Comprehensive TEI text elements require training time for practical implementation in XML
- Manual transfer of table content is prone to errors
- TEI's XML generates long code file for only one table

- OCR recognition in digital copies needs to be adjusted for tables – or image quality needs to be good enough
- TEI schema not suitable for direct table mapping in Excel *yet*
- OCR and TEI are flow text oriented → Table recognition and XML structure are correspondingly poor for structured text

  → Chicken-and-egg problem: More (XML) table templates for better text recognition (esp. wrt LLMs)?

Conclusion

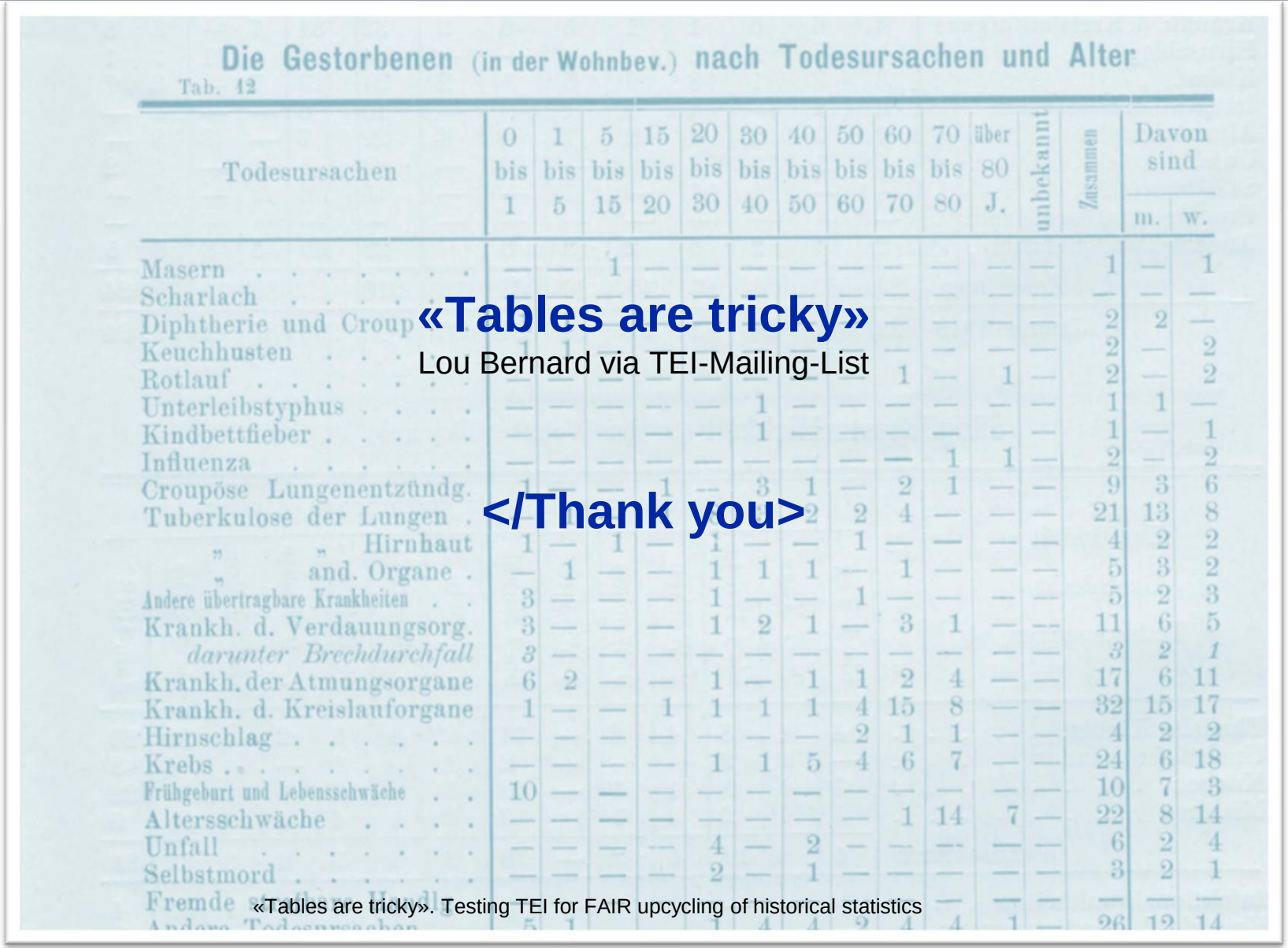
- Upcycling tables according to TEI guidelines is generally possible
- TEI offers tools that can be used to capture complex structures of historical tables in XML format
- TEI can be used to track changes in table structures
- TEI enables platform- and software-independent processing of table data – so do .csv and .txt
- TEI generally suitable for structuring continuous text, less suitable for texts where structure determines meaning

Potential Extensions

- Update to current TEI version
- Integration of additional TEI elements for formatting (e.g. separators, long dashes) and meaning (gender)

- Using LLMs to automatically do the OCR to XML procedure, BUT…
  - Image quality and OCR remain problematic
  - Proper and detailed explanation of complex structure in prompting
  - Easy transformation from xlsx to XML
- Testing different Python tools for table recognition, e.g. using Aurelius Noble's workshop

- Check developments in other digitisation projects
  - XSLT (eXtensible Stylesheet Language Transformation) for automated structuring of Excel



Die Gestorbenen (in der Wohnbev.) nach Todesursachen und Alter

Tab. 12

Todesursachen	0 bis 1	1 bis 5	5 bis 15	15 bis 20	20 bis 30	30 bis 40	40 bis 50	50 bis 60	60 bis 70	70 bis 80	über 80 J.	unbekannt	Zusammen	Davon sind m.	Davon sind w.
Masern	—	—	1	—	—	—	—	—	—	—	—	—	1	—	1
Scharlach	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
Diphtherie und Croup	—	—	—	—	—	—	—	—	—	—	—	—	2	2	—
Keuchhusten	1	1	—	—	—	—	—	—	—	—	—	—	2	—	2
Rotlauf	—	—	—	—	—	—	—	—	1	—	1	—	2	—	2
Unterleibstyphus	—	—	—	—	—	1	—	—	—	—	—	—	1	1	—
Kindbettfieber	—	—	—	—	—	1	—	—	—	—	—	—	1	—	1
Influenza	—	—	—	—	—	—	—	—	—	1	1	—	2	—	2
Croupöse Lungenentzündg.	1	—	—	1	—	3	1	—	2	1	—	—	9	3	6
Tuberkulose der Lungen	—	—	—	—	1	2	2	2	4	—	—	—	21	13	8
„ „ Hirnhaut	1	—	1	—	1	—	—	1	—	—	—	—	4	2	2
„ and. Organe	—	1	—	—	1	1	1	—	1	—	—	—	5	3	2
Andere übertragbare Krankheiten	3	—	—	—	1	—	—	1	—	—	—	—	5	2	3
Krankh. d. Verdauungsorg.	3	—	—	—	1	2	1	—	3	1	—	—	11	6	5
darunter Brechdurchfall	*3*	—	—	—	—	—	—	—	—	—	—	—	*3*	*2*	*1*
Krankh. der Atmungsorgane	6	2	—	—	1	—	1	1	2	4	—	—	17	6	11
Krankh. d. Kreislauforgane	1	—	—	1	1	1	1	4	15	8	—	—	32	15	17
Hirnschlag	—	—	—	—	—	—	—	2	1	1	—	—	4	2	2
Krebs	—	—	—	—	1	1	5	4	6	7	—	—	24	6	18
Frühgeburt und Lebensschwäche	10	—	—	—	—	—	—	—	—	—	—	—	10	7	3
Altersschwäche	—	—	—	—	—	—	—	—	1	14	7	—	22	8	14
Unfall	—	—	—	—	4	—	2	—	—	—	—	—	6	2	4
Selbstmord	—	—	—	—	2	—	1	—	—	—	—	—	3	2	1
Fremde	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
Andere Todesursachen	5	1			1	4	4	2	4	4	1	—	26	12	14

«Tables are tricky»

Lou Bernard via TEI-Mailing-List

</Thank you>